Mestrado em Informática e Sistemas

Um estudo da Diabetes Mellitus e Hipertensão Arterial baseado em técnicas de Data Mining aplicadas a dados da Administração Regional de Saúde do Centro

Dissertação apresentada para a obtenção do grau de Mestre em Informática e Sistemas (MIS) – Especialização em Tecnologias da Informação e do Conhecimento

Autor
Luís Candeias Borges

Orientador
Prof. Doutor Viriato Marques
Instituto Politécnico de Coimbra

Coimbra, dezembro de 2016

AGRADECIMENTOS

A elaboração de uma dissertação envolve sempre diversas pessoas, quer em representação institucional quer pessoal. Embora a todos se deva um sincero agradecimento pelo seu contributo, fica um agradecimento especial a algumas pessoas que foram mais além no seu apoio e contribuição.

O presente trabalho foi realizado sob a orientação do Professor Doutor Viriato Marques, a quem me cabe exprimir sincero reconhecimento pelo modo empenhado como supervisionou o trabalho, mostrando sempre disponibilidade e compreensão no decurso da dissertação ao que acresce o seu imprescindível conhecimento científico.

Durante a elaboração desta dissertação, foi fundamental recorrer ao uso de conhecimento sobre metodologia científica que adquiri do Professor Doutor Jorge Bernardino. Por aquilo que me transmitiu e por me ter demonstrado a sua importância, exprimo a minha admiração ao seu conhecimento e agradeço os seus ensinamentos.

Agradeço à Administração Regional de Saúde do Centro, e em particular ao Engenheiro Ernesto Fernandes, a colaboração prestada no projeto de investigação de que resultou o trabalho aqui apresentado, pela disponibilidade manifestada na cedência dos meios materiais e humanos que se revelaram necessários à realização dos ensaios experimentais.

Aqueles que me são mais próximos foram, sem dúvida, os que mais sentiram a evolução deste trabalho pela companhia que não lhes fiz. Apesar disso, da Helena recebi sempre compreensão, apoio e incentivo, que me permitiu acalentar a esperança de ver este trabalho terminado. O Bruno, o Filipe e o Alexandre (que nasceu durante o período de elaboração da dissertação) sempre souberam presentar-me com sorrisos, compreensão e dispensa de muitas das suas brincadeiras de criança, num período das suas vidas onde estas são fundamentais e que me permitiu alentar a ver alcançada mais esta meta.

A todos a minha eterna gratidão.

RESUMO

Esta dissertação resulta de um acordo de colaboração entre a Administração Regional de Saúde do Centro e o Instituto Superior de Engenharia de Coimbra do Instituto Politécnico de Coimbra, e visa estudar os dados sobre as doenças Diabetes Mellitus e Hipertensão, aplicando métodos de Tecnologias da Informação e do Conhecimento, integrados na área científica de Business Intelligence e Data Mining.

Na atualidade, a Diabetes Mellitus e Hipertensão são patologias incuráveis e o número de pessoas afetadas continua a agravar-se. Existe um forte interesse em abordagens realizáveis e de custo suportável, especialmente em casos não diagnosticados, para intervenção o mais cedo possível. Existe interesse em encontrar sistemas de identificação de pacientes sem recorrer a testes bioquímicos. Com a proliferação das Tecnologias de Informação na sociedade, desenvolver a baixo custo e de acesso generalizado pode fazer decrescer o número de pacientes não diagnosticados. As ferramentas devem auxiliar o processo de identificar quem poderá ser afetado para reduzir riscos preventivamente. O uso das Tecnologias de Informação pelos prestadores de cuidados de saúde em conjunto com educação dos pacientes resultará em benefícios significativos na luta contra estas doenças crónicas. É neste segmento que este estudo se inclui, adicionando as tecnologias de informação às abordagens tradicionais.

Neste trabalho aplicam-se técnicas de Data Mining para extrair conhecimento dos dados existentes no Data Warehouse da Administração Regional de Saúde do Centro. A primeira parte carateriza as doenças para determinar quais os seus aspetos mais relevantes a considerar no desenvolvimento das restantes tarefas. A segunda parte carateriza métodos e metodologias de Data Mining com o intuito de descrever as tarefas e técnicas utilizadas. Existem atualmente diversas ferramentas que implementam os diversos algoritmos de extração do conhecimento pelo que na terceira parte do trabalho é realizada uma comparação com vista à seleção informada e esclarecida da ferramenta base a utilizar neste estudo. A quarta parte é baseada nos passos comuns das metodologias aplicáveis aos estudos deste género e que consiste em compreender os dados, preparar os dados, proceder à sujeição dos dados aos algoritmos e avaliar os resultados dos modelos inferidos. A última etapa do estudo elabora algumas conclusões e sugere trabalho futuro.

A saúde é, por natureza, de importância vital para o ser humano. Devemos, por isso, evidenciar todos os esforços possíveis para que novo conhecimento possa ser gerado e usado.

Palavras-chave: Diabetes Mellitus, Hipertensão Arterial, Data Mining, Classificação, Associação.

ABSTRACT

This dissertation is the result of a collaboration agreement between the Administração Regional de Saúde do Centro and the Instituto Superior de Engenharia de Coimbra of Instituto Politécnico de Coimbra, and aims to study data on Diabetes Mellitus and Hypertension diseases by applying methods of Information and Knowledge Technology, integrated into the scientific area of Business Intelligence and Data Mining.

Currently, Diabetes Mellitus and Hypertension are incurable diseases and the number of affected people continues to worsen. There is a strong interest in realizable approaches and affordable solutions, especially in undiagnosed cases, for the earliest possible intervention. There is interest in finding patient identification systems without the use of biochemical tests. With the proliferation of Information Technology in society, develop at low cost and widespread access can decrease the number of undiagnosed patients. Tools should help the process of identifying who may be affected in order to reduce risks proactively. The use of Information Technology by health care providers in conjunction with patient education will provide significant benefits in the fight against these chronic diseases. It is in this segment that this study is included, adding Information Technology to traditional approaches.

In this work data mining techniques are applied to extract knowledge from existing data in the Data Warehouse of the Administração Regional de Saúde do Centro. The first part characterizes the diseases to determine the most relevant issues to be addressed in the development of the remaining tasks. The second part characterizes methods and data mining methodologies in order to describe the tasks and techniques used. There are several tools nowadays that implement the various knowledge extraction algorithms so on the third part of the work is carried out a comparison with the aim to have an informed and clarified selection of the base tool to use in this study. The fourth part is based on the common steps of the methodologies applicable to studies of this kind and that is to understand the data, prepare the data, subject the data to algorithms and evaluate the results of the inferred models. The last stage of the study elaborates some conclusions and suggests future work.

Health is, by nature, of vital importance for the human being. We must, therefore, show every possible effort so that new knowledge can be generated and used.

Keywords: Diabetes Mellitus, Arterial Hypertension, Data Mining, Classification, Association.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE QUADROS

SIMBOLOGIA

mg/dl – Miligramas por decilitro

mmol/l – Milimol por litro

HbA1c – Hemoglobina glicada

mm Hg – Milímetro de mercúrio

g – Gramas

n – Número de registos num conjunto de dados

O (n^3) – Símbolo de Landau de complexidade algorítmica cúbica

k – Parâmetro de grupos ou *clusters* de um conjunto de dados

NP-hard – Problemas de complexidade computacional elevada que vão até ao insolúvel em tempo útil

O (n*log n) – Símbolo de Landau de complexidade algorítmica polilogarítmica

Km^2 – Kilómetros quadrados

Kg – Kilogramas

cm – Centímetros

m - Metros

N/A – Não aplicável

Err – Erro

inf – Valores inferiores

sup – Valores superiores

ABREVIATURAS

Português		Inglês	
Abrev.	Significado	Abrev.	Significado
TI	Tecnologia (s) de Informação	IT	Information System(s)
SI	Sistema (s) de Informação	IS	Information System(s)
SAD	Sistema (s) de Apoio à Decisão	DSS	Decision Support System(s)
USD	Dólar Norte-Americano	USD	United States Dollars
IMC	Índice de massa corporal	BMI	Body-Mass Index
PIB	Produto Interno Bruto	GDP	Gross Domestic Product
SNS	Serviço Nacional de Saúde	NHS	National Healthcare System
DGS	Direção Geral da Saúde	NDH	National Department of Health
PTGO	Prova de tolerância à glicose oral	OGTT	Oral Glicose tolerance test
TAS ou PAS	Tensão Arterial Sistólica ou Pressão Arterial Sistólica	SBP	Systolic Blood Pressure
TAD ou PAD	Tensão Arterial Diastólica ou Pressão Arterial Diastólica	DBP	Diastolic Blood Pressure
OMS	Organização Mundial de Saúde	WHO	World Health Organization
DCBD	Descoberta de Conhecimento em Bases de Dados	KDD	Knowledge Discovery in Databases
PECI-MD	Processo Standard de Cruzamento Industrial de Mineração de Dados	CRISP-DM	Cross Industry Standard Process for Data Mining
AEMMA	Amostragem, Exploração, Modificação, Modelação e Avaliação	SEMMA	Sample, Explore, Modify, Model and Assess
AD	Árvore de decisão	DT	Decision Tree
RNA	Rede Neuronal Artificial	ANN	Artificial Neural Network
AG	Algoritmo Genético	GA	Genetic Algorithm
MVS	Máquina Vetorial de Suporte	SVM	Support Vector Machine
RBC	Raciocínio baseado em casos	CBR	Cases-based Reasoning

CRM	Gestão de Relacionamento com o Cliente	CRM	Customer Relationship Management
EUA	Estados Unidos da América	USA	United States of America
IBL	Aprendizagem baseada em instâncias	IBL	Instance-based Learner
ARSC	Administração Regional de Saúde do Centro	CRDH	Centre Region Department of Health
PNS	Programa Nacional de Saúde	HNP	Health National Plan
ACeS	Agrupamento de Centros de Saúde	HCG	Health Centre Group
ULS	Unidade Local de Saúde	HLU	Health Local Unit
CS	Centro de Saúde	HC	Health Center
CSP	Cuidados de Saúde Primários	PHC	Primary Health Care
UCSP	Unidades de Cuidados de Saúde Primários	PHCU	Primary Health Care Unit
USF	Unidades de Saúde Familiar	FHU	Family Health Unit
MCDT	Meios Complementares de Diagnóstico e Tratamento	CDTM	Complementary Diagnostic and Treatment Means
AVC	Acidente Vascular Cerebral	CVA	Stroke, cerebrovascular accident
DCI	Denominação Comum Internacional	INN	International Nonproprietary Name
CICP	Classificação Internacional de Cuidados Primários	ICPC	International Classification of Primary Care

Os acrónimos não foram traduzidos quando não são usados na literatura em Português.

1. INTRODUÇÃO

Esta dissertação resulta de um acordo de colaboração entre a Administração Regional de Saúde do Centro e o Instituto Superior de Engenharia de Coimbra do Instituto Politécnico de Coimbra, e visa estudar os dados sobre as doenças *Diabetes Mellitus* e Hipertensão, aplicando métodos de Tecnologias da Informação e do Conhecimento, integrados na área científica de *Business Intelligence* da qual fazem parte, e as subáreas *Data Warehousing* e *Data Mining*.

As organizações atuais têm hoje ao seu dispor tecnologias de informação que lhes permite conhecer melhor e gerir bem os seus processos de negócio. A contínua baixa de custo no armazenamento e processamento de informação levou a que as organizações possuam hoje uma enorme quantidade de dados sobre as suas atividades.

Os dados sobre o negócio estão armazenados em repositórios de informação, vulgo bases de dados, que os gestores usam no seu quotidiano. É comum que os sistemas operacionais, aqueles que são usados nas operações do dia-a-dia das organizações para registo das suas atividades, não sejam usados pelos gestores por razões de disponibilidade e é por esse motivo primordial que são usados outros repositórios de informação, conhecidos por Data Warehouses. Estes são necessariamente diferentes em estrutura visto que visam permitir à gestão aceder a informações sobre o negócio não causando significativos distúrbios às atividades operacionais. Os dados são copiados dos sistemas operacionais para o Data Warehouse sofrendo transformações para que estes se afastem do cariz técnico e se aproximem mais aos conceitos do domínio da gestão e para que a memória organizacional persista no tempo. A partir desses repositórios de informação é possível usar métodos de Data Mining e explorar os dados com o objetivo de descobrir conhecimento.

Este estudo concentra-se nas doenças Diabetes Mellitus e Hipertensão por serem doenças ainda incuráveis e pelo facto dos números de pessoas afetadas continuar a agravar-se mundialmente como o comprova a (IDF - International Diabetes Federation, 2012) e em Portugal a (Sociedade Portuguesa de Diabetologia, 2013). Em ambas as doenças existe um interesse forte em abordagens realizáveis e de custo suportável. Esta ideia é corroborada pelo estudo de (Brown, Critchley, Bogowicz, Mayige, & Unwin, 2012) especialmente em casos não diagnosticados para intervenção o mais cedo possível. Os autores referem o interesse em encontrar sistemas de identificação de pacientes sem recorrer a testes bioquímicos. Com a proliferação das TI na sociedade, desenvolver SI de baixo custo e acesso generalizado pode fazer decrescer o número de pacientes não diagnosticados.

Segundo os investigadores (Akinci, Coyne, Healey, & Minear, 2004) houve uma alteração de paradigma em gestão de doenças nos últimos anos. Os esforços de controlo passaram de doenças transmissíveis para doenças crónicas. As doenças crónicas são um problema que afeta não só o doente, mas toda a comunidade; a família e prestadores de cuidado, mas também os contribuintes que são chamados a suportar os custos. O rigoroso controlo destas doenças, quer através de terapêutica adequada quer através de redução de riscos relacionados com a adoção de corretos estilos de vida, resulta na evitação dos custos relacionados com essas complicações.

A investigação destas doenças é tradicionalmente realizada usando metodologias sobre cuidados de saúde e ensaios clínicos na área científica da medicina, mas o combate à doença é sobretudo ganho recorrendo a uma abordagem mais generalizada (Akinci, Coyne, Healey, & Minear, 2004). É neste segmento que este estudo se inclui, adicionando as tecnologias de informação às abordagens tradicionais. Para um efetivo controlo das doenças, os prestadores de cuidados de saúde procuram saber o seguinte: (1) definir objetivos de tratamento individualizado; (2) avaliar a qualidade do tratamento sugerido; (3) identificar áreas onde é requerida mais atenção; e (4) redirecionar os doentes para outros especialistas atempadamente e conforme for necessário. As tecnologias de informação facilitam esse trabalho e este estudo incidirá particularmente no primeiro ponto auxiliando os prestadores de cuidados nos aspetos de diagnóstico.

O efetivo controlo destas doenças é realizado fornecendo as ferramentas adequadas aos prestadores de cuidados de saúde, mas também, e essencialmente, através da educação dos pacientes e famílias sobre como diminuir as complicações associadas às doenças. Apesar de todos os esforços no controlo das doenças, cabe ao paciente em último caso entender que é através da sua ação preventiva ou reativa que a ação produzirá os efeitos desejados.

Qualquer entidade prestadora de cuidados de saúde tem atualmente ao seu dispor ferramentas de gestão de informação. Podemos generalizar que as entidades com maiores recursos financeiros têm acesso a melhores ferramentas de tecnologias de informação e de auxílio ao processo de decisão e diagnóstico. É de crucial importância a realização de estudos como este que criam conhecimento generalizado e usável em organizações que assim não necessitam de recorrer a elevados recursos financeiros.

Apesar de existirem vários SI que tratam de registos de pacientes, rastreio computorizado e sistemas de alerta, bem como ferramentas de auxílio à decisão e diagnóstico (SAD), ainda existe espaço para melhorar (Akinci, Coyne, Healey, & Minear, 2004). Estes autores relatam que muitos prestadores de saúde não recorrem a TI pois servem o propósito primordial de guardar informações sobre os pacientes e não lhes é dado nenhum incentivo através de meios auxiliares de diagnóstico ou definição de tratamento, contudo, apontam como razão principal para essa não adoção o custo.

Afirmam (Akinci, Coyne, Healey, & Minear, 2004) que existe uma tremenda necessidade por mais investigação no uso de SI, na área da prevenção e controlo das doenças crónicas, e que "o investimento em melhor Data Mining para precisar melhor aqueles em risco de virem a desenvolver complicações" resultará em melhores processos de cuidados e prognósticos. As ferramentas devem auxiliar o processo de identificar quem poderá ser afetado para reduzir riscos preventivamente. O uso das TI pelos prestadores de cuidados de saúde em conjunto com educação dos pacientes resultará em benefícios significativos na luta contra estas doenças crónicas.

2. DIABETES MELLITUS E HIPERTENSÃO ARTERIAL

2.1. Diabetes Mellitus

A Diabetes Mellitus, vulgo Diabetes, é a incapacidade do nosso organismo de utilizar a glicose de forma equilibrada. É uma doença crónica que se manifesta em organismos que produzem a hormona insulina de forma deficiente ou que criam uma forma de resistência a esta. Tem como resultado um aumento dos níveis de glicose (açúcar no sangue), que são a principal fonte de energia do organismo. A glicemia é a quantidade de glicose existente no sangue. Quando em excesso estamos perante uma Hiperglicemia. Quando em défice estamos perante uma Hipoglicemia. A Hiperglicemia Intermédia é uma condição em que os doentes têm níveis de glicose superiores ao normal, mas ainda assim insuficientes para serem diagnosticados como Diabetes.

A Diabetes é classificada em três tipos:

- Diabetes Tipo I;
- Diabetes Tipo II;
- Diabetes Gestacional.

A Diabetes Tipo I é aquela em que o organismo do paciente necessita de insulina vinda de fonte externa visto que não produz insulina, ou seja, quando as células Betas do pâncreas deixaram de produzir insulina por causas ainda desconhecidas e foram destruídas pelo sistema imunitário. A reposição dos níveis de insulina no organismo é muitas vezes feita através de mecanismos de injeção, mais ou menos automatizados. Alguns são meros instrumentos de injeção tradicionais e outros chegam a ser capazes de medir os níveis de glicemia e administrar doses em quantidade regulada com a necessidade. A Diabetes Tipo I surge geralmente em crianças e jovens e manifesta-se através de sintomas sérios. O surgimento em adultos não é muito frequente. A terapêutica mantém-se ao longo de toda a vida do paciente diabético.

A Diabetes Tipo II é aquela em que o organismo é capaz de produzir insulina, mas resiste à ação desta. Quando o pâncreas deixa de produzir insulina em quantidade suficiente é necessária terapêutica com administração por injeção. Os pacientes têm, frequentemente, uma herança genética causadora da doença e associados estilos de vida pouco saudáveis com alimentação hipercalórica e sedentarismo.

A Diabetes gestacional corresponde a uma anomalia do metabolismo da glicose durante a gravidez, independentemente se é necessária ou não a terapêutica com insulina. Esta condição traduz-se em riscos acrescidos no recém-nascido tais como tamanho excessivo (macrossomia), traumatismo de parto, hipoglicemia e icterícia (Sociedade Portuguesa de Diabetologia, 2013). A Diabetes Gestacional apresenta um risco futuro acrescido de complicações nos níveis de glicose para o futuro das mulheres e filhos. No passado, este tipo de Diabetes resultava em altas taxas de mortalidade, tanto para a mãe como para o filho(a) (Kaplan, Lippe, Brinkman, Davidson, & Geffner, 1982).

A terapêutica é adaptada à vida do paciente; é normal a administração de insulina de ação intermédia misturada com insulina de ação rápida, de uma a duas vezes ao dia, podendo ser mais, antes do pequeno-almoço e antes do jantar ou deitar. O importante é assegurar que o organismo tem insulina durante o dia, com um possível aumento aquando da ingestão de alimentos.

De acordo com o relatório anual da (IDF - International Diabetes Federation, 2012), a taxa de prevalência da Diabetes mundial cifrou-se em 8,3% dos adultos com idades compreendidas entre os 20-79 anos. Este valor corresponde a 371 milhões de casos dos quais 187 milhões (50,4%) estão ainda por diagnosticar: a taxa varia consoante a região mundial agravando-se nos países com rendimentos mais baixos (cerca de 80% na África Subsariana, entre 50-60% no Médio Oriente e Norte de África, cerca de 50% no Sudeste Asiático e cerca de 60% no Pacífico Ocidental). As despesas em cuidados de saúde relacionadas com doentes diabéticos totalizaram 471,6 mil milhões de USD a que corresponde 1.271,15 USD por pessoa. Foram registados 4,8 milhões de óbitos, 1,29% das pessoas com Diabetes, metade das quais com menos de 60 anos de idade.

O relatório (IDF - International Diabetes Federation, 2012) indica que na Europa a taxa de prevalência é de 8,4% a que correspondem 55 milhões de pessoas, o que se assemelha à situação mundial. As despesas de saúde relacionadas com pessoas com Diabetes na Europa quase que duplicaram (198,53%) em comparação com as despesas a nível mundial. Este número indica que apesar dos custos acrescidos na prestação de cuidados de saúde isso não se refletiu em decréscimo na taxa de prevalência, ou seja, no número de casos de pessoas com Diabetes.

Portugal tem uma taxa de prevalência da Diabetes muito superior à média Europeia. O relatório (IDF - International Diabetes Federation, 2012) indica que a taxa de prevalência para Portugal foi de 12,84%, muito acima da taxa mundial de 8,3% e da taxa europeia de 8,4%. Em termos absolutos, existiam no ano de 2012 1.031.620 pessoas com Diabetes em Portugal. A taxa de casos não diagnosticados cifra-se em 49,43% o que é o mesmo que dizer que perto de metade das pessoas com Diabetes não tem conhecimento que tem esta doença. Foram registados 7.890 óbitos, ou seja, 0,76% das pessoas com Diabetes, o que contrasta de forma positiva com a situação mundial onde a taxa é 1,29%. As despesas com cuidados de saúde relacionadas com Diabetes foram em Portugal de 2.521 USD por utente; na Europa este valor foi de 2.523,64 USD e no mundo de 1.271,15 USD.

Quadro 2:1 - Factos e Números sobre a Diabetes – distribuição geográfica (IDF - International Diabetes Federation, 2012)

Área geográfica	Taxa de prevalência	Adultos com Diabetes (milhões)	Despesa or utente (USD)
Mundo	8,3%	371	1.271,15
Europa	8,4%	55	2.523,64
Portugal	12,84%	1	2.521,48

Segundo o relatório anual da (Sociedade Portuguesa de Diabetologia, 2013), em 2011 a Diabetes afetava 12,7% da população portuguesa com idades compreendidas entre 20-79 anos de idade (aproximadamente 1.003.000 indivíduos), o que inclui 7,2% dos casos já diagnosticados e 5,5%

dos casos ainda por diagnosticar. A Diabetes tem uma incidência masculina, uma taxa de prevalência de 15,2% nos homens e 10,4% nas mulheres em relação à população com idades compreendidas entre os 20-79 anos de idade. Segundo a mesma fonte, existe uma forte correlação entre o envelhecimento dos indivíduos e a prevalência da doença; mais de um quarto da população com idades compreendidas entre 60-79 anos tem Diabetes. Relativamente ao IMC, uma pessoa obesa apresenta um risco 3 vezes superior de desenvolver a doença; a taxa de prevalência para indivíduos com IMC ≥ 30 era de 20%. Nos últimos 10 anos assistiu-se a um aumento da incidência da Diabetes de 80%. Em termos financeiros, a Diabetes representa 0,8% do PIB português ou 8% da despesa em saúde em 2011 que representa um custo direto estimado entre 1200-1450 milhões de euro.

A Diabetes Tipo II é mais frequente que a Diabetes Tipo I; a Diabetes Tipo II afeta acima de 90% dos pacientes diabéticos diagnosticados e a Diabetes Tipo I entre 5-10%; a taxa da Diabetes gestacional não é expressiva em termos absolutos (Sociedade Portuguesa de Diabetologia, 2013). O relatório indica que a Diabetes Tipo I atingia 0,14% da população portuguesa em 2011 a que correspondem mais de 3.000 indivíduos com idades entre 0-19 anos. Esta taxa foi de 0,12% em 2008, 0,13% em 2009 e 0,14% em 2010, o que revela uma tendência de ligeira progressão. Relativamente à Diabetes Gestacional, a taxa de prevalência para utentes no SNS progrediu de 3,4% em 2005 para 4,9% em 2011. A população parturiente no SNS totalizou 80% do total. A taxa de óbitos causados por Diabetes aumentou de 4,2% em 2002 para 4,4% da população portuguesa.

O relatório da (IDF - International Diabetes Federation, 2012) refere que o número de pessoas com Diabetes no mundo aumentará de 371 milhões em 2012 para 552 milhões em 2030. Na Europa este aumento será de 55 milhões em 2012 para 64 milhões em 2030. Quer isto dizer que a taxa de aumento da prevalência da Diabetes para 2030 a nível mundial será de 49% e que na Europa aumentará 16%. Este aspeto positivo na Europa contrasta com o facto de esta região ser aquela com a mais elevada taxa de prevalência da Diabetes Tipo I em crianças. Contrariamente ao que seria desejado, a taxa de prevalência da doença continua a aumentar, afetando indivíduos jovens.

Acumulam-se provas obtidas por investigação que um controlo rigoroso da Diabetes está associado com uma prevalência decrescida ou adiamento no aparecimento de certas complicações da Diabetes (Kaplan, Lippe, Brinkman, Davidson, & Geffner, 1982).

O surgimento dos sintomas da doença não afeta os organismos de forma grave, o que acarreta que os mesmos sejam com muita frequência menosprezados até que se tornem sérios. A Diabetes pode levar a situações de saúde críticas:

- A lesão da retina que pode levar à cegueira (retinopatia);
- A lesão renal que pode levar à necessidade de hemodiálise (nefropatia);
- A lesão nos nervos do organismo que pode levar a dores, sensação de calor ou frio, ou torpor nos pés e pernas (neuropatia);
- Hipoglicemia;
- Hiperglicemia;

- Doenças cérebro e cardiovasculares (angina de peito, ataques cardíacos e acidentes vasculares cerebrais);
- Alterações nos pés que, pela doença arterial, neuropatia e infeções mais difíceis de combater, leva ao aparecimento de feridas (úlceras) de difícil e prolongado tratamento que podem terminar pela necessidade de amputação (pé diabético);
- Disfunção e impotência sexual.

A adoção de estilos de vida saudáveis é, como em tantos outros casos, apontada como recomendação preventiva da doença, mas é de igual forma recomendado que se tenha atenção aos sintomas para início do tratamento o mais depressa possível. Os sinais de alerta (sintomas clássicos de descompensação) são:

- Sede constante e intensa (Polidipsia);
- Urinar em maior quantidade e com mais regularidade (Poliúria);
- Muita fome e dificuldade em saciá-la (Polifagia);
- Sensação de boca seca (Xerostomia);
- Fadiga, cansaço e perda de energia;
- Comichão no corpo (em especial nos genitais);
- Infeções recorrentes e com dificuldade em curar;
- Visão turva.

O diagnóstico é feito recorrendo a análises ao sangue após verificação que os níveis de glicemia estão acima de 126mg/dl, em jejum, ou 200mg/dl em qualquer ocasião. Visto que os sintomas vão surgindo lentamente na Diabetes Tipo II, é muito comum o diagnóstico ser feito em consultas consequentes de análises de rotina.

De acordo com a Norma da DGS N.º 2/2001 de 14 de janeiro de 2011 (Sociedade Portuguesa de Diabetologia, 2013), os critérios de diagnóstico de Diabetes são os seguintes:

- Glicemia de Jejum ≥ 126 mg/dl (ou ≥ 7,0 mmol/l); ou
- Sintomas clássicos de descompensação + Glicemia ocasional ≥ 200 mg/dl (ou ≥ 11,1 mmol/l); ou
- Glicemia ≥ 200 mg/dl (ou ≥ 11,1 mmol/l) às 2 horas, na prova de tolerância à glicose oral (PTGO) com 75g de glicose; ou
- Hemoglobina glicada A1c (HbA1c) ≥ 6,5%.

Um indivíduo que tenha pacientes diabéticos na família próxima, que tenha mais de 45 anos de idade, seja obesa, leve uma vida sedentária, apresente tensão arterial elevada ou níveis elevados de colesterol no sangue tem maior probabilidade de vir a desenvolver Diabetes. Dietas altamente calóricas, com muitos hidratos de carbono e muita gordura, estão associadas com uma diminuição de recetores de insulina enquanto dietas com fibra aparentam aumentar esses mesmos recetores de insulina (Kaplan, Lippe, Brinkman, Davidson, & Geffner, 1982).

Apesar dos esforços humanos e tecnológicos no controlo dos níveis de glicemia, a Diabetes ainda é uma doença crónica incurável. A insulina foi descoberta em 1922 e trouxe um prolongamento da vida aos pacientes diabéticos, mas não trouxe a cura. A terapêutica é necessária até que a

transplantação pancreática seja possível nos casos mais graves (Kaplan, Lippe, Brinkman, Davidson, & Geffner, 1982).

2.2. Hipertensão Arterial

A Hipertensão Arterial, vulgo Hipertensão, é causada pelo excesso de pressão nas artérias exercida pelo fluxo sanguíneo e que obriga o coração a maior esforço para realizar a normal circulação do sangue. A Hipertensão é considerada um dos fatores de risco mais significativo das doenças cardiovasculares. As doenças cardiovasculares como o Enfarte do Miocárdio e o Acidente Vascular Cerebral são das mais importantes causas de morbilidade e mortalidade em Portugal e no mundo (Macedo, et al., 2007), sendo o último a principal causa de morte em Portugal (Polónia, Ramalhinho, Martins, & Saavedra, 2006).

A tensão arterial é medida de duas formas: (1) A tensão arterial sistólica (TAS) refere-se ao momento em que o músculo cardíaco se encontra em contração e, ao invés, (2) a tensão arterial diastólica (TAD) refere-se ao momento em que este se escontra em descontração. A Hipertensão diastólica isolada é mais comum em pacientes mais novos, abaixo dos 40 anos de idade, enquanto a Hipertensão sistólica é mais comum em pacientes mais idosos (Ng, Stanley, & Williams, 2010). A causa principal do aumento da TAS com a idade deve-se a um aumento na rigidez das maiores artérias causadas por fatores associados que incluem a obesidade e sedentarismo e não existe diferença entre género sexual.

A Hipertensão está definida para valores de TAS ≥ 140 mm Hg e/ou de TAD ≥ 90 mm Hg. É ainda considerado como paciente hipertenso aquele que, mesmo não apresentando os valores acima dos referenciados, se encontrar a fazer terapêutica com medicação hipertensiva (Macedo, et al., 2007). Para pacientes de Diabetes ou doenças renais crónicas, os valores de referência baixam para TAS ≥ 130 mm Hg e/ou de TAD ≥ 80 mm Hg (Carey, 2013).

Para medir a tensão arterial, o individuo deve estar sentado e em repouso durante pelo menos 5 minutos. A medição é realizada usando um esfigmomanómetro apropriado ao adulto ou criança e devem ser recolhidas três medições em cada braço, intervaladas com pelo menos um minuto, sendo o valor resultante calculado pela média aritmética das 3 leituras (Ng, Stanley, & Williams, 2010).

A leitura da hipertensão é também realizada em ambulatório, durante 24 horas, através de um monitor eletrónico ligado em permanência ao paciente e que faz leituras em intervalos pré-selecionados, geralmente de hora a hora durante o período de atividade e de duas em duas horas durante o período de descanso (Ng, Stanley, & Williams, 2010).

Existem casos em que a medição deve ser realizada de pé como para indivíduos mais idosos, doentes diabéticos, a tomar múltipla medicação anti-hipertensiva ou ainda em pacientes conhecidos por ter uma história de hipertensão relacionada com a postura. Para pacientes com irregularidades ao nível do batimento cardíaco, também não é recomendada a medição da Hipertensão com medidores eletrónicos sendo a regra de ouro ainda o uso do esfigmomanómetro manual (Ng, Stanley, & Williams, 2010).

De acordo com (Polónia, Ramalhinho, Martins, & Saavedra, 2006) existem dados que sugerem que a prevalência e gravidade da Hipertensão é mais elevada para indivíduos de raça negra expondo-os a riscos acrescidos, comparativamente aos caucasianos.

A Hipertensão é caraterizada em dois tipos:

- Primária;
- Secundária.

A Hipertensão primária, ou essencial, é aquela em não foi possível identificar a sua causa, mas sabe-se que é causada por uma combinação de fatores genéticos e ambientais. De acordo com (Ng, Stanley, & Williams, 2010) os pacientes deste tipo de Hipertensão têm muito comummente histórias familiares ligadas à doença. Os fatores ambientais são conhecidos pelos estudos feitos às migrações onde é demonstrada a importância de apropriados estilos de vida com prática de exercício físico regular, controlo da dieta alimentar com pouca ingestão de sal e rigoroso controlo no consumo de álcool e hábitos tabágicos.

A Hipertensão secundária é aquela em que a causa é conhecida. Notavelmente, as doenças renais são responsáveis por 50% dos casos (Ng, Stanley, & Williams, 2010).

A identificação do paciente hipertenso é difícil visto que este é geralmente assintomático. A identificação é geralmente começada durante consultas de rotina ou devido a descobertas incidentais. Contudo, o diagnóstico envolve ainda uma avaliação dos riscos cardiovasculares e identificação de potenciais causas secundárias (Ng, Stanley, & Williams, 2010).

Existe ainda a categorização da hipertensão nas categorias seguintes:

- Normal, para TAS [120 – 129] mm Hg e TAD [80 – 84] mm Hg;
- Normal alto, para TAS [130 – 139] mm Hg ou TAD [85 – 89] mm Hg;
- Hipertensão Estado 1, para TAS [140 – 159] mm Hg ou TAD [90 – 99] mm Hg;
- Hipertensão Estado 2, para TAS $\geq$ 160 mm Hg ou TAD $\geq$ 100 mm Hg.

Esta categorização tem o objetivo de homogeneizar grupos em avaliação, para hierarquizar em estratos esse risco e determinação terapêutica nesses grupos, mas estes valores devem, no entanto, ser considerados flexíveis dependendo do perfil de risco cardiovascular global de cada indivíduo (Polónia, Ramalhinho, Martins, & Saavedra, 2006).

A prevalência da Hipertensão na população mundial é de 20-30%, aumentando com a idade e em determinados grupos étnicos. A taxa de prevalência mundial para Africanos/ Caribenhos é de 33% (Ng, Stanley, & Williams, 2010). A distribuição da prevalência por tipos é de 90% para Hipertensão Primária e os restantes 10% para Hipertensão Secundária (Ng, Stanley, & Williams, 2010).

Em Portugal existem alguns estudos parcelares e regionais sobre a Diabetes e escassos são os de âmbito nacional. Uma dessas exceções é o estudo realizado entre 2003 e 2004 por (Macedo, et al., 2007) e da (Sociedade Portuguesa de Hipertensão, 2012) onde se verifica que as taxas obtidas para Portugal são semelhantes aos encontrados em outros países europeus.

Em (Macedo, et al., 2007) pode-se observar que existiam em Portugal 3.311.830 indivíduos hipertensos com idade compreendida entre 18 e 90 anos, a que corresponde uma taxa de prevalência de 42,1% da população adulta portuguesa. A taxa de prevalência masculina 49,5% distribui-se por 26,2% para pacientes com idade inferior a 35 anos, 54,7% para idades superiores a 35 anos, mas inferiores a 64 anos, e 79% para maiores de 64. A taxa de prevalência feminina 38,9% era de 12,4%, 41,1% e 78,7% de acordo com os mesmos escalões etários. É significativa a diferença entre a taxa de prevalência para idades mais novas e observa-se uma aproximação das taxas à medida que a idade vai aumentando.

Quadro 2:2 - Factos e Números sobre a Hipertensão em Portugal – distribuição etária e por género sexual (Macedo, et al., 2007)

Escalão etário	Homens	Mulheres	Geral
<35 Anos	26,2%	12,4%	3.311.830 Indivíduos
≥35 e <64 anos	54,7%	41,1%	
≥ 64 Anos	79%	78,7%	

Entre os hipertensos, somente 46,1% sabiam sê-lo, 39% tomavam medicação anti-hipertensiva e 11,2% estavam controlados, ou seja, com a tensão estabelecida em parâmetros para si considerados normais. No estudo da (Sociedade Portuguesa de Hipertensão, 2012) pode observar-se que estes valores estão significativamente melhorados. Em 2012 a taxa de hipertensos que sabiam sê-lo era de 76,8% e destes 74.9% eram tratados; 42,6% dos pacientes tratados tinham a hipertensão controlada.

O estudo de (Macedo, et al., 2007) revela distribuições regionais, onde se pode observar que o Centro se encontra em linha com a média nacional para a taxa de prevalência da Hipertensão. Relativamente à taxa de prevalência conhecida, ou seja, o número de hipertensos que o sabem sê-lo, a região centro apresenta a menor taxa a nível nacional com 40,8%; a taxa nacional é de 46,1%. Sem surpresa, a taxa de prevalência tratada, ou seja, o número de hipertensos que tomam medicação anti-hipertensiva, revela que na região centro é também a menor taxa 34,3% a nível nacional 39%. Este resultado não surpreende, pois, a medicação sucede o diagnóstico.

O estudo da (Sociedade Portuguesa de Hipertensão, 2012) incide ainda sobre o consumo de cloreto de sódio (sal) avaliados em amostras de excreção de urina de 24 horas. Cada indivíduo da população portuguesa consome em média 10,7 g de cloreto de sódio por dia sendo que para hipertensos este valor é de 11,0 g e para normotensos de 10,5 g. A comparação com outros países europeus deixa Portugal muito mal visto; Espanha e Itália consomem cerca de 10 g e este valor decresce até ao mínimo de 7,9 g na Dinamarca.

Cerca de 30% dos pacientes hipertensos têm uma forma de Hipertensão chamada de “colarinho branco” que é causada por stress profissional; nestes casos é essencial para diagnóstico a medição através de monitores eletrónicos em regime de ambulatório (Ng, Stanley, & Williams, 2010).

O termo Hipertensão Resistente é usado para indicar os casos em que a condição persiste após a terapêutica anti-hipertensiva (Carey, 2013) apesar de não ser consensual se importa o número de medicamentos tomados. Após a análise de vários estudos, (Carey, 2013) conclui que a

prevalência da Hipertensão Resistente se situa cerca dos 14%. Um desses estudos foi feito em Espanha com 68.000 pacientes onde se verificou que 14,5% padeciam desta condição. Em Portugal a taxa de prevalência de Hipertensão Resistente reportada pela (Sociedade Portuguesa de Hipertensão, 2012) é de 8%.

O tratamento da Hipertensão visa a redução da tensão arterial a qualquer custo e a maioria dos pacientes necessitará de tomar 2 ou mais medicamentos (Ng, Stanley, & Williams, 2010). A terapêutica recomendada deverá ter em linha de conta a vida do paciente, mas também que este está devidamente educado quanto às consequências dos seus atos, como por exemplo a não aderência à mesma, quais os seus direitos e responsabilidades. Deve ser considerado o risco tendo em conta caraterísticas étnicas, socioeconómicas e culturais (Polónia, Ramalhinho, Martins, & Saavedra, 2006).

O prognóstico da doença depende de fatores simples como aderência à terapêutica recomendada, mas também de adoção de estilos de vida e ambientais adequados.

Segundo a OMS, está previsto um agravamento significativo da taxa de prevalência da Hipertensão na população mundial até 2025. Em Portugal, a conclusão dos autores do estudo (Macedo, et al., 2007) aponta para a "necessidade de desenvolvimento de estratégias nacionais para melhorar a prevenção, a deteção e tratamento" da doença no nosso país e este estudo vai ao encontro desse objetivo. Estes autores revelam ainda no seu estudo que existe uma taxa muito significativa de hipertensos muitos jovens que não sabem que padecem desse problema e como tal não estão controlados. As TI podem ser determinantes visto que são muito bem aceites por populações mais jovens.

2.3. A relação Diabetes e Hipertensão

Existe um elevado grau de relação entre a Diabetes e a Hipertensão; a sua coexistência é muito frequente. De acordo com (Sowers, Epstein, & Frohlich, 2001) os pacientes diabéticos hipertensos são duas vezes mais frequentes que os pacientes diabéticos normotensos, os pacientes hipertensos têm maior predisposição de vir a padecer também de Diabetes do que os pacientes normotensos e cerca de ¾ das doenças cardiovasculares com Diabetes são causadas por Hipertensão.

Os regimes terapêuticos prescritos para doentes hipertensos reduzem a incidência de sintomas associados com a Diabetes, como o atraso de progressão da nefropatia diabética, entre outros (Polónia, Ramalhinho, Martins, & Saavedra, 2006).

3. DATA MINING

3.1. Introdução

Os dados de uma organização são o que constitui a sua memória, a sua identidade. Sem os dados, as organizações não funcionam, pois, as ações tomadas necessitam de fundamentação que só é possível havendo factos criados a partir dos dados existentes.

As organizações começaram a adoção em massa de TI durante os anos 80 do século passado. Começaram por informatizar os processos de negócios financeiros e de gestão de recursos humanos. Por causa do sucesso alcançado, o âmbito foi alargado a toda a organização e a realidade atual é que qualquer organização tem hoje todos os seus processos de negócio cobertos ou assistidos por TI.

O sucesso na adoção de TI trouxe benefícios muito significativos às organizações, mas não sem custos. As organizações atuais possuem hoje demasiados dados para que uma análise para tomada de decisão possa ser realizada de forma manual.

Os dados sobre o negócio são armazenados em bases de dados de Sistemas Operacionais ou em repositórios de dados especializados como Data Warehouses. Um Data Warehouse vai mais além do que um repositório de dados, permite que os utilizadores acedam aos dados para a realização de análises mais orientadas (Young, 2000). A sua exploração permite responder a questões mais direcionadas para a gestão operacional e tática, ou seja, a gestão do quotidiano ou ações com efeitos esperados a curto e médio prazo. O Data Mining permite a exploração dos dados para descoberta de conhecimento mais usado nas decisões de gestão estratégica que afetarão a organização a longo prazo pois modificam modelos de negócio necessários a um reposicionamento da oferta da organização no mercado.

3.2. História

O Data Mining pode ser considerado como tecnologia e metodologia recentemente desenvolvida, ganhando proeminência apenas em 1994 (Trybula, 1997) e (Koh & Tan, 2005). A 1ª conferência em Descoberta de Conhecimento e Data Mining (SIGKDD) foi realizada pela ACM nos Estados Unidos em 1995 (Yoo, et al., 2012).

3.3. Definição

Não existe uma definição universal para o termo Data Mining. Uma revisão da literatura permite encontrar várias definições que se descrevem de seguida. Apesar de equivalentes no significado, é possível encontrar a definição vista de várias perspetivas o que explica a universalidade da aplicação desta.

Data Mining tem como objetivo identificar correlações e padrões em dados, que sejam válidos, novos, potencialmente úteis e compreensíveis (Chung & Gray, 1999), através de técnicas que

percorrem conjuntos de dados muitas vezes grandes para encontrar padrões que são demasiado subtis ou complexos para serem detetados por humanos (Kreuze, 2001).

Data Mining é um processo exploratório de descoberta de conhecimento orientado pelos dados em que o foco está na descoberta e extração de padrões de informação úteis a partir de bases de dados grandes e complexas (Berry & Linoff, 2000) que acrescenta ainda a importância da informação ser atempada.

Data Mining pode ser definida como um processo de seleção e exploração de dados armazenados em vastos repositórios de dados e o uso dessa informação para a construção de modelos preditivos com vista à descoberta de padrões e tendências previamente desconhecidas em bases de dados (Koh & Tan, 2005) (Khajehei & Etemady, 2010).

O Data Mining tem como objetivo identificar e validar correlações e padrões em dados, potencialmente úteis e previamente desconhecidos, percorrendo conjuntos de dados (*Datasets*) para descobrir padrões demasiado subtis ou complexos de serem detetados pelo ser humano (Koh & Tan, 2005) sem recorrer a essas mesmas ferramentas.

Data Mining é a análise de conjuntos de dados observacionais (muitas vezes grandes) para encontrar relacionamentos insuspeitos e sumariar os dados de novas maneiras que são ambas compreensíveis e úteis para o proprietário dos dados (Hand, et al. 2001). Data Mining permite a criação de hipóteses científicas de grandes conjuntos de dados experimentais e de literatura biomédica (Yoo, et al., 2012).

Data Mining amadureceu para uma forma de lidar com a crescente disponibilidade de dados digitais e da diferença entre a disponibilidade de dados e do uso do conhecimento derivado destes (Fayyad, et al. 1996) e (Berger & Berger, 2004).

Reunindo as partes, é possível definir Data Mining como a extração de conhecimento interessante (não-trivial, implícito, previamente desconhecido e de utilidade potencial) que através de algoritmos específicos descobre padrões, tendências nos dados e mecanismos de regras (associações entre dados aparentemente dissociados).

3.4. Metodologias

O termo Data Mining é muitas vezes confundido com o processo de Descoberta de Conhecimento em Bases de Dados (KDD[1]), a metodologia CRISP-DM e a organização lógica SEMMA. Como se explica em seguida, os termos não são inteiramente sinónimos, sendo que Data Mining é somente uma parte das etapas em todo o processo.

O processo de Descoberta de Conhecimento em Bases de Dados descrito por (Fayyad, et al. 1996) é iterativo e sequencial e inclui as seguintes etapas:

[1] Sigla KDD provém do termo em inglês "Knowlegde Discovery in Data"

1. Seleção dos dados – Após a compreensão do domínio de aplicação os dados são selecionados e recolhidos;
2. Pré-processamento dos dados – Correção dos dados, de lacunas ou inconsistências;
3. Transformação dos dados – Conversão de dados de um formato requerido pelo armazenamento eletrónico para o formato requerido para a sua interpretação por algoritmos de Data Mining;
4. Data Mining – Aplicação de algoritmos que produzem modelos matemáticos usados na inferência de descoberta de conhecimento e padrões nos dados;
5. Interpretação e avaliação dos resultados – Aplicação de técnicas de visualização aos modelos e resultados do Data Mining para compreensão por parte dos utilizadores do domínio de aplicação.

Os investigadores (Khajehei & Etemady, 2010) sugerem que as etapas 1 e 2 são a mesma, porém os próprios chamam a etapa de "Seleção e Limpeza de Dados". Logo, a limpeza de dados só faz sentido após a seleção, mesma que ocorra registo a registo. Apesar da existência de um percurso linear, nada impede a necessidade de retorno a etapas anteriores se identificada essa necessidade.

Esta metodologia não surgiu dos meios académicos, mas sim de um consórcio de empresas que a desenharam para suprir as suas necessidades e viram a sua potencialidade comercial para fornecer orientação em projetos de Data Mining. A metodologia CRISP-DM não é um livro de instruções para iniciantes, a sua correta aplicação é baseada no conhecimento e experiência dos métodos e técnicas de Data Mining (Shearer, 2000). A metodologia de realização de projetos de Data Mining CRISP-DM inclui diversas fases que são as seguintes:

1. Compreensão do negócio – Centra-se na análise do domínio de aplicação para compreensão dos objetivos do projeto;
2. Compreensão dos dados – É feita a recolha dos dados e uma análise qualitativa dos mesmos;
3. Preparação dos dados – Correção dos dados de omissões, erros ou inconsistências e transformados de acordo com o requerido na modelação a realizar em seguida;
4. Modelação – São empregues as técnicas de modelação de Data Mining apropriadas e obtidos os modelos;
5. Avaliação – Os modelos são revistos e validados em relação aos objetivos;
6. Operacionalização – Apresentação dos resultados aos utilizadores finais e colocação em uso para alcance dos objetivos.

SEMMA não é bem uma metodologia de Data Mining, mas sim uma organização lógica em torno das ferramentas do *software* (SAS Enterprise Miner, 2013) para efetuar as tarefas centrais de Data Mining. Apesar disso muitos são os que usam a SEMMA como metodologia visto esta aproximar-se muito a uma metodologia completa. As fases do processo definido pelo SEMMA são as seguintes:

1. Amostragem – Extração de um conjunto de dados significativo e representativo do universo em análise;
2. Exploração – Os dados são visualizados geralmente através de técnicas gráficas para a sua melhor compreensão;

3. Modificação – Criação, seleção e transformação dos dados para se dar início à modelação;
4. Modelação – Seleção e aplicação de técnicas de Data Mining para obtenção dos modelos;
5. Avaliação – Visto que a metodologia se baseia numa amostragem, a avaliação requer a tentativa de uso tanto para a amostra como, eventualmente, para os restantes dados.

Apesar da semelhança dos processos acima mencionados ser notória, fica claro que Data Mining é somente uma etapa em todo o processo de Descoberta de Conhecimento em Dados. Data Mining inclui a aplicação de algoritmos específicos para extração de padrões ou desvios de dados previamente preparados, mas exclui as etapas de compreensão dos dados e do negócio bem como a operacionalização de resultados.

3.5. Estatística

Data Mining é uma área multidisciplinar com fundações que intersecta os campos da Estatística e Aprendizagem Automática e que se expandiu para incluir métodos e técnicas (algoritmos) oriundas de Inteligência Artificial, Reconhecimento de padrões, Bases de dados, Data Warehousing, Matemática, Sistemas Periciais e Visualização de Dados (Yoo, et al., 2012).

O Data Mining recorre ao uso extensivo de métodos estatísticos, mas é diferente da Estatística em muitos aspetos. Conforme (Seifert, 2004), "Data Mining representa uma diferença [da Estatística] em tipo ao invés de grau".

Em estatística verificam-se hipóteses ou mede-se a quantidade exata de uma relação entre duas ou mais variáveis enquanto em Data Mining o objetivo é salientar e descrever padrões gerais e regras de diferenças que requerem muitas vezes apenas resultados gerais para indicar que existem essas mesmas relações (Castellani & Castellani, 2003). Na estatística é fundamental saber o valor exato enquanto em Data Mining o valor é apenas indicativo para conhecimento que a relação existe.

Embora ambos os domínios se baseiem em rigor na Matemática, a Estatística prefere a adoção de abordagens exatas enquanto o Data Mining adota parcialmente heurísticas (Yoo, et al., 2012). Muitas vezes o objetivo é detetar uma tendência ou padrão e não quantificar essa relação com precisão.

Por razões históricas motivadas pela dificuldade de cálculo de grandes conjuntos de dados, a Estatística é baseada no cálculo de amostras significativas recolhidas da população. O Data Mining usa também extensivamente a técnica de amostragem, mas em determinados casos devem ser usados todos os dados da população como a aglomeração de dados ou descoberta de padrões escondidos para obtenção do melhor resultado (Yoo, et al., 2012).

Outra diferença entre Estatística e Data Mining prende-se com o tipo de dados que são trabalhados; a Estatística trabalha dados numéricos e o Data Mining é capaz de lidar com outros

tipos de dados como imagens, sons, vídeos e texto, bem como variáveis categóricas (Yoo, et al., 2012).

Como é prática comum nas ciências modernas, baseadas no método científico, a teoria começa numa hipótese baseada em observações e são *a posteriori* coletados dados que depois de testados corroboram (ou não) a teoria científica. A Estatística faz uso deste método; sendo uma disciplina hipotético-dedutiva, analisa os dados com base na hipótese previamente definida. Podemos afirmar que avançamos do geral para o específico. Em Data Mining acontece muitas vezes o oposto, avançamos de um conjunto de dados (do geral) para a descoberta de regras, padrões, desvios (o específico) (Yoo, et al., 2012).

3.6. Tarefas

Realizar Data Mining é extrair conhecimento de bases de dados, mas esta realização pode assumir diversas formas que estão classificadas em tarefas. As tarefas de Data Mining estão relacionadas com os objetivos que se pretendem alcançar. Existem tarefas de Classificação, Estimação, Segmentação (*clustering*), Modelação de dependências (Associação e Sequenciação, incluindo temporal), Sumarização, Descrição e Visualização e Deteção de desvios (*outliers*) (Santos & Ramos, 2009) (Santos & Azevedo, 2005) (Koh & Tan, 2005) (Shearer, 2000).

A realização de tarefas de Data Mining implica o uso de técnicas que se classificam em aprendizagem descritiva ou preditiva (Yoo, et al., 2012): (1) A aprendizagem descritiva é exploratória por natureza e permite segmentar os dados através de análises à sua semelhança ou descoberta de padrões e relacionamentos para que os utilizadores compreendam os dados que são, regularmente, de elevada dimensão. (2) A aprendizagem preditiva infere regras ou valores através da construção de modelos que têm por base dados de treino e que são aplicados a dados de teste para avaliação do valor do modelo construído.

3.6.1. Classificação

O objetivo da tarefa de classificação é predizer a classe de novos casos, ou seja, a qualificação de casos a determinadas classes pré-estabelecidas de acordo com os seus atributos. Em termos estatísticos, a classe é a variável dependente definida em função dos atributos que são a variável independente.

Em termos de aplicação a classificação é usada, por exemplo, pela banca na atribuição de empréstimos (classe) tendo em conta os rendimentos, bens e garantias do proponente (atributos) ou em saúde na definição de diagnóstico e prognóstico tendo por base os sintomas, historial e condições ambientais do paciente.

A classificação é uma tarefa preditiva de aprendizagem supervisionada (*supervised learning*) que significa que os atributos e classes que vão conduzir ao processo de determinação da classe são conhecidos à partida. É normal o algoritmo dividir os exemplos conhecidos em dois: Um dos conjuntos é usado para o treino (construção) do modelo e é por isso chamado de conjunto de dados de treino (Trainingset), o outro conjunto é usado para a sua avaliação e é chamado de conjunto de dados de teste (Testset). O algoritmo de classificação usa os casos do conjunto de

treino e constrói um modelo matemático que, depois de treinado, deverá ser capaz de classificar objetos descritos por valores de atributos que tenham ou não ocorrido nos exemplos de treino. Para avaliação da qualidade do modelo são usados os exemplos do conjunto de teste por comparação da classe predita pelo modelo e a classe real.

3.6.2. Regressão

O objetivo da regressão é estimar o valor de uma variável contínua tendo por base a tendência dos valores observados. Tal como a classificação, também a regressão envolve modelação com aprendizagem preditiva. A regressão designa-se multivariada quando aplicada a várias dimensões.

Esta técnica é usada no domínio da saúde, por exemplo, para determinar a evolução do custo de prestação dos serviços em função da idade dos utentes.

3.6.3. Clustering

Em Clustering (segmentação, aglomeração) o objetivo é agrupar elementos que partilhem um determinado grau de semelhança. Tomando em consideração todos os elementos de um conjunto de dados, pretende-se formar grupos de elementos (clusters) em que os atributos sejam os mais semelhantes possíveis aos outros elementos do mesmo grupo e o mais desigual dos elementos constituintes de outros grupos.

O Clustering é uma tarefa descritiva de aprendizagem não supervisionada (*unsupervised learning*) pois apenas são tratados os dados das variáveis independentes; visto que não se destina a classificar os casos, a classe (variável dependente) não é relevante. O Clustering reveste-se de um estudo essencialmente exploratório dos dados com vista à sua descrição e compreensão pois os dados trabalhados são normalmente muito grandes e sobre os quais existe um desconhecimento profundo.

Em termos de aplicação o Clustering é utilizado, por exemplo, na banca para segmentação de clientes tendo em conta os seus perfis socioeconómicos, na biologia para determinação de taxonomias tendo em conta as caraterísticas físicas ou na saúde na determinação de grupos de risco tendo em conta fatores bioquímicos, físicos, socioeconómicos, demográficos, ambientais entre outros.

3.6.4. Descrição e Visualização

A tarefa de Descrição e Visualização trata de interpretar e analisar os dados de forma a serem compreendidos. Seja de forma textual ou através de gráficos, o importante é mostrar o que os dados representam numa forma que um humano e não um computador seja capaz de interpretar. Os conjuntos de dados podem ser grandes e pela simples observação é impossível a sua interpretação ou deteção de padrões escondidos especialmente em dados complicados que contêm interações complexas e não-lineares (Koh & Tan, 2005).

3.6.5. Modelação de Dependências (Associação)

O pretendido pela tarefa de Modelação de Dependências (Santos & Ramos, 2009) ou Associação ou Dependência (Santos & Azevedo, 2005) é descobrir padrões de associação ou sequenciação nos dados. Formalmente é encontrar um modelo que descreva dependências significativas entre variáveis. As associações são descobertas através de análise de correlação das variáveis, representadas quer através de modelos gráficos quer através de regras calculadas com medidas numéricas chamadas de suporte e confiança.

A associação consegue determinar, por exemplo, que face a determinados comportamentos de risco o utente está mais predisposto a contrair outros pelo facto de esta regra acontecer numa quantidade significativa de outros utentes. A sequenciação está normalmente associada a um fator temporal, mas também ocorre pela sua ordenação nos conjuntos de dados.

A sequenciação modela acontecimentos como o surgimento de uma ocorrência de um sintoma de uma doença dentro de um determinado intervalo de tempo ou que depois de acontecer este sintoma seguir-se-á outro tal como aconteceu noutros pacientes.

3.6.6. Deteção de Desvios (Outliers)

A deteção de desvios (anomalias, outliers) tem como objetivo determinar os casos de um conjunto que são díspares da grande maioria, a níveis suspeitos de serem gerados por mecanismos diferentes dos demais.

Inicialmente considerados como ruído nos dados e perturbadores de aplicação das técnicas de Data Mining, eles são atualmente considerados como um dado valioso pois permitem, por exemplo, realizar a deteção de fraudes na banca ou o surto de uma doença.

3.7. Técnicas

As tarefas de Data Mining são concretizadas através de técnicas que visam a obtenção de modelos. Os modelos, obtidos pelas tarefas de Data Mining, são representações matemáticas que visam a compreensão e o estudo dos dados.

As técnicas de Data Mining são concretizadas através de algoritmos; cada ferramenta de Data Mining implementa uma versão desses algoritmos. As técnicas de Aprendizagem Automática (*Machine Learning*) mais usadas em Data Mining são: Árvores de Decisão (AD), Regras de Associação e Sequenciação Temporal, Regressão, Redes Neuronais Artificiais (RNA), Algoritmos Genéticos (AG), Clustering, Classificadores Bayesianos e *Support Vector Machines* (SVM).

Generalizando a questão para um caso prático, supondo que estamos interessados em analisar a qualidade de tratamento de saúde a determinados pacientes, começamos por juntar dados pertinentes com o nosso objetivo, mas que de início poderão conter erros devidos a má interpretação, má entrada de dados, diferenças culturais, entre outros fatores. Começa-se por atravessar as fases de refinamento de dados, colmatando ou eliminando dados erróneos, incompletos ou em falta, e aplicam-se técnicas de exploração de dados de onde são obtidos

resultados. Depois de analisados os resultados, é possível voltar atrás e repetir o processo. É muitas vezes através deste processo por iterações que se obtém o resultado final.

A aplicação dos algoritmos aumenta o conhecimento qualitativo extraído da quantidade de dados; potencia o processo qualitativo permitindo aos investigadores analisar até grandes quantidades de dados. Os conhecimentos resultantes obtidos pela análise quantitativa necessitam de fundamentação em métodos qualitativos de análise e conceção de sistemas, como entrevistas ou etnografia, para entender os resultados dentro do respetivo domínio de aplicação.

Existem diversas técnicas de aplicação de Data Mining possíveis a que correspondem muitos algoritmos e variantes destes. Consoante o objetivo assim é a escolha do algoritmo que será usado na extração de conhecimento a partir dos dados existentes. Todos os algoritmos são construídos para tarefas específicas e todos têm aspetos positivos e negativos a considerar (Castellani & Castellani, 2003).

3.7.1. Árvores de Decisão

A árvore de decisão é uma representação matemática de um modelo com regras estruturadas numa hierarquia de classes ou valores (Santos & Azevedo, 2005) aplicáveis a tarefas de classificação ou regressão, consoante se trate de classes discretas ou contínuas, respetivamente. A representação gráfica assemelha-se a uma árvore natural sendo a representação mais comum invertida pois o topo simboliza a raíz e na base encontramos as folhas.

A estrutura da árvore de decisão é composta por três elementos: (1) Nós internos, (2) folhas e (3) ramos. A raíz da árvore é um nó interno. Os nós internos correspondem à avaliação de um atributo. Dos nós internos saem dois ou mais ramos que correspondem às possibilidades de resultados da avaliação do nó interno. Quando essa avaliação pode resultar num só resultado, é representado um nó terminal puro chamado de folha a que corresponde a classe.

O processo de indução é realizado por um algoritmo, e.g. o mais referido pela literatura científica *C4.5* (Quinlan, 1993) ou o *Ant-Tree-Miner* (Otero, Freitas, & Johnson, 2012) inspirado no funcionamento de uma colónia de formigas. O algoritmo *greedy* constrói a árvore de cima para baixo, ou seja, da raíz para as folhas, tendo por base um conjunto de dados de treino devidamente classificados. Este facto quer dizer que a indução da árvore de decisão é uma aprendizagem supervisionada. O processo de seleção do atributo a colocar no nó interno é chamado de Método de Seleção do Atributo (Yoo, et al., 2012) e é o fator mais importante para a performance do modelo depois da qualidade dos dados do treino. Existem várias técnicas de seleção do atributo, mas todas têm como objetivo encontrar aquele cujos valores melhor contribuem para dividir os registos de acordo com as classes. Por outras palavras, tentam encontrar o atributo em que a variabilidade dos valores mais se assemelhe à variabilidade das classes. Os critérios de seleção dos atributos são usados para calcular os mais significativos, os que mais contribuem para definir a classe. Este processo continua tendo por base o novo nó, dividindo o conjunto de dados até que seja encontrada a classe predominante que dá origem à folha. Se o conjunto de dados possui muitos atributos e classes possíveis a árvore pode ficar demasiado complexa e prejudicar a performance do modelo.

Quando terminado o processo de indução, a árvore encontra-se ajustada aos dados oferecendo uma estrutura demasiado rígida para classificar exemplos não treinados ou simplesmente porque a estrutura contém mais do que o necessário. Para resolver estes problemas é necessário fazer a poda (*prunning)* da árvore que é, no fundo, uma técnica que implica a eliminação de nós internos que não contribuem de forma decisiva para o resultado final. A poda pode ser feita durante ou após a aprendizagem do modelo e baseia-se em cálculos de previsão de uma taxa de erro. A poda resulta em árvores menores com melhor potencial e precisão (Quinlan, 2001 e Han & Kamber, 2001). Para realizar a poda de forma correta e simplificar estruturas complexas podem ser necessários muitos dados bem como a seleção dos mais apropriados.

As árvores de decisão são de construção simples e rápida, de representação compreensível pelos utilizadores e onde estes podem visualizar claramente os atributos que mais contribuem para a definição da classe.

3.7.2. Regras de Classificação

As regras de classificação são proposições do tipo "Se {(atributos, valores)} -> Então {classe}" que estão popularizadas na formulação de hipóteses. Esta técnica tem uma performance comparável com as árvores de decisão e são de fácil visualização e interpretação.

Podem ser induzidas através de uma abordagem indireta, sendo extraídas, por exemplo, a partir das Árvores de Decisão, ou por abordagem direta, pela avaliação de cada classe separadamente, tentando verificar se existem valores predominantes para os atributos em subgrupos de instâncias vizinham que partilham a mesma classe. Fazem parte desta abordagem os algoritmos de referência *CN2* (Clark & Boswell, 1991) e *RIPPER* (Cohen, 1995).

Os algoritmos de indução direta de regras, do qual fazem parte os algoritmos de aprendizagem baseados em instâncias (IBL), têm uma abordagem sequencial; começam por gerar uma regra candidata, removem as instâncias afetadas por essa regra, se necessário é feita a poda da regra, a regra é adicionada ao conjunto de regras e o processo é repetido. A geração das regras pode ir do geral para o específico ou vice-versa, sendo que no primeiro caso o antecedente (pares de {atributo, valor}) vai sendo acrescentado e no segundo caso diminuído. Esta variação é feita consoante medidas que testam a melhoria da qualidade da acurácia (Equação 1) (o número de instâncias que confirmam a regra dentro do conjunto de instâncias com os mesmos atributos que confirmam ou não a regra) e da sua cobertura (o número de instâncias com os atributos que confirmam a regra dentro do conjunto total de instâncias). Estas duas medidas trabalham em conjunto pois uma regra pode ser de grande acurácia, mas pouca cobertura. Por exemplo, o *RIPPER* usa a medida *Foil's Information Gain* visto esta só ser elevada quando as regras tenham alto suporte e alta cobertura. A eliminação das instâncias afetadas pela regra serve o propósito de não se repetir a geração da mesma regra na iteração seguinte. Depois de gerada a regra esta passa pelo processo de poda idêntico às Árvores de Decisão. Consoante os atributos do caso a classificar, podem ser "disparadas" várias regras pelo que é determinante haver um critério de desempate, por exemplo a primeira regra a ser disparada. É por isso comum as regras serem ordenadas pelo antecedente ou consequente (a {classe}) ou usando medidas como as instâncias afetadas nessas regras.

3.7.3. Case-Based Reasoning

A técnica de *Case-Based Reasoning* (CBR) (Schank, 1982) pode ser vista como o uso de conhecimento de casos passados para classificação de novos casos. Um caso é constituído pela definição do problema, a sua solução e anotações relevantes. O modelo de aprendizagem requer a construção de uma base de conhecimento para decidir sobre usar os casos conhecidos ou criar um novo. O modelo tem quatro etapas:

1. Recuperação – onde são identificados os casos na base de conhecimento que são relevantes para o problema;
2. Reutilização – é desenvolvida a solução do problema com possível adaptação das soluções dos casos relevantes;
3. Revisão – aplicação da solução encontrada e, se necessário, eventual alteração para uso futuro;
4. Retenção – Salvaguarda da solução encontrada na base de conhecimento.

Esta técnica é simples de compreender pois mimetiza o funcionamento neurológico e a cognição usada no conhecimento humano. Como faz base no conhecimento empírico, é usada com frequência na saúde.

3.7.4. Redes Neuronais Artificiais

As Redes Neuronais Artificiais (RNA) são uma técnica computacional inspirada na neurobiologia em que o algoritmo imita o funcionamento das redes neuronais existentes no cérebro. Essencialmente estas redes possuem elementos que efetuam uma pequena parte do processamento requerido (nós) e outros que transportam a informação (as ligações) e são responsáveis pela interligação dos nós. O processamento vai sendo realizado à medida que a rede é percorrida combinando os resultados intermédios recebidos, processando mais um pouco e passando esse resultado ao nó seguinte até ao resultado final. Os nós da rede dividem-se por camadas: (1) A primeira camada (entrada/*input*) é responsável por receber os dados para a rede, (2) as ligações levam esses dados para camadas intermédias (escondidas/*hidden*) e (3) a última camada (saída/*output*) é responsável pela demonstração dos resultados.

Existem várias arquiteturas de RNA: (1) Redes totalmente conectadas em que cada nó está ligado diretamente a todos os outros, (2) Redes de camada única aplicáveis a problemas linearmente separáveis (não se considera a camada de entrada visto que esta serve apenas para receber os dados) e (3) Redes multicamada com uma ou mais camadas escondidas e camada de saída.

Os neurologistas sabem que o cérebro aprende alterando a resistência da ligação sináptica entre neurónios após estimulação repetida pelo mesmo impulso elétrico. Computacionalmente este comportamento de aprendizagem é realizado por um algoritmo que computa os valores dos nós e pesos das ligações na rede. Um nó (neurónio artificial) pode realizar qualquer tipo de processamento aos valores de entrada, porém o comum é a aplicação de uma função matemática chamada função de ativação, resultando dessa operação os valores de saída. As funções de ativação mais comuns são: degrau, sinal e sigmoide.

O algoritmo de aprendizagem define o tipo de RNA, sendo a mais utilizada do tipo percetrão multicamada com retropropagação porque a sua performance é considerada superior aos outros tipos (Yoo, et al., 2012) Um percetrão simples soma os valores de entrada e a este valor aplica a função degrau resultando numa de duas classes (1 ou 0), sendo por isso aplicável somente a problemas linearmente separáveis e capazes de representar uma reta no espaço de decisão.

Uma rede de retropropagação é uma rede multicamada e deriva do modelo percetrão. O treino modifica os fatores de ponderação chamados coeficientes sinápticos (pesos), ajustáveis, colocados entre nós, e que simulam a resistência da ligação sináptica. A retro propagação é o ajuste dos pesos realizado tendo por base o desvio entre o valor obtido pela RNA e o valor real dos exemplos classificados usados para o treino. A função de ativação é normalmente a sigmoide pois é capaz de representar arcos. A combinação de nós permite representar diversos espaços geométricos no espaço de decisão e ser aplicado a problema complexos não linearmente separáveis.

As RNA do tipo percetrão possuem a vantagem de ter uma construção simples, mas a desvantagem de apenas tratarem problemas linearmente separáveis. Outros tipos ultrapassam essa dificuldade, mas são mais complexos visto que adicionam parâmetros variáveis como o número de camadas e os nós para cada uma e não existe forma de calcular o valor ótimo dos parâmetros senão de forma empírica, tornando a rede muito sensível à parametrização pois o treino requer muita computação tornando-se por isso demorado. Devido às camadas escondidas e ao ajuste dos pesos, as RNA são de difícil compreensão e visualização para os utilizadores especialistas do domínio do problema; são vistas como caixas negras em que são conhecidos os valores de entrada e de saída, mas não o que levou a esses resultados.

Os algoritmos de aprendizagem baseiam-se no cálculo de um valor e avaliação do resultado. Se incorreto, ajustam o valor e testam de novo. Procedem assim até encontrar um valor que permita uma classificação correta. As RNA foram durante muito tempo vistas como o melhor algoritmo de classificação, mas a introdução das Árvores de Decisão e Support Vector Machines veio alterar esse paradigma (Yoo, et al., 2012).

3.7.5. Naïve Bayes

Um classificador *Naïve Bayes* é uma representação matemática de um modelo que classifica casos assumindo a incerteza existente na natureza e concretizada através de modelos estatísticos probabilísticos. Serve para classificar, ou seja, atribuir um valor de probabilidade, dados os valores dos atributos. Tem por base a fórmula de Bayes da probabilidade condicionada (teorema desenvolvido por Laplace com origem no trabalho de Bayes). É robusto relativamente a atributos irrelevantes e tuplos com ruído ou com valor desconhecido, dado que os valores erróneos não afetam de forma significativa o cálculo das probabilidades condicionais.

Os atributos correlacionados podem degradar significativamente a performance do modelo pois esta técnica assume que as variáveis têm independência condicional, e daí o nome de *naïve*. A independência condicional reduz o esforço computacional a uma multiplicação de probabilidades e necessita por isso poucos exemplos de treino. A rapidez do algoritmo perfaz a sua maior

vantagem e torna esta na técnica de classificação mais rápida (Yoo, et al., 2012). A simplicidade do algoritmo torna-o adequado para conjuntos de dados com muitos atributos. A desvantagem desta técnica é que os atributos estão normalmente correlacionados e esse facto degrada a performance da técnica a ponto de ser preterida por outras.

3.7.6. Support Vector Machine

Uma Support Vector Mahine (*SVM)* é uma representação matemática de um modelo de classificação tipicamente binária. Os casos são representados num espaço e o classificador calcula então uma fronteira (híper-plano) que separa ambas as classes de forma equitativa. Para classificar novos casos, basta verificar em que lado da fronteira este é representado. Apesar de inicialmente só ser possível uma classificação através de uma função linear, ou seja, a fronteira entre classes ser uma linha reta, desenvolvimentos mais recentes tornam possível a classificação não-linear através de funções polinomiais e de base radial.

A maior vantagem desta técnica encontra-se na capacidade de atingir bons resultados de classificação na generalidade dos casos de problemas de classes binárias: dos 21 testes realizados por (Meyer, Leisch, & Hornik, 2003) a SVM encontrava-se no pódio em 19. Este facto não deve ser confundido como sendo esta a melhor técnica de classificação pois o mesmo depende muito da estrutura dos conjuntos de dados.

A técnica SVM é muito dependente da função selecionada e do número de classes. Como os SVM são classificadores binários, os problemas que envolvam classes ternárias ou superiores são reduzidos a problemas de classes binárias tratados por associação de classificadores. Quando comparada com outras técnicas de classificação, a SVM tem um treino lento e requer significativos recursos computacionais (Meyer, Leisch, & Hornik, 2003).

3.7.7. Ensemble

A técnica de *Ensemble* é na prática uma associação de classificadores com o objetivo de minimizar a taxa de erro e assim permitir obter resultados melhores do que o faz uma técnica de forma individual. A técnica começa por construir diversos modelos classificadores, necessariamente diferentes uns dos outros. A diferença passa por ser usada uma das seguintes formas de manipulação:

- Do conjunto de dados de treino – Cada classificador usará um conjunto de dados de treino diferente, criados através de diferentes métodos de amostragem. As abordagens mais conhecidas são as seguintes:
  - *Bootstrap Aggregation (Bagging)* – São gerados vários conjuntos de dados de treino com base no conjunto de treino original através de amostragem aleatória com reposição. Os exemplos do conjunto de dados de teste são classificados por todos os modelos sendo a classificação final obtida por voto maioritário, ou seja, a classe predominante no resultado de todos os classificadores.
  - *Boosting* – gera um conjunto de dados de treino com base no anterior (e não no original) em que os casos mais difíceis de classificar vêem o seu peso acrescido para que surjam com maior frequência na recolha da amostra e que os mais

fáceis vêem o seu peso decrescido para que outros casos novos sejam escolhidos na sua vez. O método de *boosting* mais conhecido é o *Adaptive Boosting* (*Adaboost*) (Freund & Schapire, 1997) que usa um coeficiente de ponderação para cada classificador construído iterativamente calculado com base na taxa de erro e que é usado para determinação da classificação na votação final;

- Dos atributos – Os conjuntos de dados são criados através da seleção de atributos por métodos de análise da sua correlação, através de técnicas de segmentação, por recomendação de peritos ou mesmo de forma aleatória. A abordagem mais conhecida é a *Random Forests* (Breiman, 2001) que induz diversas Árvores de Decisão com base em conjuntos de treino gerados a partir do original que diferem nos atributos e que por isso geram diferentes modelos. O resultado é obtido por votação;
- Das classes – Os conjuntos de dados de treino são criados para que as classes concorram entre si. As abordagens da combinação entre classes múltiplas mais conhecidas são as seguintes:
  - *One against all (OAA, 1-R)* – Cada classe concorre contra um grupo constituído pelas demais existindo tantos classificadores como classes. A votação final é obtida por maioria;
  - *One against one (OAO, 1-1)* – Cada classe concorre somente com outra. São criados tantos classificadores quantas forem as combinações possíveis de pares de classes. Neste caso os exemplos de outras classes são ignorados na indução do modelo. A determinação da classe final é decidida por voto maioritário;
  - *Error-Correcting Output Coding (ECOC)* (Dietterich & Bakiri, 1995) – Cada classe é codificada por uma *codeword* binária. São construídos tantos classificadores quantos os bits desta *codeword*. Cada classificador prevê um dos bits da *codeword* e a classificação final é obtida através da comparação da distância de Hamming (Hamming, 1950) obtida e das *codewords* de cada classe, ganhando aquela em que a distância é menor, onde houver menos variação.
- Do algoritmo de treino – Neste caso o que muda não diz respeito aos dados, mas sim aos algoritmos de indução dos modelos que vêem os seus parâmetros alterados e que por esse motivo resultam em classificadores diferentes. A classificação obtém-se por voto maioritário.

A diferença nos modelos classificadores fará com que se obtenha um resultado individual com maior precisão e confiabilidade. Um caso notável é o *Adaboost* que atinge normalmente resultados superiores a uma SVM individual (Morra, et al., 2010) (Situ, et al., 2010) (Douglas, et al.,2010) (Lopes, et al. 2011). O estudo de (Douglas, et al.,2010) revela ainda que a *Random Forest* melhorou em 1% o resultado do *Adaboost*. Existe investigação que mostra ainda a possibilidade de criar uma combinação de combinações de classificadores (Ensemble de Ensembles) que poderão ainda melhorar os resultados, contudo os resultados tendem para um limite inultrapassável independentemente do número de combinações realizado (Ahn, et al., 2007).

3.7.8. Clustering

O clustering hierárquico é uma técnica de aglomeração que se baseia no princípio que os objetos mais próximos fazem parte do mesmo cluster (grupo) e objetos mais afastados fazem parte de outros clusters. A proximidade é determinada através de funções de cálculo de distância, como a de Minkowski (Euclidiana ou Manhattan), Chebyshev ou Mahalanobis. A técnica pode assumir dois tipos, que se distinguem pela indução do modelo: (1) Divisivo (abordagem de cima para baixo) que consiste em considerar todos os objetos num cluster e iterativamente dividir um cluster em dois, requerendo a decisão sobre o cluster a ser dividido e a forma de realização da divisão, ou a forma mais comum (2) Aglomerativo (abordagem de baixo para cima) que considera de início cada objeto como um cluster (*singleton*) e que de forma iterativa junta os dois clusters mais similares com base em medidas de distância (semelhança, proximidade) entre atributos.

Além da escolha da métrica da distância para avaliar a proximidade entre objetos, a utilização desta técnica pressupõe ainda a escolha do método de determinação dos clusters a aglomerar. As técnicas aglomerativas mais utilizadas podem ser categorizadas de acordo com o respetivo método de cálculo:

- *Single Linkage* ou *Nearest Neighbor* ou *Minimum Method* (Florek, et al. 1951, Sneath, 1957, Lance & Williams, 1967, Johnson, 1967 e McQuitty, 1967) e (Sneath & Sokal, 1973) – O algoritmo seleciona como representantes entre dois clusters o par de objetos que se encontre mais próximo e junta os clusters cujo par tenha a menor distância;
- *Complete Linkage* ou *Farthest Neighbor* ou *Maximum Method* (Lance & Williams, 1967 e Johnson, 1967) e (Sneath & Sokal, 1973) – É igual ao anterior, mas é escolhido o par de objetos que se encontre mais afastado;
- *Unweighted Pair-Group Method using Arithmetic Averages (UPGMA)* ou *Group Average* (Sokal & Michener, 1958) e (Sneath & Sokal, 1973) – O algoritmo seleciona todos os pares de objetos a partir de dois *clusters* e calcula a média de todas as distâncias possíveis entre objetos. Depois de calcular as distâncias médias entre os clusters selecionáveis, os dois grupos com a menor distância são combinados num só cluster;
- *Weighted Pair-Group Method using Arithmetic Averages (WPGMA)* (Sneath & Sokal, 1973) – É igual ao anterior, mas usa como fator de ponderação (peso) o número de objetos contido em cada cluster;
- *Unweighted Pair-Group Method using the centroid average (UPGMC)* ou *Centroid* (Sokal & Michener, 1958) e (Sneath & Sokal, 1973) – O algoritmo calcula o centróide (ponto que define o centro geométrico ou ponto de gravidade) de cada cluster selecionável e junta os clusters cuja distância entre centróides seja menor;
- *Weighted Pair-Group Method using the centroid average (WPGMC)* ou *Median* (Sneath & Sokal, 1973) – É igual ao anterior, mas usa como fator de ponderação (peso) o número de objetos contido em cada cluster;

- *Ward's Method* (Ward, 1963) – O algoritmo baseia-se na otimização de uma função objetivo que tem por base a soma total do erro quadrático (soma do quadrado da distância dos objetos ao centróide). Em cada passo o algoritmo junta os clusters em que resulte a menor variância intra-cluster resultando numa maior concentração de objetos junto do centróide.

Os algoritmos de clustering hierárquico têm geralmente uma computação de complexidade temporal cúbica $O(n^3)$ o que os torna desaconselhados para conjuntos de dados grandes. A representação do clustering hierárquico é vulgarmente realizada recorrendo à técnica gráfica de dendrograma que é de fácil visualização e compreensão pelos utilizadores. Através da observação da hierarquia do dendrograma é razoavelmente possível deduzir o número de clusters existentes no conjunto de dados, facto que separa esta técnica de clustering das demais.

O clustering partitivo é uma técnica em que o algoritmo aloca os objetos do conjunto de dados a um dos possíveis clusters. Todos os algoritmos desta técnica necessitam da parametrização do número de clusters a computar (normalmente designado pela letra *k*) antes da sua execução. Existem três tipos de clustering partitivo:

- Clustering baseado em centróides ou *k-means* clustering;
- Clustering baseado na distribuição;
- Clustering baseado na densidade.

O algoritmo mais conhecido do clustering baseado em Centróides é o algoritmo de (Lloyd, 1982) muitas vezes referido como *k-means*. Sendo do tipo partitivo, o utilizador começa por definir o número de clusters pretendido. O algoritmo define de forma aleatória no espaço de dados os centróides destes clusters que podem ou não coincidir com objetos existentes. Em cada iteração o algoritmo junta os objetos aos centróides mais próximos (minimização da soma do erro quadrático) através de cálculos com funções de distância ou semelhança e recalcula a posição dos centróides tendo em conta os objetos que estão a estes associados. O algoritmo termina quando não existirem um mínimo de objetos a mudar de cluster. Sendo este um problema de otimização *NP-hard*, o algoritmo procura soluções aproximadas encontrando ótimos locais. Uma vez que o algoritmo calcula os centróides iniciais de forma aleatória, cada corrida do algoritmo poderá resultar em soluções diferentes. O *k-means* tem dificuldade em detetar clusters reais que não sejam circulares ou que tenham tamanhos ou densidades incomuns. A existência de casos fora do normal (outliers) é conhecida por degenerar as soluções. Existem diversas variações do algoritmo para tentar melhorar a solução e que passam por definir métodos de cálculo da melhor posição dos centróides iniciais. É vulgar correr o algoritmo por diversas vezes e aceitar o melhor resultado encontrado.

O clustering baseado na distribuição usa métodos probabilísticos de cálculo de pertença dos objetos aos clusters através da aproximação do conjunto de dados a uma distribuição de probabilidade pré-definida, ou seja, os clusters vão sendo definidos em função do cálculo da verosimilhança de um determinado objeto lhe pertencer. O método probabilístico mais conhecido e usado desta técnica é o *Gauss Mixture Model*, que usa o algoritmo *Expectation-Maximization* (Dempster, Laird, & Rubin, 1977) onde cada cluster é representado por um vetor que inclui a

média e a matriz de covariância. De início os valores do número de clusters e do vetor representativo (posição inicial) podem ser estimados ou aleatórios, tal como no *k-means*. Em cada iteração é feito o cálculo de pertença dos objetos (*E-step*) através do teorema de cálculo da probabilidade de Bayes e em seguida são recalculados os parâmetros dos vetores representativos de cada cluster (*M-step*) para serem usados na iteração seguinte. O algoritmo pára quando a variação entre os valores verificados nos vetores representativos das classes entre iterações fica abaixo de um limiar definido, o que significa uma maximização da verosimilhança dos objetos aos clusters finais. Esta técnica tem geralmente bons resultados, mas uma convergência lenta causada pelos cálculos requeridos e obriga o utilizador a algum estudo preparatório como escolher corretamente o número de clusters e a distribuição de probabilidade que melhor representa o conjunto de dados (Couvreur, 1996). É vulgar correr o algoritmo com diversas distribuições e aceitar o melhor resultado.

O clustering baseado na densidade é do tipo partitivo e constrói clusters avaliando os objetos que estão próximos uns dos outros usando funções de cálculo de distância tal como no *k-means* só que avalia também se existem objetos em número suficiente (densidade) para serem considerados como parte desse subconjunto. O algoritmo mais conhecido desta técnica é o *DBSCAN* (Martin, Kriegel, Sander, & Xu, 1996) que, para definir se um objeto faz parte desse conjunto, usa o parâmetro *Eps* que não é mais que um raio de ação a partir de um objeto e avalia se dentro dessa área circular existem um determinado número mínimo de objetos vizinhos, o parâmetro *MinPts*. O algoritmo é sensível a estes parâmetros: um valor de *Eps* baixo não encontra vizinhos e num valor de *Eps* elevado todos farão parte do mesmo cluster. Já o *MinPts* define se o ponto é nuclear, de fronteira ou outlier: (1) nuclear é quando tem tantos ou mais vizinhos que o valor especificado, (2) fronteira é quando tem menos, mas tem pelo menos um ponto nuclear na sua vizinhança e (3) outlier se não for nenhum dos outros casos. O algoritmo é de computação de complexidade temporal *O(n*log n)* o que oferece uma vantagem sobre outros métodos de clustering descobrindo essencialmente os mesmos resultados. Esta técnica permite definir clusters com formas geométricas variadas, separando-se assim das formas tradicionais de clusters globulares mas não é recomendada quando o conjunto de dados é de alta dimensionalidade pois dificultam o cálculo das distâncias, nos casos de clusters com densidades muito diferentes visto que o parâmetro *MinPts* é sempre o mesmo, e ainda o problema dos pontos de fronteira que ficam na cauda de duas distribuições Gaussianas (clusters), pois o algoritmo analisa o número de vizinhos e não o decréscimo progressivo deste valor podendo mesmo definir um ponto de fronteira como nuclear, sendo preferível nestes casos o uso da técnica de clustering baseado na distribuição.

Recentemente têm sido desenvolvidas melhorias para colmatar as desvantagens dos algoritmos de referência de clustering. As técnicas de clustering partitivo são muitas vezes preferidas às de clustering hierárquico porque são capazes de lidar com conjuntos de dados maiores e a acurácia dos algoritmos de clustering partitivo é geralmente maior do que os de clustering hierárquico (Yoo, et al. 2007).

3.7.9. k-Nearest Neighbours

Os algoritmos da técnica Nearest Neighbours (NN) baseiam-se em medidas de distância entre objetos representados num plano, tal como os mencionados nas técnicas aglomerativas (clustering). Visto que a classificação se baseia em distância, estes algoritmos são ideiais para atributos numéricos normalizados e não para atributos categóricos. A predição baseia-se na premissa de que objetos que estão perto são mais similares do que objetos mais distantes. O parâmetro *k* indica o número de objetos que são considerados na determinação da classe. Cada objeto que participa na vizinhança vota para a determinação da classe sendo esta obtida por maioria. É também usual considerar pesos diferentes na votação consoante a proximidade dos objetos sendo que objetos mais perto têm maior contribuição que objetos mais afastados. Uma das vantagens destes algoritmos é o facto de serem *lazy*, isto é, não processarem dados na fase de treino, fazendo apenas memorização dos dados, e deixando todo o trabalho para a predição. Num modelo complexo com muita dimensionalidade (muitos atributos) a predição pode ser computacionalmente custosa na determinação da distância, recorrendo por isso a subconjuntos de atributos relevantes. Os algoritmos desta técnica são incrementais pois a adição de novos objetos não obriga a recálculo do modelo.

3.7.10. Regras de Associação

As regras de associação são uma técnica que procura relações entre itens que se encontram dispersos por uma grande quantidade de transações de dados, o que dificulta a sua descoberta por mera observação. Tendo como exemplo as transações de vendas realizadas num supermercado, é possível através da técnica de associação determinar produtos que são vulgarmente comprados quando outros também o são, conhecida por *Market Basket Analysis*. Segundo (Agrawal, Imielinski, & Swami, 1993) a descoberta das regras de associação faz-se em duas fases:

1. Descoberta dos cestos frequentes - As transações individuais são convertidas em compras por cliente. Através de métodos de análise combinatória, são gerados todos os possíveis conjuntos de itens comprados em conjunto (cesto). Cada cesto candidato é analisado através da proporção do número de vezes que aparece relativamente ao número total de transações existentes no conjunto de dados (medida de suporte). Este parâmetro é essencial para definir que apenas são frequentes aqueles cestos cujo aparecimento ocorre com maior frequência ou dito de outra forma, que esteja acima de um limiar (parâmetro de suporte mínimo) definido pelo utilizador;
2. Descoberta das regras interessantes – Tendo por base os itens de cada cesto frequente, são geradas as regras através de análise combinatória. Consideremos como exemplo que a partir do cesto {pão, leite, ovos, farinha} foi criada a regra {pão, leite, ovos} -> {farinha}, i.e., quem compra pão, leite e ovos também costuma comprar farinha. Esta regra é validada por uma medida de interesse, e.g. confiança que representa a proporção do número de ocorrências do item consequente {farinha} em cestos que contêm no antecedente os itens {pão, leite, ovos}. Só serão consideradas interessantes as regras cujo valor da medida de interesse esteja acima de um limiar (e.g. parâmetro de confiança mínima) definido pelo utilizador.

Como é fácil adivinhar, a análise combinatória gera todas as combinações possíveis e nem todas são desejadas por razões de eficiência computacional, tendo por sido isso criadas diversas técnicas que permitem eliminar as desnecessárias.

O algoritmo de referência *A-priori* (Agrawal & Srikant, 1994) faz uso do princípio que quando um cesto é frequente, todos os cestos possíveis de realizar combinando alguns dos mesmos itens (sub-cestos) também o são e o seu corolário que quando um cesto não é frequente, todos os cestos onde estes itens existam no conjunto também o não são (super-cestos). Também no que diz respeito às regras o algoritmo faz uso do teorema que se uma regra não satisfaz a condição imposta pelo parâmetro de confiança mínima, então qualquer regra composta pelo menos com os mesmos itens no antecedente também não satisfaz a referida condição. Manter a informação de todos os cestos frequentes em memória pode diminuir a performance. É possível compactar o espaço requerido para manter os cestos frequentes visto que estes podem ser gerados a partir de conjunto mínimo de cestos que possuam determinadas propriedades. Destas formas o algoritmo elimina a computação de cestos não frequentes e regras desinteressantes, acelerando a sua performance.

O algoritmo *FP-Growth* (Han, Pei, & Yin, 2000) vai mais longe na compactação da memória requerida para a geração dos cestos frequentes pela geração de uma estrutura de dados em árvore (*FP-tree*) representativa dos dados.

No final da indução, os nós contêm o número de transações em que o item aparece e os ramos representam os cestos com um item em cada nó, ordenados pela forma de inserção na árvore. A poda é possível através da soma das contagens dos itens e por comparação com o suporte mínimo; se for superior a este, tratam-se dos cestos frequentes. Em casos normais em que o conjunto de dados siga uma distribuição Gaussiana (com a maioria dos cestos com alguma repetição de itens), esta compactação traduz uma performance significativamente superior à do algoritmo *A-priori*. Apesar da existência de medidas objetivas de avaliação do interesse das regras extraídas, continua a ser um fator determinante a opinião subjetiva dos utilizadores conhecedores do domínio de aplicação.

3.7.11. Conjuntos Difusos

A teoria dos conjuntos tem sido o método tradicional de classificar e aglomerar elementos. Neste método os elementos têm uma relação ao conjunto através de um grau binário: essa relação existe ou não existe. Assumindo a incerteza natural, o professor (Zadeh, 1965) da Universidade de Berkeley propôs uma teoria complementar denominada Conjuntos Difusos[2] em que a relação dos elementos ao conjunto não é classificada em binário, mas com um grau contínuo. Nos Conjuntos Clássicos os elementos relacionam-se ou não a um conjunto e nos Conjuntos Difusos os elementos relacionam-se a um conjunto através de um determinado grau de pertença situado no intervalo [0,1] (variável dependente contínua), designada por função de pertença. Os Conjuntos Difusos generalizam os Conjuntos Clássicos quando a classe só assume o grau 0 ou 1. Desta

[2] Tradução do termo em inglês “Fuzzy Sets”.

forma podemos classificar objetos sujeitos a graus de subjetividade, como por exemplo se um quadro é bonito ou a dor que um paciente sente. Quer isto dizer que em vez dos valores possíveis da classe serem "sente dor" ou "não sente dor", na teoria dos Conjuntos Difusos poderíamos usar outros valores, conhecidos por termos linguísticos difusos, como "não sente dor", "sente uma dor ligeira, mas não restritiva", "sente uma dor forte, mas não restritiva", "sente uma dor forte e restritiva", ou ainda um valor numérico que expressasse essa dor, sendo assim uma classificação mais próximo dos valores reais.

A construção de modelos de classificação por Conjuntos Difusos passa por três passos: (1) a difusão[3] onde através da aplicação de funções de pertença aos valores de entradas são descritos os Conjuntos Difusos envolvidos, seguido pelo passo de (2) inferência com base em regras do tipo {Se->Então}, obtidas por métodos automáticos ou por conhecimento de especialistas e por fim o processo de (3) concisão[4] onde é atribuído um valor discreto à classe de saída com base novamente numa função de pertença. As funções de pertença (e.g. triangular, trapezoidal, *singleton*, gaussiana, etc.) servem para mapear os atributos de entrada em termos linguísticos difusos e destes em classes de saída consoante se trate do passo de difusão ou concisão, respetivamente. O processo de inferência é obtido por operações lógicas a que são sujeitas todas as regras de forma individual e de posterior operação de combinação. Em todos os passos mencionados por este método podem ser aplicados diversos algoritmos.

Esta técnica possui uma enorme vantagem que é a possibilidade de adaptação dos termos da linguagem natural aos valores numéricos dos sistemas computorizados, o que facilita também a compreensão dos modelos de classificação ou aglomeração obtidos. A desvantagem maior desta técnica é a multiplicidade de algoritmos possíveis para o processo de indução do modelo, que são de elevada complexidade e inibem a compreensão da origem do modelo obtido.

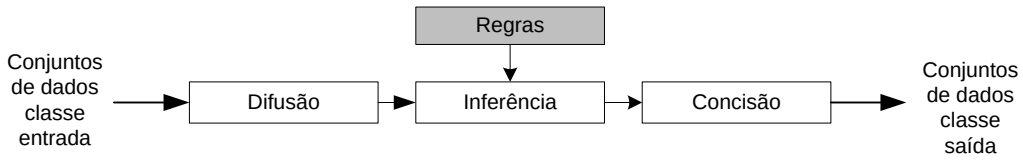

Figura 1 – Indução de modelos de Conjuntos Difusos (Zadeh, 1965)

Na mesma linha de pensamento sobre a incerteza natural dos Conjuntos Difusos surgiu a teoria dos Conjuntos Aproximados (Pawlak, 1982), onde a fronteira dos conjuntos também não é precisa. Segundo a terminologia do autor deste método, a relação dos elementos e conjuntos é chamada indiscernibilidade. Um objeto é indiscernível de outro quando ambos definem a mesma variável dependente, o atributo de decisão segundo a terminologia dos Conjuntos Aproximados.

A inferência das regras de classificação começa pela determinação de classes equivalentes que são conjuntos de casos que partilham os mesmos valores nas variáveis independentes ou atributos

[3] Tradução do termo em inglês "Fuzzyness".
[4] Tradução do termo em inglês "Defuzzification".

de condição, segundo a terminologia dos Conjuntos Aproximados. Os casos que pertencem a uma determinada classe equivalente podem ter ou não o mesmo atributo de decisão. Em segundo lugar é criada uma matriz de discernibilidade em que os eixos são as classes equivalentes e o valor do cruzamento dos eixos (célula) contém os atributos de condição que diferenciam essas classes equivalentes. Usando esta matriz podemos extrair as funções de discernibilidade relativa que possui o conjunto mínimo de atributos de condição que diferenciam esta classe equivalente das demais. Em seguida, às funções de discernibilidade relativa são retirados atributos de condição supérfluos usando operações de lógica booleana. Das reduções relativas é possível extrair as regras de classificação.

Se numa classe equivalente todos os casos que a constituem apontarem o mesmo atributo de decisão, esse também vai ser o valor da regra extraída. Nas classes equivalentes onde existem vários atributos de decisão possíveis, chamadas de classes vagas, o retorno é dado pela função de pertença aproximada que resulta num par de valor, o valor da probabilidade e o atributo de decisão. O valor de probabilidade é geralmente obtido pela proporção do número de regras existentes na classe equivalente que apontam para um determinado atributo de decisão a dividir pelo número total de regras dessa classe equivalente. Uma regra de uma classe vaga diz-se pertencer a vários atributos de decisão de acordo com os valores de probabilidade. Pode entender-se para fins de classificação o atributo de decisão com um valor de probabilidade maior, mas no entendimento puro sobre conjuntos aproximados, um objeto com estas caraterísticas será sempre classificado como pertencente a várias classes em determinada probabilidade.

A vantagem do método dos Conjuntos Aproximados é que não necessita de informação adicional aos dados pois é determinístico, mas devido ao seu formalismo matemático é por vezes difícil de compreender o processo de indução do modelo.

3.7.12. Algoritmos Genéticos

Os algoritmos genéticos são de inspiração biológica e mimetizam o processo de evolução natural das espécies descrito por Charles Darwin, ou seja, através de processos de seleção dos mais adaptados com recombinação e mutação genética. Os algoritmos genéticos são utilizados em problemas de otimização visto que resultam em procuras aleatórias no espaço de soluções. Soluções alternativas são codificadas no que se designa por cromossoma que são descritos por símbolos chamados genes.

Os algoritmos genéticos começam por selecionar cromossomas de forma aleatória que constituem a população. Os escolhidos são avaliados com a função de *fitness* e aos melhores é atribuída uma maior probabilidade de virem a ser novamente escolhidos (seleção natural). Se a condição de paragem não for atingida, são escolhidos novos cromossomas (nova geração) criados através de processos de recombinação e mutação dos genes e o processo de reavaliação é repetido até ser encontrada a condição de paragem da otimização. Os cromossomas da geração anterior são geralmente eliminados. A recombinação é um processo onde um par de cromossomas troca genes. A mutação é um processo que modifica os genes de um cromossoma. Ambos os processos causam perturbação na seleção introduzindo novas soluções não disponíveis em gerações anteriores. Pela verificação da função de fitness das soluções encontradas é obtida a solução

ótima. A paragem pode ser feita pelo número de gerações ou quando a solução obtém um valor desejado. Apesar das operações dos algoritmos genéticos fornecerem a oportunidade de encontrar a solução ótima, podem falhar (Sapna, Tamilarasi, & Kumar, 2012), especialmente quando o tamanho dos cromossomas for grande.

3.8. Ferramentas

As ferramentas de *software* de Data Mining conjugam fundamentos, teorias, métodos e algoritmos. Estas aplicações baseiam a sua operação em algoritmos que procuram padrões de conhecimento combinando um conjunto de ferramentas de interrogação e exploração dos dados com ferramentas que permitem a visualização de resultados e geração de relatórios. Existem atualmente muitas aplicações de descoberta de conhecimento e Data Mining disponíveis na forma mais tradicional plataforma proprietária, mas também em *open-source* e de livre utilização na Internet.

As ferramentas atuais de Data Mining são baseadas na execução de algoritmos genéricos e não incorporam conhecimento do domínio. Segundo (McCarthy, 2000) a ciência e o senso comum dizem-nos que os factos sobre o mundo não são diretamente observáveis, mas podem ser inferidos a partir de observações sobre os efeitos de ações, e.g. o facto de que alguém padece de Diabetes Mellitus é mais estável do que um padrão de compras particular que pode permitir inferir que o comprador padece desta condição. Os factos sobre um fenómeno são mais preditivos de comportamento futuro que os factos observáveis. Este autor defende a inclusão do conhecimento do domínio nas ferramentas, mas isso não é ainda o estado da arte atual. O que é comum realizar com as ferramentas atuais é ter especialistas do domínio a validarem *a posteriori* o conhecimento extraído.

A avaliação de um modelo induzido através de técnicas de Data Mining pode ser muito positiva, mas isso não invalida a sua desatualização com o passar do tempo. É necessária a atualização dos modelos com a periodicidade devida caso-a-caso. Os modelos podem ser atualizados de diversas formas, periodicamente recriando o modelo novamente com base nos dados anteriores e nos novos dados adicionados, ou consoante o surgimento de novos casos através de sistemas mais especializados, como a inclusão de modelos induzidos através de CBR (Huang, Chen, & Lee, 2007) que permitem a adição de novos casos à base de conhecimento à medida que os mesmos vão surgindo. As ferramentas de Data Mining mais comuns permitem a criação de modelos, mas são os SAD criados a partir destes que suportarão os processos críticos das organizações.

3.9. Domínios de Aplicação

Em termos de aplicação, o Data Mining já foi aplicado com sucesso: por instituições financeiras, para classificação de clientes com vista à concessão de crédito e deteção de fraudes; para *marketing* direto, *cross-selling* e *up-selling*; no comércio a retalho, para segmentação de mercado e disposição de produtos no local de venda; por fabricantes, para controlo de qualidade e agendamento de manutenção (Koh & Tan, 2005).

Nas vendas a retalho é importante conhecer padrões de consumo para poder realizar campanhas de *marketing* e publicidade. Através da extração de regras de associação a partir dos produtos vendidos é possível encontrar produtos que normalmente são vendidos ao mesmo tempo que outros. Aumentando a proximidade de localização desses produtos ou direcionando descontos em proporção à quantidade é possível dinamizar as vendas.

A forma tradicional de aplicação de Data Mining tem sido a junção de conhecimento de especialistas do domínio com técnicas de modelação para a solução de problemas específicos. Contudo, a evolução desta disciplina tem criado novos desafios: (1) Os conjuntos de dados começam a ser de elevada cardinalidade e são mesmo um novo ramo de investigação conhecido por *Big Data*, (2) a concorrência exige ferramentas que produzam modelos de rápida construção onde começam a surgir os modelos atualizados em permanência através de *streaming* e (3) os utilizadores começam a entender a disciplina e a exigir que os resultados dos modelos sejam facilmente interpretáveis e acionáveis no seu contexto.

A aplicação de Data Mining já provou ser de alta eficácia e eficiência no tratamento de problemas importantes (Apte, Liu, Pednault, & Smyth, 2002). Este domínio continuará a ser fundamental na construção de SAD, o que irá requerer a melhoria das técnicas subjacentes às ferramentas de Data Mining para melhor apresentação e compreensão dos modelos construídos. A melhoria passa por automatizar, escalar e tornar mais confiável o modelo gerado.

3.10. Data Mining na Saúde

O campo de estudo da Informática da Saúde[5] tem cerca de duas décadas e é definido como a aplicação de ciência da computação, tecnologias de comunicação e gestão de bases de dados para a organização, entrega e análise de toda e qualquer informação relevante para os cuidados de saúde (Cios & Moore, 2002).

O Data Mining em saúde está progressivamente a tornar-se essencial e os fatores de motivação são vários (Koh & Tan, 2005):

- A existência de fraude e abuso nos seguros de saúde está a levar as companhias de seguros a usar ferramentas de Data Mining para detetar quem ofende;
- Os grandes volumes de dados gerados pelas transações de saúde impedem o processamento e análise através de métodos tradicionais;
- A possibilidade de extração de conhecimento a partir dos volumosos repositórios de dados leva a que as próprias instituições de saúde possam conhecer melhor os utentes e assim direcionar os seus serviços – *CRM*[6];
- Por razões de análise de custo e faturação, auxiliar no agrupamento dos utentes pois têm geralmente tratamentos relativamente constantes.

[5] Tradução do termo em inglês "Health Informatics"

[6] Acrónimo em inglês que abrevia "Customer Relationship Management" e que designa software com o propósito de desenvolver os relacionamentos de uma entidade com os seus clientes.

A aplicação bem-sucedida de Data Mining fornece novo conhecimento biomédico e de cuidados de saúde que podem ser usados de forma efetiva no suporte à tomada de decisão no processo de diagnóstico, escolha de opções terapêuticas e predição de prognóstico, mas também na tomada de decisão de aspetos administrativos como a definição das equipas médicas, realização de seguros de saúde, tendências de mercado e demografia, garantia de qualidade e eficiência de processos (Yoo, et al., 2012). Em suma, todo o conhecimento que possa ser aplicado no fornecimento de melhores cuidados de saúde aos pacientes.

A aplicação de Data Mining em saúde pode ser agrupada em quatro áreas (Koh & Tan, 2005):

- Eficiência de tratamento – obtida pela comparação de sintomas apresentados pelos pacientes e o resultado do tratamento prescrito pelos profissionais de saúde;
- Gestão de cuidados de saúde – identificação e seguimento de pacientes de doenças crónicas ou pacientes de alto risco, intervenções apropriadas e redução de admissões e custos hospitalares;
- CRM – gestão de interações entre organizações de saúde e pacientes através de *call centres*, consultórios, serviços de faturação e pagamento;
- Deteção de fraude e abuso – aplicações que estabelecem normas e identificam desvios ou padrões anormais para evidenciar prescrições indevidas e potencialmente perigosas bem como ativação de seguros fraudulentos;

Este estudo incide na área de gestão de cuidados de saúde embora estas áreas estejam dependentes umas das outras.

O Data Mining na saúde tem especificidades próprias que o diferenciam de outros domínios de aplicação:

- Heterogeneidade dos dados médicos – Relacionada com o alto volume e complexidade dos dados e falta de estruturação conceptual;
- Questões éticas, legais e sociais – Propriedade e proteção dos dados usados em contexto comercial, questões éticas que se levantam com os métodos e descobertas científicas aplicadas ao ser humano;
- Filosofia estatística subjacente – Questões que se levantam com o uso de métodos estatísticos que tendem a generalizar em contraponto com a importância de cada vida humana.

A questão mais pertinente que se coloca hoje no Data Mining na Saúde é a capacidade de levantar as questões apropriadas. Tal só poderá ser possível aumentando o conhecimento das capacidades do Data Mining pelos profissionais de saúde e a melhoria da qualidade dos dados. Resolvendo estas questões, quanto mais exaustivamente esta área for estudada mais nos aproximaremos das respostas relevantes.

3.11. Dados

A saúde é considerada como o domínio mais difícil em Data Mining (Roddick et al., 2003) pelo facto de o conhecimento empírico ser fundamental e por não estar armazenado em repositórios informáticos, mas muitas vezes ser conhecimento intrínseco dos profissionais de saúde.

A prática clínica é um procedimento complexo em que os profissionais de saúde devem deduzir diagnósticos e definir terapias com base em conhecimento empírico. É de extrema importância que os dados clínicos existentes possam providenciar conhecimento para casos difíceis e/ou similares.

Os dados usados em Informática da Saúde são criados por e para prestadores de cuidados de saúde, não sendo muitas vezes óbvio para quem cria os dados dos benefícios que a correta entrada desses mesmos dados trará para quem os estuda e que com eles melhora a prática dos cuidados de saúde. Como tal, os dados estão com frequência em falta ou com erros e a estrutura dos dados é muitas vezes inadequada obrigando a transformações que podem comprometer resultados. Especialistas no mesmo domínio podem utilizar terminologia diferente para descrever a mesma coisa, o que obriga a estruturação de conceitos.

A estruturação de registos médicos eletrónicos é ainda uma matéria de investigação (Los et al., 2004). Não havendo um consenso abrangente leva a que muitos dados existam na forma de texto livre (não estruturado). A introdução de dados nestes registos é ainda uma tarefa essencialmente manual. Estas condições levam a que os dados com registos médicos contenham erros de introdução.

Os dados considerados para prática de cuidados de saúde são de elevado tamanho, quer em cardinalidade quer em dimensionalidade: incluem condições de doenças como sintomas e sinais, mas também aspetos ambientais e fatores sociais (Xuezhong, et al., 2010). A indução de conhecimento é difícil devido à elevada rede de interligações entre sintomas, diagnósticos e terapias.

A complexidade na obtenção de conhecimento extraído de bases de dados sobre saúde deve ser devidamente comparada com os benefícios daí resultantes. A aplicação de técnicas de Data Mining na saúde tem sido alvo de intensa investigação nos últimos anos. A revisão feita por (Khajehei & Etemady, 2010) enumera como exemplo estudos na deteção de cancro, genética, psicologia, problemas do sono, Diabetes e nefrologia com as técnicas de SVM, *Text Mining*, RNA, AD, Regressão e Ensemble.

A qualidade de conhecimento extraído é proporcionalmente igual à qualidade dos dados existentes, i.e., "*garbage in, garbage out*"[7]. A fraca qualidade dos dados, a representação inconsistente devido à falta de uniformização, a necessidade de reconverter dados criados por outrem e a complexidade do domínio da saúde fazem com que extrair conhecimento de bases de dados sobre saúde seja uma tarefa complexa.

Para combater a Diabetes Mellitus, o governo de Singapura iniciou em 1992 um rastreio dos pacientes nos hospitais públicos (Apte, Liu, Pednault, & Smyth, 2002). Com esta base de conhecimento pretende-se a criação de regras que ajudem os prestadores de cuidados a melhorar o prognóstico dos pacientes. Contudo, existem questões limitativas à sua aplicação. Os dados

[7] Expressão Inglesa utilizada no domínio da computação que indica que a qualidade dos dados resultantes deriva da qualidade dos dados em que se baseiam esses resultados.

possuem erros de introdução, valores em falta, erros de conversão numérica, redundância e erro de formato para aplicação das técnicas. Visto que o conjunto de dados é grande assim são também as regras extraídas e os especialistas estão demasiado ocupados para avaliarem a qualidade das regras. Quanto ao primeiro problema, foram desenvolvidas tarefas semiautomáticas para a correção de alguns erros dos dados. Quanto ao facto de serem extraídas demasiadas regras, a equipa de trabalho começou por analisar as regras com valores mais significativos e usar ferramentas de visualização. Os médicos confirmaram que as regras estavam de acordo com a prática por si observada, mas ficaram surpreendidos com algumas exceções que desconheciam previamente.

As bases de dados em saúde têm com frequência falta de valores causados por vários fatores (Yoo, et al., 2012): (1) Pacientes com a mesma doença não fazem necessariamente os mesmos exames por terem caraterísticas biológicas ou socioeconómicas diferentes; (2) Os SI que gerem as unidades prestadoras de saúde estão orientados para a contabilidade e faturação e não para a prestação de auxílio à prática médica; (3) A manutenção de dados em papel continua a ser prática comum (mais que não seja em termos de dados passados).

3.12. Limitações e Fatores Críticos de Sucesso

A acessibilidade aos dados é muitas vezes problemática pois não existe uniformização de sistemas entre os ambientes que os usam como o hospital, o laboratório, as clínicas e outros (Koh & Tan, 2005). A solução pode passar por construir um Data Warehouse, mas muitas vezes o custo pode ser proibitivo para o projeto. Seja como for, os dados necessitam de serem processados antes da aplicação das técnicas e este trabalho é muitas vezes realizado por um analista de dados não especializado na área da saúde que pode inadvertidamente introduzir dados erróneos. É recomendável que este passo seja acompanhado ou validado por um especialista no domínio de aplicação.

Os dados são fonte de uma miríade de problemas: o armazenamento é diverso, em grande quantidade e localizado em várias fontes. A não existência de um vocabulário médico padronizado leva a registos com dados complexos, heterogéneos e com falta de representação matemática e forma canónica. Poderão ser fatores limitativos também as questões éticas, legais e morais (Koh & Tan, 2005). Existem técnicas de Text Mining que poderão ser úteis para dados existentes. Relativamente ao futuro, muitos destes problemas podem ser resolvidos recorrendo ao uso adequado de técnicas de Interação Pessoa-Máquina no desenvolvimento aplicacional, por exemplo construindo controlos de validação adequados e restritivos como listas que evite a introdução textual para campos numéricos. A qualidade da informação vai sempre depender da qualidade dos dados introduzidos.

As conclusões do estudo de (Koh & Tan, 2005) alertam ainda para o problema de as técnicas de Data Mining extraírem predições assentes em flutuações aleatórias e não reprodutíveis causadas pelo excesso ou escassez de quantidade de dados. As técnicas de Data Mining são parametrizáveis para limitar a informação resultante e a quantidade de dados não deve ser aquela

de onde seja possível extrair informação, mas aquela que se encontrar disponível e que tenha interesse para o trabalho.

Data Mining é uma área multidisciplinar que requer conhecimentos de domínios como a estatística, a informática, a aprendizagem automática, entre outras, e da área onde os trabalhos são realizados, como a saúde no caso deste trabalho. Não é, portanto, fácil encontrar pessoas que tenham conhecimentos profundos em todas estas áreas. O trabalho é normalmente resultado de um esforço coletivo e onde são importantes técnicas de gestão destes recursos humanos e de gestão de projetos.

As organizações de saúde que se encontrem a desenvolver trabalhos de Data Mining necessitam de usar recursos significativos: de pessoas, de tempo e de capital (Koh & Tan, 2005). É por isso imprescindível que o trabalho tenha suporte da gestão de topo, expetativas realistas, gestores experientes, prazos adequados e recursos disponíveis. Os prestadores de serviços de saúde devem estar convencidos da utilidade do trabalho e dispostos a colaborar numa possível restruturação de processos.

Numa perspetiva mais técnica, os algoritmos possuem com frequência a necessidade de parametrização. Porém, os utilizadores não têm, geralmente, suficiente informação sobre a correta parametrização. Existe uma correlação de qualidade entre os resultados dos modelos produzidos pelos algoritmos e os parâmetros e por isso a importância de estes estarem corretamente definidos.

Embora uma acurácia elevada de um modelo produzido possa parecer interessante para um investigador, no domínio da saúde o erro pode ser muito dispendioso e acima de tudo com pesadas consequências para o ser humano. A fraca qualidade de um modelo pode ser causada pela falta de dados, pelo uso de dados erróneos, pela escolha errada do algoritmo, pela má parametrização, pela errada metodologia ou pela fraca interpretação e aplicação dos resultados; tudo pode ser um fator limitativo e carece de trabalho adicional, muitas vezes com base no conhecimento empírico dos prestadores de cuidados de saúde.

Tanto quanto foi possível investigar, não existe uma ferramenta de Data Mining dedicada em exclusivo ao domínio da saúde. As ferramentas são generalistas por razões económicas e esse facto é limitativo pois não incorporam conhecimento do domínio durante o processo de descoberta de conhecimento.

3.13. Pré-processamento

A preparação dos dados para processamento é a primeira fase a realizar num processo de aplicação de Data Mining. Considerada pouco glamorosa, é, contudo, a fase mais importante e muitas vezes a razão determinante do sucesso. Os repositórios em saúde possuem dados que são muitas vezes confusos, com distribuições anormais, com dados em falta e variáveis grosseiramente definidas (Castellani & Castellani, 2003). De acordo com a tarefa pretendida, existem algumas orientações que podem ser seguidas para minimizar a questão do ruído nos dados.

Em tarefas de classificação, o objetivo é descobrir os atributos e seus valores que são determinantes na definição da classe. Os conjuntos de dados contêm vulgarmente atributos que não servem este objetivo e nada mais são do que ruído nos dados que pioram a performance dos modelos; dados redundantes como a idade e a data de nascimento ou irrelevantes como o género sexual num modelo de classificação de cancro da próstata, devem ser corrigidos antes da aplicação do algoritmo. Existem técnicas como a análise de correlação ou deteção dos atributos mais relevantes que permitem lidar com dados erróneos, porém estas têm limites a partir dos quais é comprometida a qualidade dos resultados (Castellani & Castellani, 2003). Estas técnicas analisam atributos de forma individual e embora um atributo só por si possa ser considerado ruído, a sua combinação com outros pode ter muita relevância. Há, por isso, que ter algum cuidado com o uso de técnicas que eliminam atributos considerados ruído.

Para uma tarefa de Data Mining podem existir muitos algoritmos com muitas variantes. Com a enorme variedade de ferramentas existentes, realizar uma tarefa de classificação de forma exaustiva poderia significar centenas de modelos gerados. No teste realizado por (Borges, Marques, & Bernardino, 2013) é possível observar que, quando sujeitos aos mesmos algoritmos e ferramentas, diferentes conjuntos de dados podem determinar diferentes melhores algoritmos classificadores. Na quase impossibilidade de se realizarem todos os testes para um conjunto de dados, é sempre possível existir um algoritmo que melhore a classificação. Não é praticável testar todos os algoritmos, mas é recomendável correr aqueles com os quais é normalmente possível obter, se não o melhor resultado, pelo menos um valor muito aproximado.

Sempre que os conjuntos de dados originais são de elevada dimensão é comum serem usadas técnicas estatísticas de amostragem. Sobre a amostra (conjunto de dados) é ainda realizado um particionamento de dados que divide a amostra para o treino e teste do modelo. É possível que os casos do conjunto de treino sejam de classificação muito fácil ou muito difícil (Yoo, et al., 2012). Por esse motivo e porque não é aceitável testar o modelo com os casos usados no seu treino, deve ser usada uma técnica de particionamento como as seguintes:

- *k-fold cross-validation* – esta técnica consiste em particionar o conjunto de dados em *k*-subconjuntos, normalmente 5 ou 10. São criados *k*-modelos e em cada um são usados *k-1* subconjuntos para treino deixando de fora um dos subconjuntos para teste, por isso também chamada de "*leave-one-out*". Cada modelo gerado é testado com um subconjunto diferente. A média dos resultados dos *k*-modelos é habitualmente considerada o resultado do modelo final;
- *Holdout* – neste modo de particionamento é comum falar-se em percentagens. Um *percentage split* de 70:30 significa que 70% dos casos são aleatoriamente selecionados para integrar o conjunto de dados de treino e os restantes 30% para teste.

O *k-fold cross-validation* é mais usado em conjuntos de dados pequenos e permite que os casos sejam usados tanto para treinar o modelo como para o testar. É importante lembrar que o objetivo é classificar casos novos não representados no conjunto de dados.

Relativamente a tarefas de clustering, quando o conhecimento sobre os dados é pouco ou nenhum, devemos começar por sujeitar o conjunto de dados a algoritmos de clustering

hierárquicos pois o dendrograma resultante indiciará o número de clusters existentes que é um parâmetro de entrada nos demais algoritmos. Uma vez que o clustering hierárquico é de elevada exigência computacional, os conjuntos de dados de elevada dimensão deverão ser previamente sujeitos a um processo estatístico de amostragem. O dendrograma resultante será de menor dimensão e melhor compreensão. Após a determinação do número de clusters, poderá ser usado o método *bisecting k-means* pois diversos estudos mencionam-o como aquele que atinge melhores resultados de acurácia, onde a classe está presente (Yoo, et al., 2012).

Os outliers podem ter interessante em diversos domínios de aplicação, mas em clustering podem degenerar soluções e devem ser eliminados, caso se trate de ruído, ou considerados relevantes caso um especialista do domínio assim o entenda.

Alguns algoritmos apenas tratam dados numéricos, o que causa a necessidade de conversão dos dados categóricos. Esta conversão pode, no entanto, resultar em distorções quando avaliada a sua distância. Tomemos como exemplo os diagnósticos para os três tipos de Diabetes: Diabetes gestacional, Diabetes Tipo I e Diabetes Tipo II. Quando convertidos para um valor numérico resultaria em 1,2 e 3, respetivamente. Usando uma medida de distância unidimensional, significaria que a relação entre a Diabetes Gestacional e a Diabetes Tipo I era menor ($|1-2|=1$) que a Diabetes Gestacional e a Diabetes Tipo II ($|1-3|=2$). É essencial que a conversão respeite os níveis de correlação existentes nos dados.

Considerando a regra de associação "quem compra pão e manteiga habitualmente compra leite", verificamos que na realidade estes produtos não existem nas prateleiras do supermercado visto que são categorias de produtos e não produtos per si. A geração de regras com todos os produtos eleva a complexidade de descoberta de regras interessantes pela sua dimensão (muitas regras) e falta de representatividade (casos esparsos). A categorização dos dados permite reduzir as regras a dimensões interpretáveis e capazes de serem descartadas ou não por especialistas do domínio. Este método é chamado de mineração de regras de associação multinível.

As quatro maneiras mais comuns de lidar com dados em falta são:

- Apagar os registos – é de resolução fácil, mas pode eliminar valores noutros atributos que podem ser importantes;
- Substituir valores em falta pela média – pode enviesar os resultados introduzindo dados que não são reais;
- Substituir valores em falta por zeros – os algoritmos entendem o valor como sendo um valor real enviesando os resultados;
- Substituir valores em falta pelo valor do vizinho mais próximo ou pela média dos *k* vizinhos mais próximos – dependendo dos dados, este valor poderá estar corretamente aproximado ou completamente errado.

3.14. Seleção de atributos

O processo de determinação dos atributos é muito relevante no processo de Data Mining pois tem uma relação direta sobre a qualidade dos dados. A acurácia e consistência são as medidas mais importantes (Patil, Joshi, & Toshniwal, 2010). Se um atributo estiver mal definido, tem que

ser corrigido ou melhorado, caso contrário deve ser descartado. Visto que os modelos são sensíveis aos valores dos atributos, existem técnicas que os permitem incluir ou excluir.

A seleção de atributos envolve a comparação dos registos de um conjunto analisando o valor da sua contribuição para a tarefa em causa. Este método é conhecido por análise de sensibilidade. O seu objetivo é reduzir o custo e complexidade para melhorar a acurácia do modelo e a sua visualização e compreensão.

Existem diversas formas de seleção de atributos, mas todas visam a minimização dos conjuntos de dados mantendo os atributos com maior contribuição para o modelo e descarte dos demais. Os critérios de seleção de atributos usados neste trabalho são definidos no artigo de (Novakovic, 2010): *Information Gain*, *Gain Ratio*, *Symmetrical Uncertainty*, *Relief-F*, *One-R* e Qui-Quadrado.

Um dos métodos possíveis de cálculo dos atributos mais significativos é pegar em cada um e verificar o seu contributo individual para a qualidade do modelo. Após ser determinado o atributo mais significativo, combina-se este com os outros e verifica-se o contributo de cada conjunto. Procede-se sucessivamente dessa forma à medida que o modelo melhorar. O processo termina quando a adição do atributo piorar o resultado do modelo. Este método foi o utilizado por (Ban, Heo, Oh, & Park, 2010) para a identificação de casos de Diabetes Tipo 2 a partir de dados genéticos através de um classificador obtido com SVM. Esta estratégia visa a otimização do modelo e é dependente dos dados usados, não garantindo o resultado para diferentes casos.

Outro dos métodos possíveis de cálculo dos atributos mais significativos é através de cálculos estatísticos como o teste do *Qui-Quadrado*. Este método foi utilizado por (Meng, Huang, Rao, & Liu, 2012) e que lhes permitiu avaliar a importância estatística dos atributos para o seu conjunto, tendo rejeitado os atributos com um nível de significância superior a 0,05. Estes autores vão mais longe ao definir ordem nos atributos mais significativos: no modelo de regressão logística baseiam-se nos valores dos coeficientes, nas RNA pelos valores dos pesos e no caso da AD no valor obtido pelo indicador de entropia *Information Gain*.

No tratamento da Diabetes Mellitus são muitos os atributos que são registados, mas apenas um número reduzido é utilizado, contribuindo assim para a existência de dados com ruído, irrelevantes e redundantes. Os algoritmos de aprendizagem com base em instâncias vizinhas têm a vantagem de serem tolerantes com dados com ruído. Por exemplo, o *Relief-F* (Kira & Rendell, 1992) que estima a qualidade dos atributos de acordo com a sua capacidade de distinguir entre as instâncias que estão perto umas das outras. O algoritmo *Fssmc* (Huang, McCullagh, Black, & Harper, 2007) é uma otimização do *Relief-F* e foi usado para redução do conjunto de dados do estudo destes autores de 410 para 8 atributos. Na prática, qualquer algoritmo de classificação pode ser usado como método de seleção de atributos para redução do conjunto de dados.

3.15. Privacidade e segurança de dados

O trabalho em bases de dados em saúde implica o acesso a dados protegidos visto que retratam de forma direta condições de indivíduos. É fundamental assegurar que seja possível obter o

conhecimento desejado sem pôr em causa as questões de privacidade e segurança. O Data Mining em saúde envolve inevitavelmente as questões de privacidade e segurança dos indivíduos, o que é uma diferença considerável com outras áreas.

Apesar de parecer óbvio que os dados de saúde de uma pessoa pertencem-lhe, a verdade é que a definição da propriedade dos dados é mais complicada. Os dados são mantidos tanto pelas instituições prestadoras de cuidados de saúde como de empresas seguradoras. Estas empresas poderão, em condições de garantia de anonimato, vender os dados ou o conhecimento dele extraído?

Para assegurar a devida proteção de dados, há que acautelar o respeito pelo estabelecido legalmente e vigiado pela Comissão Nacional de Proteção de Dados. Esta proteção implica a preservação da confidencialidade através de estratégias de remoção de identificação e garantia de anonimato dos utentes bem como medidas de segurança de acesso aos dados mais sensíveis. Embora desejável, esta regulação retira informação que pode ser valiosa aos analistas de dados e é importante encontrar o equilíbrio entre as duas partes.

Embora um assunto controverso, a verdade é que os dados guardados em bases de dados da saúde podem ainda esconder práticas médicas incorretas. Enquanto noutros domínios este facto poderia ser considerado menosprezável ou um desvio aceitável, em saúde poderá ter consequências graves e indesejáveis em indivíduos. Por este facto os prestadores de cuidados de saúde podem não ver o Data Mining com bons olhos. O Data Mining não pretende expor estas questões ou escolher entre mostrar ou esconder o conhecimento por motivos éticos. O Data Mining limita-se a analisar os dados para descobrir conhecimento.

3.16. Investigação

Existe muita investigação feita no desenvolvimento e melhoria de algoritmos de Data Mining que visa melhorar os modelos produzidos e utilizados nas mais diversas áreas de conhecimento humano, incluindo a saúde e em particular a Diabetes Mellitus e a Hipertensão. De acordo com o artigo de (Kaplan, 2008) estas doenças eram as duas mais significativas no custo do plano de saúde da empresa médica estado-unidense Segal em 2007, totalizando 12% e 9%, respetivamente. O investimento em Data Mining permitiu ganhar controlo sobre os fatores que elevam o custo e assim evitando a passagem dos custos para os funcionários e a manutenção do seu plano de saúde.

Uma rápida procura nos repositórios de literatura científica pelos termos “Data Mining” e “*Health*” devolve dezenas de milhares de artigos sobre modelos preditivos e descritivos criados nos campos da Biomedicina e Saúde sobre deteção de fraudes, pacientes não diagnosticados, custo de prestação de cuidados, diagnóstico e prognóstico de doenças, duração de internamento e associações entre doenças e medicamentos. Descrevem-se de seguida alguns destes casos com o intuito de demonstrar a possibilidade de utilização de técnicas de Data Mining nesta área de conhecimento científico, tanto na prestação de cuidados como na investigação.

A companhia de seguros norte-americana Highmark construiu um modelo de classificação para detetar reclamações fraudulentas em tempo quase real. De acordo com (Yoo, et al., 2012) estas fraudes atingem cerca de 10% das despesas de saúde anuais nos Estados Unidos da América, de acordo com os números oficiais das agências governamentais. Com este classificador, a Highmark reduziu o tempo necessário na condução das investigações e na tomada de decisão. A Highmark continuou a aperfeiçoar o modelo de acordo com os novos dados. Esta atualização constante originou proveitos na ordem dos 11,5 milhões de USD em 2005. Através de um conjunto de dados que inclui os sintomas de pacientes, historial médico e aspetos demográficos, a Highmark construiu um modelo de árvores de decisão que lhes permitiu identificar pacientes que não estavam diagnosticados. A redução do tempo no diagnóstico permitiu aumentar o número de reembolsos exigido às organizações públicas de saúde e assim aumentar os proveitos em milhões de USD e as condições de saúde oferecidas aos pacientes.

A empresa estado-unidense Healthways presta serviços de saúde para melhoria do bem-estar dos pacientes através de maximização dos seus investimentos. Nesse sentido desenvolveu modelos de RNA com o objetivo de identificar pacientes em maior risco de desenvolver determinadas patologias e determinar se estes têm capacidade para cumprir com os métodos de cuidados apropriados (Yoo, et al., 2012). De acordo com essa capacidade, foram definidas ações de intervenção direcionada e prevenção planeada, reduzindo o custo na prestação dos cuidados aos pacientes.

Na Bioinformática são estudados, entre outras coisas, algoritmos que permitem comparar a sequenciação dos nucleótidos de ADN de diversas amostras para identificar desigualdades causadoras de doenças genéticas. Através da técnica biomolecular conhecida por *microarray*, é possível recolher amostras de ADN e distinguir entre doenças muito similares. Através de clustering, (Golub, et al., 1999) construíram um modelo classificador de *microarrays* capaz de distinguir entre duas classes de Leucemia que à data eram difíceis de diagnosticar com precisão e assim contribuir na melhoria do tratamento.

(Hu, et al., 2006) usaram 7 conjuntos de dados sobre saúde para realizar uma comparação de resultados dos modelos produzidos com técnicas de SVM, Árvores de Decisão e métodos de Ensemble. O método de validação usado foi o *Cross-Validation* com *10-fold* e a métrica de avaliação a acurácia. O melhor resultado foi obtido em 4 modelos baseados na técnica de SVM, mas devido à má performance dos restantes, o valor médio desta técnica foi ultrapassado pelos métodos de Ensemble.

(Harper, 2005) usou 4 conjuntos de dados sobre saúde (um dos quais sobre Diabetes Mellitus) para realizar uma comparação de modelos produzidos por algoritmos de classificação com técnicas de Regressão, Árvores de Decisão e RNA. O modelo com a melhor acurácia foi construído com o algoritmo *CART* (Árvore de Decisão) e o pior do estudo foi a RNA.

(Potter, 2007) aplicou 56 algoritmos de classificação e Ensemble existentes no *software* Weka v.3.5.5 a 2 conjuntos de dados de saúde sobre tumores para avaliar o seu diagnóstico e prognóstico. Para tal usou o método de avaliação *Cross-Validation* com *10-fold*. Os resultados

demonstraram que o melhor algoritmo para um dos conjuntos de dados nem aparece no *top* 5 do outro conjunto de dados. Concluiu ainda que quando o conjunto de dados é alterado através de métodos de seleção de atributos, o mesmo contribui para que o melhor algoritmo seja também outro.

(Delen, et al., 2005) usou uma base de dados sobre o cancro da mama para prever a sobrevivência superior a 5 anos após o diagnóstico, criando modelos de RNA, Árvores de Decisão (*C5.0*) e Regressão Logística. A experiência é interessante pelo elevado número de registos que totalizam 433,272 casos cada um com 72 atributos. O método de validação usado foi o *Cross-Validation* com *10-fold*. A Árvore de Decisão obteve o melhor resultado de acurácia, sendo seguido pela RNA e por último a Regressão Logística.

De acordo com os dados da Associação Americana de Sistemas de Dados Renais, em 2007 foram usados acima de 8 mil milhões de USD em tratamentos de hemodiálise para 341,264 pacientes em último estágio de doença renal. O número de pacientes está a aumentar sendo que para tal contribui em 54% a Diabetes Mellitus e em 33% a Hipertensão, doenças em que a taxa de prevalência continua a aumentar. A esperança média de vida de um paciente em último estágio de doença renal é de 3 anos. Estes alarmantes números levaram (Shah, et al., 2003) a usarem técnicas de Conjuntos Aproximados e Árvores de Decisão para preverem os fatores significativos na contribuição para o aumento da esperança média de vida destes pacientes. Os fatores críticos identificados foram: altura do diagnóstico, tempo de diálise, desvio do peso-alvo, valores de tensão arterial, níveis de cálcio e potássio na solução de diálise, o volume total de sangue, o rácio de fluxo arterial e pressões venosas.

A classificação pode ainda ser utilizada para a previsão de custos na prestação de cuidados de saúde, tal como o demonstraram (Bertsimas, et al., 2008). Estes investigadores usaram os dados de seguradoras Estado-Unidenses e que incluem despesas médicas e farmacêuticas de 2866 organizações referentes a 838.242 pacientes, ocorridas durante o período de 8 de janeiro de 2004 a 31 de julho de 2007. Os dados dos 2 primeiros anos foram usados para treino e o último ano para teste. Os custos foram agrupados em *5 bins* e os atributos foram sujeitos a estratégias de seleção, tendo sido obtido uma acurácia de 84,6%.

O clustering é geralmente utilizado quando pouco se conhece sobre os dados. Por esse facto, os investigadores (Van't Veer, et al., 2002) usaram 98 *microarrays* de ADN com cancro da mama e produziram dois clusters através da técnica de clustering hierárquico para revelarem aqueles em que surgiriam metástases distantes dentro de 5 anos, revelando que a qualidade do prognóstico através de análise por *microarray* é superior à predição por parâmetros clínicos (condições de saúde e sintomas).

As regras de associação permitem revelar relações escondidas nos dados, em especial quando os conjuntos são transacionais e de cardinalidade elevada. A Organização de Saúde Sul Coreana sem fins lucrativos (KMIC) regista de forma bianual a informação demográfica, biomédica e ambiental dos seus utentes para determinar programas de saúde governamentais. Os investigadores (Chae, et al., 2001) usaram técnicas de extração de regras de associação de

Árvores de Decisão para ajudar à formulação de um programa governamental de gestão da Hipertensão.

Reconhecendo a existência de associação entre doenças (comorbidade) e medicamentos e doenças e testes laboratoriais, os investigadores (Wright, Chen, & Maloney, 2010) verificaram a possibilidade de uso de Regras de Associação para comparar os registos médicos eletrónicos de uma base de conhecimento com elevada cardinalidade com bases de dados de referência. O conjunto de dados dizia respeito a uma amostra de 100.000 pacientes, 272.749 doenças, 442.658 medicamentos e 11.801.068 resultados de testes laboratoriais. As hipóteses do seu trabalho incluíam ainda a comparação do custo de criação e manutenção das regras com uma base de conhecimento de gestão manual. Aplicaram o algoritmo *A-priori* em *software* criado pelos próprios, com suporte=5% e confiança=10%, e avaliaram as regras obtidas com as medidas suporte, confiança, qui-quadrado, interesse e convicção. Um dos aspetos relevantes nas suas conclusões foi que os melhores resultados variam consoante as medidas de interesse e que por esse motivo um estudo completo deve incluir várias. A aplicação da metodologia que utilizaram tem o problema de existirem medicamentos que apontam relações de comorbidade, e.g. a insulina e a hipertensão que é uma relação transitiva através da Diabetes Mellitus. Os autores deixam ainda a recomendação de usar esta técnica em conjunto com técnicas de gestão de bases de conhecimento, pela sua complementaridade; a desvantagem de uma é uma vantagem na outra, mas as técnicas não são mutuamente exclusivas.

A comorbidade da Perturbação de Hiperatividade e Défice de Atenção foi investigada por (Tai & Chiu, 2009). Estes autores aplicaram o algoritmo *A-priori* à base de dados de investigação do Seguro de Saúde Nacional Tailandês. O conjunto de dados tinha 18,321 casos de pacientes com menos de 10 anos de idade em 2001. Os parâmetros usados foram suporte=4% e confiança=90%. A sua conclusão aponta para a compatibilidade das regras obtidas com estudos relacionados, validando que esta técnica pode ser usada como alternativa em conjuntos de elevada cardinalidade.

Em termos de tarefas de classificação é muito relevante o uso de Árvores de Decisão, que se deve à fácil compreensão dos modelos produzidos e aos elevados resultados obtidos. Os investigadores (Chae, Kim, Tark, Park, & Ho, 2003) analisaram os fatores críticos de mortalidade de 8.405 pacientes internados usando para tal um modelo de Árvore de Decisão inferido com o algoritmo *CHAID*. Segundo os investigadores, a escolha por esta técnica deveu-se ao facto de que neste modelo é fácil retirar as regras de classificação e com elas criar um SAD usado para controlo dos indicadores de qualidade.

3.17. Diabetes Mellitus

A investigação assente em bases de dados sobre Diabetes Mellitus é extensa. De acordo com (Breault, 2001) a investigação teve início há vinte anos onde uma base de dados foi utilizada para extração de conhecimento através de *queries*. Nos anos seguintes a mesma técnica foi usada para melhoria de tratamento e como ferramenta de compreensão e divulgação entre profissionais de saúde. Em 1997 um hospital Britânico juntou os registos dos seus pacientes com os dados do

registo central dos Serviços Nacionais de Saúde de onde foi possível concluir que a análise das certidões de óbito é, só por si, insuficiente para deduzir conclusões sobre a mortalidade relacionada com a Diabetes Mellitus. Em 1998 foi criada uma base de dados sobre diabetes nos veteranos militares dos EUA que inclui as medições bioquímicas, medicamentos prescritos e internamentos hospitalares de 139.646 pacientes. Em 1999 foi criado o Registo de Diabetes do Reino da Bélgica onde era obrigatória a introdução de todos os incidentes de Diabetes Tipo 1 bem como dos familiares em primeiro grau com menos de 40 anos de idade, o que permitiu estudos epidemiológicos e genéticos.

A Diabetes Mellitus é uma doença que se prolonga no tempo e por isso a importância da análise da evolução temporal. A Agência de Saúde Local Italiana de Pavia tem recolhido dados dos seus cerca de 530.000 pacientes: (1) desde 2002, relacionados com aspetos administrativos como internamentos, prescrições farmacológicas e consultas em ambulatório e (2) desde 2007, sobre dados clínicos sobre as patologias mais prevalentes, como a Diabetes Mellitus. Os dados administrativos são, por natureza, representados como uma sequência de eventos enquanto os dados clínicos surgem como séries temporais de valores numéricos. Os dados foram pré-processados para uniformizar a sua representação e a codificação posta de acordo com *standards* da indústria. Sobre estes dados, os investigadores (Concaro, Sacchi, Cerra, & Bellazzi, 2009) desenvolveram um algoritmo de análise temporal de regras de associação com o intuito de avaliar prestação de cuidados de saúde nesta doença. Segundo os mesmos, o algoritmo é baseado no *A-priori* pois usa os parâmetros de confiança e suporte para a descoberta das regras interessantes, mas foi modificado para aplicação num contexto temporal em vez de puramente transacional.

A deteção precoce da Diabetes Mellitus de tipo 2 através de técnicas de Data Mining foi o objeto do estudo de comparação realizado pelos investigadores (Bayu, Rodiyatul, & Hermansyah, 2011). O conjunto de dados estudado do Hospital Público Indonésio de Palembang descrevia 435 casos, após pré-processamento, sendo que cerca de 80% eram de pacientes da doença e por isso a amostra não revela a taxa de prevalência real da doença. A relevância do estudo prende-se com o emprego das técnicas de SVM (*SMO*), Regras de Classificação (*IBk*), Árvores de Decisão (*J48*), Naïve Bayes e estes três últimos com *boosting* (*Adaboost*). A avaliação dos modelos foi feita com o método de *Holdout* com *10-fold cross-validation* e a métrica usada foi a acurácia. A ferramenta usada foi a Weka. O ponto mais interessante deste estudo prende-se com a mistura de avaliação quantitativa (a percentagem de acurácia) com qualitativa através da opinião de especialistas. Das 19 regras interessantes extraídas no estudo através da Árvore de Decisão, 6 foram escolhidas pelos especialistas com base na sua experiência. Os autores do estudo revelam que o *boosting* não melhorou os resultados dos modelos e que a Árvore de Decisão e as Regras classificam pior que o classificador Naïve Bayes, contrariando resultados de estudos similares, mas não revelam a razão para esse acontecimento. O *boosting* foi feito com apenas dois classificadores o que é insuficiente para que os pesos se ajustem na seleção da amostra e esta técnica é também conhecida por ser sensível ao ruído dos dados, típico dos dados em saúde. Embora algumas técnicas se revelem superiores a outras mais vezes, os estudos apontam para o facto de o resultado depender dos dados e ser, portanto, fundamental testar várias técnicas e ajustar a parametrização dos algoritmos se o objetivo for atingir o melhor resultado possível.

De acordo com a Organização Mundial de Saúde (World Health Organization, 2014), os modos de tratamento da Diabetes Mellitus são: Medicamentos, Dieta, Redução de peso, Cessação Tabágica, Exercício Físico e Insulina. Os investigadores (Aljumah, Ahmad, & Siddiqui, 2013) utilizaram regressão nas ferramentas de Data Mining da (Oracle , 2014) para predizer o modo de tratamento mais apropriado consoante a idade do paciente. As suas conclusões indicam que os modos de tratamento mais indicados para pacientes jovens (15-44 anos) são, por ordem de preferência: (1) Dieta, (2) Redução de Peso, (3) Exercício Físico e Medicamentos, *ex aequo*, (5) Cessação Tabágica e (6) Insulina. Para os pacientes seniores (45-64 anos) são: (1) Dieta, (2) Medicamentos, (3) Exercício Físico, (4) Redução de Peso, (5) Cessação Tabágica e (6) Insulina. Apesar das diferenças, recomendam a combinação dos modos de tratamento.

A Diabetes Mellitus está associada com fatores de hereditariedade; as pessoas com familiares próximos pacientes desta doença têm uma predisposição genética agravada para também virem a padecer dela. A investigação genética também usa técnicas de Data Mining para prognosticar esta patologia. Os investigadores (Ban, Heo, Oh, & Park, 2010) usaram SVM para classificar dados genéticos (sequências de ADN) no diagnóstico de Diabetes Tipo 2.

As técnicas de Data Mining são também utilizadas com vista a atingir um prognóstico mais favorável da Diabetes Mellitus. É o caso das bombas de insulina usadas no controlo e monitorização dos níveis de glucose no sangue. Os investigadores (Marling, Wiley, Bunescu, Shubrook, & Schwartz, 2012) usaram um modelo de CBR para desenvolver uma aplicação informática que permite detetar problemas na gestão da glucose no sangue. Foram usados 50 casos de 20 pacientes na sua criação. Os pacientes concordaram com os resultados obtidos e afirmaram que aceitariam as recomendações aplicáveis. Outro estudo destes investigadores foi realizado tendo por base as leituras realizadas pelas bombas. Neste foram usados 300 casos onde classificadores diferenciaram os níveis de glucose como sendo muito ou pouco variável. As classificações foram comparadas com análises realizadas por médicos. O classificador Naïve Bayes obteve o resultado mais semelhante com os dos médicos, 85% das vezes. Os investigadores foram mais longe e usaram um modelo de regressão temporal combinado com um classificador SVM para administrar insulina de forma preditiva, ou seja, o *software* da bomba que desenvolveram tomaria uma ação corretiva ao prever que o paciente estava na iminência de ter um pico de glicémia e não depois de este ocorrer. Foram desenvolvidos modelos individualizados para cada paciente e um genérico para os novos pacientes.

As técnicas de classificação e de geração de regras de associação podem ser usadas para identificar os atributos decisivos de um conjunto de dados e os seus efeitos na Diabetes Mellitus. Numa primeira fase os investigadores (Nuwangi, Oruthotaarachchi, Tilakaratna, & Caldera, VCON 2010, 2010) usaram regras de associação para determinar os fatores que de forma combinada mais significativamente contribuíam nessa doença e descobriram que ao contrário do que se acreditava, o género sexual é determinante no processo de diagnóstico. Sujeitaram o registo clínico de 10.000 pacientes no Sri Lanka a uma análise de regras de associação pelo *software* Weka do qual extraíram as 3 regras com maior valor de confiança para cada tipo de Diabetes Mellitus. Numa segunda fase (Nuwangi, Oruthotaarachchi, Tilakaratna, & Caldera, 2010) elaboraram árvores de decisão tendo por base o consequente da regra como atributo de

classificação para cada uma das 3 regras encontradas. Desta forma encontraram que a combinação de alguns atributos anteriormente desconhecidos pode ser significativa no processo de diagnóstico e descreveram em termos de valores de probabilidades a contribuição de cada atributo usando uma árvore de decisão para diagnóstico. As descobertas realizadas foram avaliadas por especialistas no domínio médico embora, como os próprios autores afirmam, careçam ainda de prova através de análise mais abrangente.

Embora a maioria dos conjuntos de dados estudados em Data Mining sobre Diabetes Mellitus se concentrem em atributos relacionados com os dados sócios-demográficos, bioquímicos e estilos de vida, a maioria também conclui que o IMC é um atributo decisivo para o diagnóstico. O estudo dos investigadores (Su, Yang, Hsu, & Chiu, 2006) demonstra a possibilidade de construir modelos de predição da Diabetes Tipo 2 através de dados antropométricos recolhidos com técnicas de imagiologia em 3 dimensões. Utilizaram técnicas de AD, RNA, Regressão Logística e Conjuntos Aproximados. Após pré-processamento, o estudo incidiu sobre dados de 6.023 pacientes, dos quais 3.076 do sexo feminino e revela os atributos físicos mais relevantes para o diagnóstico. Os autores não revelam a quantidade de pessoas pacientes da doença, somente as suas medidas antropométricas. Os modelos foram otimizados através da eliminação de atributos que contribuíam negativamente para o resultado. O treino foi realizado com 80% dos casos do conjunto de dados escolhidos de forma aleatória sendo os restantes para teste. A acurácia das técnicas AD e Conjuntos Aproximados foi superior à das outras técnicas, mas todas foram acima dos 80%. Embora usando os atributos mais comuns de um conjunto de dados de 1.487 pessoas, das quais 735 padeciam de Diabetes Mellitus, os autores (Meng, Huang, Rao, & Liu, 2012) obtiveram um resultado similar concluindo que a AD obteve melhores resultados do que o modelo de Regressão Logística e da RNA.

Apesar da maioria dos estudos usarem como variável dependente o diagnóstico da existência ou não da doença, existem outras possibilidades. O estudo de (Huang, McCullagh, Black, & Harper, 2007) incide sobre a distinção de pacientes com um bom e mau controlo da glucose no sangue através de testes bioquímicos de HbA1c. Os dados registados durante o período de 2000 a 2004, existiam no processo clínico de 3857 utentes do Hospital de Ulster, Dundonald, Belfast, Irlanda do Norte, Reino Unido. No conjunto original existiam 410 atributos pelo que foi realizada uma redução da dimensionalidade através do algoritmo *Fssmc* da sua autoria. As técnicas usadas para induzir o modelo foram Naïve Bayes, *IB1* e *C4*.5 e foram testados diferentes atributos. O melhor resultado (95,26%) foi obtido com *IB1* e 8 atributos, mas notavelmente os autores não referem o *software* usado no teste.

Contrariando o facto da maioria dos estudos existentes serem baseados em técnicas de classificação através de AD, RNA, SVM ou IBL, (Suh & Vudumula, 2011) usaram clustering aglomerativo para desenvolver uma ferramenta de diagnóstico da Diabetes Mellitus. Realizaram a descrição da base de dados em tabelas de atributos, tabela de conceitos e tabelas de relações. A tabela de conceitos descreve os diferentes tipos da doença. A tabela de atributos descreve os sintomas, os estilos de vida e recomendações de prevenção da doença. A tabela de relações descreve o mapeamento entre a tabela de conceitos e as três de atributos. Através de clustering aglomerativo foi possível calcular a probabilidade de ocorrência dos sintomas relativamente aos

diferentes tipos da doença, bem como dos estilos de vida e as recomendações de prevenção. Foi ainda calculada a probabilidade de ocorrência de sintomas, estilos de vida e tipos de doença combinados. Apesar dos bons resultados, esta técnica é semiautomática visto que as hierarquias conceptuais (associação das tabelas) são geradas por especialistas do domínio.

Na prática clínica os médicos seguem recomendações oriundas da investigação. Contudo, surgem casos não descritos nas recomendações para os quais os médicos necessitam de prescrever tratamento ou terapêutica medicamentosa. Na sua essência as recomendações são regras de classificação extraídas através de uma AD. Tendo por base esta premissa, os investigadores (Toussi, Lamy, Toumelin, & Venot, 2009) compararam as recomendações clínicas da Autoridade Nacional de Saúde Francesa para a Diabetes Mellitus tipo 2 com uma AD realizada com a experiência dos clínicos e outra ainda extraída com o algoritmo *C5.0* de uma base de dados com 463 pacientes do Hospital Avicenne de Bobigny, França da qual foram criadas 27 regras. Embora limitado pela cardinalidade do conjunto de dados, foi possível concluir que este método pode ser usado para complementar a criação das recomendações e como avaliação de desvios entre a recomendação e as decisões terapêuticas realizadas.

É do conhecimento médico a relação que existe entre a Diabetes Mellitus e a Hipertensão; os pacientes de Diabetes Mellitus têm um risco acrescido (quádruplo) de dano cardiovascular (Parthiban, Rajesh, & Srivatsa, 2011). Estes autores construíram um modelo de diagnóstico para a Hipertensão construído a partir de atributos de um conjunto de dados sobre a Diabetes Mellitus. A técnica que utilizaram foi um classificador Naïve Bayes, segundo afirmam, pela vantagem de este requerer poucos exemplos de treino para estimação dos parâmetros média e variância. Uma vez que os atributos são considerados independentes, somente a variância necessita de ser conhecida para cada classe. O conjunto de dados foi extraído da base de dados do Instituto de investigação em Diabetes Mellitus de Chennai, India e possui instâncias de 500 pacientes. Os atributos são aqueles mais verificados nos estudos científicos, salientando a tensão arterial e os níveis de colesterol. O *software* usado foi o Weka com *10-fold Cross-Validation* e a acurácia do modelo obtido foi de 74%. O contributo essencial deste estudo é a ideia de usar os atributos de uma doença para diagnosticar outra doença onde exista uma correlação forte.

3.18. PIMA

O conjunto de dados de Diabetes das Índias Pima do Arizona nos EUA, foi desenvolvida pelo (National Institute of Diabetes and Digestive and Kidney Diseases, 2013) e cedida em 1990 ao repositório da Universidade da Califórnia, Irvine. O conjunto de dados descreve 768 pacientes do género sexual feminino com idades compreendidas entre 21 e 81 anos, divididos em 500 casos que não padecem de Diabetes Mellitus e os restantes sim. O conjunto de dados tem 9 atributos:

1. N.º de gravidezes;
2. Plasma do Teste Oral de Tolerância à Glicose de 2 horas;
3. TAD;
4. Espessura da dobra cutânea tricipital;
5. Nível de Insulina em soro de 2 horas;
6. Índice de Massa Corporal;

7. Função "*pedigree*" Diabetes;
8. Idade;
9. Surgimento de Diabetes no espaço de 5 anos. Esta é a classe de predição e pode assumir o valor 0=não diabético e 1=diabético.

Apesar de ser o conjunto de dados mais usado na comunidade científica, este padece de alguns problemas. De acordo com (Knowler, et al., 1978) e (Hanson, et al., 1998) este conjunto de dados refere-se a uma população específica com uma prevalência acima do habitual devido a uma predisposição genética. Aliado a este facto está ainda o problema de os registos não estarem equilibrados entre classes. Se assumirmos um modelo que classifique qualquer caso a um paciente não diabético temos uma acurácia de 65%, o que é equivalente a dizer que temos um erro de classificação inaceitável de 35%. Também em relação aos dados existem alguns erros a notar: 5 pacientes têm no nível de glicose o número 0, tal como 11 pacientes no índice de massa corporal, 28 pacientes na TAD e 140 pacientes nos níveis de insulina em soro. Não sendo possíveis tais valores, deduz-se que os valores em falta foram preenchidos com 0 provocando um erro. Se eliminarmos os registos com este erro, o conjunto de dados tem 392 casos.

O conjunto de dados de Diabetes das Índias Pima tem sido alvo de extensa análise e uso na investigação científica em Data Mining.

Quadro 3:1 - Investigação baseada no conjunto de dados PIMA – técnicas e resultados

Autor (es)	**Metodologia**	**Algoritmo**	**Acurácia**
Smith et al. 1988	Criação de Rede Neuronal com n=576 casos aleatórios de treino	ADAP	76%
Wahba, et al. 1992	Aplicação de 2 algoritmos de Regressão para n=500 casos de treino	PSA	72%
		GLIM	74%
Quinlan, 1993	Modelo baseado na Árvore de Decisão	C4.5	71,1%
Michie, et al. 1994	Teste de 22 algoritmos com 12-fold cross-validation:	Discrim	77,5%
		Quaddisc	73,8%
		Logdisc	77,7%
		SMART	76,8%
		ALLOC80	69,9%
		k-NN	67,6%
		CASTLE	74,2%
		CART	74,5%
		IndCART	72,9%
		NewID	71,1%
		AC^2	72,4%
		Baytree	72,9%
		NaiveBay	73,8%
		CN2	71,1%
		C4.5	73%
		Itrule	75,5%
		Cal5	75%
		Kohonen	72,7%
		DIPOL92	77,6%

		Backprop	75,2%
		RBF	75,7%
		LVQ	72,8%
Oates, 1994	Aplicação de um algoritmo em 2/3 dos casos usados para treino	MSDD	71,33%
Bioch, et al. 1996	Comparação de 2 modelos de redes neuronais com n=500 casos de treino	Std RNA	75,4%
		Bayes RNA	79,5%
Ripley, 1996	Teste de 7 algoritmos com n=532 casos totais, dos quais 200 casos de treino ou n=300 casos de treino para os algoritmos capazes de lidar com atributos com zero	Logistic Regression	80,2%
		MARS	77,4%
		PPR	77,4%
		RNA	77%
		k-NN (k=9)	75,3%
		OLVQ	78,9%
		CART	77,7%
Carpenter & Markuzon, 1998	Criação de modelo de Rede Neuronal (ARTMAP-IC) e comparação com outros algoritmos, para n=576 casos aleatórios de treino.	ARTMAP-IC	81%
		Basic fuzzy ARTMAP	66%
		Logistic Regression	77%
		k-NN	77%
Khan, 1998	Rede Neuronal com um conjunto de dados balanceado e que inclui todos os casos diabéticos, apesar dos valores em falta. n=268 casos de treino	MFN	78%
Eklund & Hoang, 1998	Teste de 5 algoritmos com 80% dos casos usados para treino, resto para teste	C4.5	71,02%
		C4.5 rules	71,55%
		ITI	73,16%
		LMDT	73,51%
		CN2	72,19%
Liu, 1998	Junção de modelo de classificação com regras de associação (CAR) e comparação com outro algoritmo (C4.5 rules)	CAR	73,1%
		C4.5 rules	75,5%
King, et al., 1998	Teste de 14 algoritmos, retiraram os casos com os valores de insulina em falta ficando com n=532.	CART	76%
		Scenario	30%
		See5	73%
		S-Plus	79%
		WizWhy	74%
		DataMind	69%
		DMSK	67%
		NeuroShell2-	77%
		Neural	81%
		PcOLPARS	80%
		PRW	77%
		MQ Expert	78%
		NeuroShell2-	81%
		PolyNet	78%
		Gnosis	
		KnowledgeMiner	

Wang, 2000	Algoritmo de classificação pela agregação de padrões emergentes	CAEP	75%
Breault, 2001	Conjuntos Aproximados	Johnson Reducer	73,9%
(Temurtas, Yumusak, & Temurtas, 2009)	Rede Neuronal com função sigmóide não linear com 10-fold cross-validation	MLNN com LM PNN	79,62% 78,05%
	Rede Neuronal com função sigmóide com validação tradicional (um treino e um teste)	MLNN com LM PNN	82,37% 78,13%
(Patil, Joshi, & Toshniwal, 2010)	Modelo hibrido com k-Means e C4.5. Eliminaram registos cuja classificação pelo k-Means estivesse incorreta. Eliminaram registos com zeros nos atributos e normalizaram valores. Usaram 10 fold-cross validation no C4.5.	HPM	92,38%
(Ganji & Abadeh, 2011)	Regras de Classificação difusas geradas com algoritmo de colónia de formigas com 10-fold cross validation	FCS-ANTMINER	84,24%
(AlJarullah, 2011)	Árvore de Decisão com 10-fold cross validation	J48	78,18%
(Karthikeyani, Begum, Tajudin, & Begam, 2012)	Teste de 10 algoritmos de classificação com 10-fold cross validation	C4.5	86%
		SVM	74,8%
		kNN	78%
		PNN	67%
		BLR	75%
		MLR	75%
		CRT	85%
		CS-CRT	86%
		PLS-DA	76%
		PLS-LDA	73%

Conforme se pode observar, os valores do rácio de acurácia obtidos estão entre os 66% e os 92,38%. Os modelos híbridos apresentam bons resultados, mas torna-se óbvio que se retirarmos os registos de mais difícil classificação os resultados serão melhores. Poderemos estar perante um caso de *overfitting* do modelo. Os casos de difícil classificação também existem na realidade.

Pela metodologia usada nos estudos observa-se que não são usados os mesmos conjuntos de dados na produção e teste dos modelos, dificultando a comparação. Os resultados são na maior parte obtidos com uma só execução do algoritmo o que aumenta a possibilidade de o modelo ter sido obtido por mero acaso pela escolha dos casos e não é um indicador qualitativo do algoritmo usado.

Modelos preditivos podem ser ferramentas valiosas, mas algumas condições devem ser acauteladas: (1) o modelo deve incluir todos os dados relevantes, (2) o modelo deve ser testado numa amostra independente e (3) o modelo deve ser relevante e útil. Muitos dos modelos existentes na literatura não satisfazem estas condições (Meng, Huang, Rao, & Liu, 2012).

Fica assim demonstrada a diversidade de técnicas usadas na investigação e possíveis para a aplicação nos conjuntos de dados de saúde. Seria impraticável testar todas as técnicas em todos os conjuntos de dados. De uma maneira geral, as conclusões dos estudos apontam para a inexistência de uma técnica que seja melhor que as demais em todos os conjuntos de dados e que por isso, consoante a tarefa seja necessária a construção de modelos usando todas a técnicas possíveis e a comparação dos seus resultados.

4. CLASSIFICAÇÃO PARA DIAGNÓSTICO

A pesquisa médica recorre com muita frequência à tarefa de classificação porque esta permite a realização de diagnósticos, ação essencial em medicina. Por esse motivo importa conhecer os detalhes da classificação e no que à sua aplicação diz respeito.

Em seguida é demonstrado um estudo comparativo de técnicas e ferramentas de *software* de Data Mining (KNIME, Orange, RapidMiner e Weka), selecionadas em função da sua capacidade para classificar dados (Borges, Marques, & Bernardino, 2013).

4.1. Revisão da literatura

Considerando uma oferta de preços entre 75 e 25.000 USD, (King, et al., 1998) realizaram um estudo comparativo de catorze ferramentas. O processo de avaliação foi realizado por três tipos de grupos de utilizadores: (1) quatro estudantes universitários inexperientes em Data Mining, (2) um estudante licenciado com experiência em Data Mining, e (3) um consultor profissional em Data Mining. Os testes foram realizados com quatro Datasets. Para testar a flexibilidade e capacidade das ferramentas, os seus tipos de saída variaram: Dois Datasets com classificações binárias (um com dados em falta), um Dataset multiclasse e um Dataset sem ruído. Dois terços dos casos foram escolhidos de forma aleatória para o Trainingset e o outro terço para o Testset. Os autores desenvolveram uma lista de critérios para avaliação de ferramentas de Data Mining. As ferramentas corriam sobre sistemas operativos Microsoft Windows 95, NT ou Macintosh 7.5 e usavam diferentes tipos de técnicas: árvores, regras e redes. Os resultados demonstram um relatório técnico que detalha o procedimento de avaliação e a pontuação de todos os componentes para todos os critérios. Os autores também mostram que a escolha de uma ferramenta depende de uma pontuação ponderada de várias categorias tais como orçamento e experiência de utilizador. Os autores concluem também que o preço da ferramenta está relacionado com a sua qualidade.

O trabalho realizado por (Giraud-Carrier & Povel, 2003) descreve um esquema geral para a caraterização de ferramentas de *software* para Data Mining. Os autores descrevem um modelo para a caraterização através de um número de dimensões complementares, em conjunto com uma base de dados de 41 das mais populares ferramentas de Data Mining. A proposta de caraterização, orientada para negócio, é definida segundo o objetivo de negócio, tipo de modelo, caraterísticas dependentes do processo, caraterísticas da *interface* do utilizador, requisitos de sistema e informação do fabricante. Usando estas, os autores avaliam as 41 ferramentas. Os autores concluem que com a ajuda de um esquema padronizado e uma correspondente base de dados, os utilizadores são capazes de selecionar uma ferramenta, no que diz respeito à sua capacidade de ir ao encontro dos objetivos de negócio.

O trabalho realizado por (Carey & Marjaniemi, 1999) apresenta um esquema de trabalho (*framework*) para avaliação de ferramentas de Data Mining e descreve a metodologia para a sua aplicação. Esta metodologia é baseada em experiências em primeira mão em Data Mining usando

Datasets comerciais de uma variedade de indústrias. A experiência sugere quatro categorias para a avaliação das ferramentas: Performance, funcionalidade, usabilidade e suporte de atividades auxiliares. Os autores demonstram que a metodologia de avaliação tira vantagem de conceitos de matrizes de decisão para objetivar um processo inerentemente subjetivo. Usando uma aplicação de folha de cálculo, o esquema de trabalho proposto é facilmente automatizado e assim mais facilmente posto em prática e de verossímil empregabilidade. Os autores mostram que não existe uma ferramenta que se destaque das outras, em todos os aspetos relacionados com as técnicas de Data Mining, mas que existem algumas ferramentas que, entre si, partilham a liderança neste segmento de mercado.

No trabalho realizado por (Abbott, Matkovsky, & Elder, 1998) são comparadas cinco das mais aclamadas ferramentas de Data Mining para aplicação em deteção de fraudes (outliers). Os autores empregaram duas fases na seleção sucedidas por uma avaliação pormenorizada. Para a primeira fase, mais de 40 ferramentas foram avaliadas em seis diferentes qualidades. As 10 melhores passaram para a segunda fase de seleção onde foram avaliadas em mais caraterísticas adicionais. Depois de serem selecionadas as 10 melhores ferramentas, os autores usaram critérios periciais e reavaliaram as suas caraterísticas donde resultaram as 5 melhores selecionadas para uma avaliação pormenorizada. As ferramentas selecionadas foram SPSS Clementine, Darwin (hoje Oracle Data Miner), SAS Enterprise Miner, IBM Intelligent Miner e PRW. As propriedades das ferramentas avaliadas incluíam as áreas de conformidade com a arquitetura cliente-servidor, capacidades de automação, a abrangência dos algoritmos implementados, a facilidade de utilização e a acurácia na deteção de fraudes nos dados de teste. Os resultados mostram que todas as 5 ferramentas avaliadas pelos autores possuíam excelentes propriedades, mas que cada uma seria melhor que as demais em ambientes específicos. Os autores concluíram que o Intelligent Miner tinha a vantagem em ser o líder de mercado. O Clementine é excelente na ajuda que fornece e em facilidade de utilização. O Enterprise Miner será o mais ajustado para um ambiente de análise de dados. O Darwin foi o melhor quando a largura de banda era cara. O PRW é a melhor escolha quando não é obvio qual o algoritmo mais apropriado ou quando os analistas estão familiarizados com folhas de cálculo, visto este basear-se nelas.

No trabalho realizado por (Lee & Hen, 2008) foram comparadas e analisada a *performance* de 5 ferramentas: IBM Intelligent Miner, SPSS Clementine, SAS Enterprise Miner, Oracle Data Miner e Microsoft Business Intelligence Development Studio. Foram usadas 38 métricas para comparar a *performance* das ferramentas. Os dados foram minerados segundo diversas técnicas e algoritmos suportados pelas ferramentas, incluindo algoritmos de classificação, regressão, segmentação e associação. Os autores sugerem na conclusão que deveria ser criado um *middleware* capaz de interagir com todas as ferramentas e assim ser capaz de aproveitar as caraterísticas mais fortes de cada uma.

No trabalho realizado por (Wahbeh, Al-Radaideh, Al-Kabi, & Al-Shawakfa, 2010), os autores comparam 4 ferramentas de Data Mining de utilização livre para revelar aquela que melhor se ajusta à tarefa de classificação. As ferramentas testadas foram KNIME, Orange, Tanagra e Weka. Para realizar a avaliação, os autores optaram pela métrica de precisão. A conclusão obtida foi a que não existe uma ferramenta que seja melhor que as outras em todos os casos; o resultado

depende muito dos dados do Dataset utilizado e da forma como os algoritmos foram implementados nas ferramentas. Em geral, o WEKA revelou vantagem sobre os restantes atingindo valores de precisão maiores em mais casos, mesmo quando foi alterado o método de amostragem de *percentage split* para *cross-validation*.

4.2. Metodologia

A preparação do estudo consiste em três etapas preparatórias: (1) A seleção das ferramentas de Data Mining a testar, (2) a seleção dos Datasets a empregar e (3) a seleção dos algoritmos de classificação a avaliar. O estudo consistirá em testar todas as combinações possíveis com os Datasets, as ferramentas, as técnicas, os modos de particionamento e algoritmos, e avaliar a sua classificação pela métrica da acurácia.

4.3. Ferramentas

Na primeira etapa da metodologia é realizada a seleção das ferramentas a testar. Tendo por base as 5 melhores ferramentas de Data Mining (Auza, 2010), a escolha obedeceu ao critério de *user-friendliness*. Todas possuem uma *interface* gráfica, mas somente 4 são possíveis de utilizar sem recurso a *scripting*; em JHepWork é imprescindível conhecer a linguagem de programação (neste caso, Jython). Foram selecionadas para estudo aquelas que um analista (não programador) seja capaz de utilizar:

- KNIME 2.6.0 (Knime.com AG, 2012) – O nome vem de Konstanz Information Miner pois foi inicialmente desenvolvido na Universidade de Konstanz, Alemanha. É um *software* escrito em Java, com uma *interface* gráfica baseada no IDE eclipse. Permite executar tarefas de Data Mining e Machine Learning tais como pré-processamento, classificação, regressão, segmentação, associação e visualização. Os utilizadores podem criar fluxos (*pipelines*) graficamente arrastando os nós de um repositório para a janela de projeto, definir o fluxo e as propriedades para cada nó. Após o processamento, podem ser obtidos visualmente os resultados da tarefa através de nós de avaliação.
- Orange 2.6a1 (Demšar, Curk, & Erjavec, 2013) - É um *software* modular baseado em componentes para Data Mining e Machine Learning que possui uma *interface* gráfica de programação. Contém um conjunto completo de módulos para tarefas de pré-processamento, visualização, classificação, regressão, avaliação, segmentação e associação. Está escrito em C++ e Python. A sua interface gráfica permite que os componentes (*widgets*) sejam arrastados para uma tela, sejam interligados e definidos os seus parâmetros, produzindo assim o fluxo necessário à conclusão da tarefa.
- RapidMiner 4.6.000 (Rapidminer, 2012) – Antes conhecido por YALE (Yet Another Learning Environment), é um ambiente para experimentação de tarefas de Data Mining e Machine Learning, que é utilizado tanto para investigação como em aplicações reais. Permite que as experiências sejam feitas através de uma *interface* gráfica e que os projetos sejam compostos por um grande número de operadores, arbitrariamente aninhados e detalhados em arquivos XML. O RapidMiner oferece

mais de 500 operadores de todos os principais procedimentos de Machine Learning e também combina sistemas de aprendizagem e avaliadores de atributos do ambiente Weka. Está disponível como uma ferramenta independente para a Análise de Dados e como um motor de Data Mining que pode ser integrado noutros produtos.

- Weka 3.6 (Machine Learning Group at the University of Waikato, 2012) – é uma *suite* de *software* para tarefas de Data Mining e Machine Learning, realizada pela Universidade de Waikato, Nova Zelândia. Contém uma grande coleção de algoritmos escritos em Java para tarefas de pré-processamento de dados, regressão, classificação, segmentação, associação e visualização. As técnicas são baseadas na hipótese que os dados estão disponíveis em *flat files* ou bases de dados relacionais. A sua *interface* de utilizador principal é o Explorer, mas a mesma funcionalidade pode ser acedida por uma linha de comandos.

4.4. Datasets

O segundo passo da metodologia consiste em definir os Datasets a testar. Todos os Datasets escolhidos e que a seguir se descrevem foram descarregados do repositório UCI da Universidade da Califórnia, Irvine, California, EUA (Bache & Lichman, 2013).

4.4.1. Adult

O Dataset é uma extração de dados do censo realizado em 1994 nos EUA, com incidência em dados socioeconómicos. A classe é binária e tem o objetivo de predizer se o rendimento excede ou não os 50.000 USD/ano. A distribuição da classe é de aproximadamente ¼ para rendimentos superiores a 50.000 USD/ano e ¾ para rendimentos inferiores. O Dataset tem 7% de instâncias com valores em falta em três dos atributos: workclass (1836), occupation (1843) e native-country (583). Os dados estatísticos dos atributos são:

Quadro 4:1 - Dados estatísticos do Dataset Adult

Atributo	Tipo	Estatística	Limites/Distribuição
age	real	média = 38,582 +/- 13,640	[17,000 - 90,000]
workclass	nominal	moda = Private (22696), menor = Never-worked (7)	Private (22696), Self-emp-not-inc (2541), Self-emp-inc (1116), Federal-gov (960), Local-gov (2093), State-gov (1298), Without-pay (14), Never-worked (7)
fnlwgt	real	média = 189.778,367 +/- 105.549,978	[12.285,000 - 1.484.705,000]
education	nominal	moda = HS-grad (10501), menor = Preschool (51)	Bachelors (5355), Some-college (7291), 11th (1175), HS-grad (10501), Prof-school (576), Assoc-acdm (1067), Assoc-voc (1382), 9th (514), 7th-8th (646), 12th (433), Masters (1723), 1st-4th (168), 10th (933), Doctorate (413), 5th-6th (333), Preschool (51)
education-num	real	média = 10,081 +/- 2,573	[1,000 - 16,000]

marital-status	nominal	moda = Married-civ-spouse (14976), menor = Married-AF-spouse (23)	Married-civ-spouse (14976), Divorced (4443), Never-married (10683), Separated (1025), Widowed (993), Married-spouse-absent (418), Married-AF-spouse (23)
occupation	nominal	moda = Prof-specialty (4140), menor = Armed-Forces (9)	Tech-support (928), Craft-repair (4099), Other-service (3295), Sales (3650), Exec-managerial (4066), Prof-specialty (4140), Handlers-cleaners (1370), Machine-op-inspct (2002), Adm-clerical (3770), Farming-fishing (994), Transport-moving (1597), Priv-house-serv (149), Protective-serv (649), Armed-Forces (9)
relationship	nominal	moda = Husband (13193), menor = Other-relative (981)	Wife (1568), Own-child (5068), Husband (13193), Not-in-family (8305), Other-relative (981), Unmarried (3446)
race	nominal	moda = White (27816), menor = Other (271)	White (27816), Asian-Pac-Islander (1039), Amer-Indian-Eskimo (311), Other (271), Black (3124)
sex	nominal	moda = Male (21790), menor = Female (10771)	Female (10771), Male (21790)
capital-gain	real	média = 1.077,649 +/- 7.385,292	[0,000 - 99.999,000]
capital-loss	real	média = 87,304 +/- 402,960	[0,000 - 4.356,000]
hours-per-week	real	média = 40,437 +/- 12,347	[1,000 - 99,000]
native-country	nominal	moda = United-States (29170), menor = Holand-Netherlands (1)	United-States (29170), Cambodia (19), England (90), Puerto-Rico (114), Canada (121), Germany (137), Outlying-US(Guam-USVI-etc) (14), India (100), Japan (62), Greece (29), South (80), China (75), Cuba (95), Iran (43), Honduras (13), Philippines (198), Italy (73), Poland (60), Jamaica (81), Vietnam (67), Mexico (643), Portugal (37), Ireland (24), France (29), Dominican-Republic (70), Laos (18), Ecuador (28), Taiwan (51), Haiti (44), Columbia (59), Hungary (13), Guatemala (64), Nicaragua (34), Scotland (12), Thailand (18), Yugoslavia (16), El-Salvador (106), Trinadad&Tobago (19), Peru (31), Hong (20), Holand-Netherlands (1)

O atributo “fnlwgt” é um índice que demonstra o corelacionamento de aspetos socioeconómicos com a população vizinha, ou seja, pessoas com caraterísticas demográficas similares deverão apresentar valores similares. De acordo com a informação prestada pelos autores, os testes de classificação apontam para uma acurácia de cerca de 85%.

4.4.2. Breast-cancer

O Dataset foi obtido no Centro Médico Universitário do Instituto de Oncologia de Ljubljana, Jugoslávia. Incide sobre o domínio das ciências da vida e especificamente sobre o cancro da mama. A classe é binária e tem o objetivo de predizer recorrências do cancro. A distribuição da classe é de aproximadamente ¼ para casos de recorrência e ¾ para não recorrência. O Dataset

tem 3% de instâncias com valores em falta em 2 dos atributos: node-caps (8) e breast-quad (1). Os dados estatísticos dos atributos são:

Quadro 4:2 - Dados estatísticos do Dataset Breast-cancer

Atributo	**Tipo**	**Estatística**	**Limites/Distribuição**
age	nominal	moda = 50-59 (96), menor = 20-29 (1)	20-29 (1), 30-39 (36), 40-49 (90), 50-59 (96), 60-69 (57), 70-79 (6)
menopause	nominal	moda = premeno (150), menor = lt40 (7)	ge40 (129), lt40 (7), premeno (150)
tumor-size	nominal	moda = 30-34 (60), menor = 45-49 (3)	0-4 (8), 5-9 (4), 10-14 (28), 15-19 (30), 20-24 (50), 25-29 (54), 30-34 (60), 35-39 (19), 40-44 (22), 45-49 (3), 50-54 (8)
inv-nodes	nominal	moda = 0-2 (213), menor = 24-26 (1)	0-2 (213), 3-5 (36), 6-8 (17), 9-11 (10), 12-14 (3), 15-17 (6), 24-26 (1)
node-caps	nominal	moda = no (222), menor = yes (56)	no (222), yes (56)
deg-malig	nominal	moda = 2 (130), menor = 1 (71)	1 (71), 2 (130), 3 (85)
breast	nominal	moda = left (152), menor = right (134)	left (152), right (134)
breast-quad	nominal	moda = left_low (110), menor = central (21)	central (21), left_low (110), left_up (97), right_low (24), right_up (33)
irradiat	nominal	moda = no (218), menor = yes (68)	no (218), yes (68)

De acordo com a informação prestada pelos autores, os testes de classificação apontam para uma acurácia de cerca de 78%.

4.4.3. Car Evaluation

O Dataset descreve veículos automóveis pelo seu preço, tecnologia e conforto. A classe é quaternária e tem o objetivo de predizer a aceitabilidade. A distribuição da classe é de aproximadamente 70% para inaceitável, 22% para aceitável, 4% para bom e 4% para muito bom. O Dataset não tem valores em falta. Os dados estatísticos dos atributos são:

Quadro 4:3 - Dados estatísticos do Dataset Car Evaluation

Atributo	**Tipo**	**Estatística**	Limites/Distribuição
buying	nominal	moda = v-high (432), menor = v-high (432)	v-high (432), high (432), med (432), low (432)
maint	nominal	moda = v-high (432), menor = v-high (432)	v-high (432), high (432), med (432), low (432)
doors	nominal	moda = 2 (432), menor = 2 (432)	2 (432), 3 (432), 4 (432), 5-more (432)
persons	nominal	moda = 2 (576), menor = 2 (576)	2 (576), 4 (576), more (576)
lugboot	nominal	moda = small (576), menor = small (576)	small (576), med (576), big (576)
safety	nominal	moda = low (576), menor = low (576)	low (576), med (576), high (576)

4.4.4. Credit Approval

O Dataset descreve situações de análise financeira para concessão de cartões de crédito. Por motivos de proteção de dados, os autores optaram por descrever os atributos através de códigos e não é conhecido o seu significado. De qualquer maneira, este Dataset é interessante visto que possui uma mistura de tipos de variáveis: contínuas, nominais com baixo valor e nominais com elevado valor. A classe é binária e tem o objetivo de predizer a concessão do crédito. A distribuição da classe é de aproximadamente 56% para os casos negativos e 44% para os positivos. O Dataset tem 5% de instâncias com valores em falta em 7 dos atributos: A1(12), A2(12), A4(6), A5(6), A6(9), A7(9) e A14(13). Os dados estatísticos dos atributos são:

Quadro 4:4 - Dados estatísticos do Dataset Credit Approval

Atributo	Tipo	Estatística	Limites/Distribuição
A1	nominal	moda = b (468), menor = a (210)	a (210), b (468)
A2	real	média = 31,568 +/- 11,958	[13,750 - 80,250]
A3	real	média = 4,759 +/- 4,978	[0.000 - 28.000]
A4	nominal	moda = u (519), menor = l (2)	l (2), u (519), y (163)
A5	nominal	moda = g (519), menor = gg (2)	g (519), gg (2), p (163)
A6	nominal	moda = c (137), menor = r (3)	aa (54), c (137), cc (41), d (30), e (25), ff (53), i (59), j (10), k (51), m (38), q (78), r (3), w (64), x (38)
A7	nominal	moda = v (399), menor = o (2)	bb (59), dd (6), ff (57), h (138), j (8), n (4), o (2), v (399), z (8)
A8	real	média = 2,223 +/- 3,347	[0,000 - 28,500]
A9	nominal	moda = t (361), menor = f (329)	f (329), t (361)
A10	nominal	moda = f (395), menor = t (295)	f (395), t (295)
A11	real	média = 2,400 +/- 4,863	[0,000 - 67,000]
A12	nominal	moda = f (374), menor = t (316)	f (374), t (316)
A13	nominal	moda = g (625), menor = p (8)	g (625), p (8), s (57)
A14	real	média = 184,015 +/- 173,807	[0,000 - 2.000,000]
A15	real	média = 1.017,386 +/- 5.210,103	[0,000 - 100.000,000]

4.4.5. Iris

Este Dataset pertence ao domínio da Botânica, um dos ramos da Biologia. Trata-se da descrição de uma subespécie de tulipas através de alguns dos seus atributos físicos observáveis. A classe é ternária e tem o objetivo de predizer o sub-tipo da planta. A distribuição da classe é de exatamente 1/3 para cada valor da classe. O Dataset não tem instâncias com valores em falta. Os dados estatísticos dos atributos são:

Quadro 4:5 - Dados estatísticos do Dataset Iris

Atributo	Tipo	Estatística	Limites/Distribuição
sepal length	real	média = 5,843 +/- 0,828	[4,300 - 7,900]
sepal width	real	média = 3,054 +/- 0,434	[2,000 - 4,400]
petal length	real	média = 3,759 +/- 1,764	[1,000 - 6,900]
petal width	real	média = 1,199 +/- 0,763	[0,100 - 2,500]
iris	nominal	moda e menor = todas por igual	Iris-setosa(50), Iris-versicolor(50), Iris-virginica(50)

Estes dados são particularmente interessantes para exercícios de classificação por terem uma classe linearmente separável das outras duas, mas estas duas não são linearmente separáveis uma da outra. Esta particularidade e a simplicidade de interpretação levam a que este Dataset seja o mais estudado na literatura científica em Data Mining.

4.4.6. Lung-cancer

O Dataset incide sobre o domínio das ciências da vida. Por motivos de proteção de dados, os autores optaram por descrever os atributos através de códigos e não é conhecido o seu significado. A classe é ternária e tem o objetivo de predizer um de 3 tipos de cancro no pulmão. A distribuição da classe é de aproximadamente 28% para o tipo 1, 41% para o tipo 2 e 31% para o tipo 3. O Dataset tem cerca de 16% de instâncias com valores em falta em 2 dos atributos: a4(4) e a 38(1). Os dados estatísticos dos atributos são:

Quadro 4:6 - Dados estatísticos do Dataset Lung-cancer

Atributo	Tipo	Estatística	Limites/Distribuição
a1	nominal	moda = 0 (31), menor = 1 (1)	0 (31), 1 (1)
a2	nominal	moda = 2 (18), menor = 1 (1)	1 (1), 2 (18), 3 (13)
a3	nominal	moda = 3 (13), menor = 0 (4)	0 (4), 1 (4), 2 (11), 3 (13)
a4	nominal	moda = 1 (15), menor = 0 (1)	0 (1), 1 (15), 2 (12)
a5	nominal	moda = 0 (23), menor = 1 (9)	0 (23), 1 (9)
a6	nominal	moda = 2 (14), menor = 1 (6)	1 (6), 2 (14), 3 (12)
a7	nominal	moda = 2 (14), menor = 1 (7)	1 (7), 2 (14), 3 (11)
a8	nominal	moda = 3 (18), menor = 1 (5)	1 (5), 2 (9), 3 (18)
a9	nominal	moda = 1 (29), menor = 2 (1)	1 (29), 2 (1), 3 (2)
a10	nominal	moda = 1 (20), menor = 3 (1)	1 (20), 2 (11), 3 (1)
a11	nominal	moda = 1 (19), menor = 3 (3)	1 (19), 2 (10), 3 (3)
a12	nominal	moda = 1 (15), menor = 3 (2)	0 (11), 1 (15), 2 (4), 3 (2)
a13	nominal	moda = 2 (12), menor = 3 (9)	1 (11), 2 (12), 3 (9)
a14	nominal	moda = 2 (14), menor = 1 (6)	1 (6), 2 (14), 3 (12)
a15	nominal	moda = 2 (13), menor = 1 (6)	1 (6), 2 (13), 3 (13)
a16	nominal	moda = 1 (23), menor = 3 (1)	1 (23), 2 (8), 3 (1)
a17	nominal	moda = 2 (28), menor = 1 (4)	1 (4), 2 (28)
a18	nominal	moda = 2 (28), menor = 1 (4)	1 (4), 2 (28)
a19	nominal	moda = 0 (24), menor = 1 (1)	0 (24), 1 (1), 2 (7)
a20	nominal	moda = 2 (16), menor = 1 (3)	0 (13), 1 (3), 2 (16)
a21	nominal	moda = 2 (24), menor = 1 (8)	1 (8), 2 (24)
a22	nominal	moda = 2 (26), menor = 1 (6)	1 (6), 2 (26)
a23	nominal	moda = 2 (24), menor = 1 (8)	1 (8), 2 (24)
a24	nominal	moda = 1 (24), menor = 3 (3)	1 (24), 2 (5), 3 (3)
a25	nominal	moda = 2 (19), menor = 3 (2)	1 (11), 2 (19), 3 (2)
a26	nominal	moda = 2 (12), menor = 3 (9)	1 (11), 2 (12), 3 (9)
a27	nominal	moda = 2 (22), menor = 3 (10)	2 (22), 3 (10)
a28	nominal	moda = 2 (22), menor = 1 (4)	1 (4), 2 (22), 3 (6)
a29	nominal	moda = 2 (17), menor = 1 (5)	1 (5), 2 (17), 3 (10)
a30	nominal	moda = 1 (22), menor = 3 (4)	1 (22), 2 (6), 3 (4)
a31	nominal	moda = 3 (19), menor = 2 (4)	1 (9), 2 (4), 3 (19)

a32	nominal	moda = 3 (19), menor = 2 (4)	1 (9), 2 (4), 3 (19)
a33	nominal	moda = 3 (26), menor = 1 (3)	1 (3), 2 (3), 3 (26)
a34	nominal	moda = 1 (14), menor = 3 (5)	1 (14), 2 (13), 3 (5)
a35	nominal	moda = 1 (16), menor = 3 (4)	1 (16), 2 (12), 3 (4)
a36	nominal	moda = 2 (20), menor = 3 (1)	1 (11), 2 (20), 3 (1)
a37	nominal	moda = 1 (14), menor = 3 (4)	1 (14), 2 (14), 3 (4)
a38	nominal	moda = 2 (18), menor = 3 (2)	1 (11), 2 (18), 3 (2)
a39	nominal	moda = 2 (25), menor = 3 (3)	1 (4), 2 (25), 3 (3)
a40	nominal	moda = 2 (26), menor = 1 (3)	1 (3), 2 (26), 3 (3)
a41	nominal	moda = 2 (15), menor = 3 (2)	1 (15), 2 (15), 3 (2)
a42	nominal	moda = 2 (18), menor = 3 (3)	1 (11), 2 (18), 3 (3)
a43	nominal	moda = 2 (28), menor = 1 (1)	1 (1), 2 (28), 3 (3)
a44	nominal	moda = 2 (24), menor = 3 (2)	1 (6), 2 (24), 3 (2)
a45	nominal	moda = 2 (27), menor = 3 (2)	1 (3), 2 (27), 3 (2)
a46	nominal	moda = 2 (26), menor = 3 (2)	1 (4), 2 (26), 3 (2)
a47	nominal	moda = 2 (30), menor = 3 (2)	2 (30), 3 (2)
a48	nominal	moda = 2 (30), menor = 3 (2)	2 (30), 3 (2)
a49	nominal	moda = 2 (28), menor = 1 (2)	1 (2), 2 (28), 3 (2)
a50	nominal	moda = 2 (28), menor = 1 (2)	1 (2), 2 (28), 3 (2)
a51	nominal	moda = 2 (24), menor = 1 (4)	1 (4), 2 (24), 3 (4)
a52	nominal	moda = 2 (25), menor = 3 (1)	1 (6), 2 (25), 3 (1)
a53	nominal	moda = 2 (25), menor = 3 (1)	1 (6), 2 (25), 3 (1)
a54	nominal	moda = 2 (18), menor = 1 (14)	1 (14), 2 (18)
a55	nominal	moda = 2 (26), menor = 1 (6)	1 (6), 2 (26)
a56	nominal	moda = 2 (23), menor = 1 (9)	1 (9), 2 (23)

Este Dataset é interessante para o estudo de classificação por ter muitos atributos (56). A alta dimensionalidade de um Dataset influi na sua dificuldade de classificação.

4.4.7. Wine

Este Dataset insere-se no domínio da produção agrícola, em particular na qualificação de vinho. Os casos são derivados de diferentes viticultores e os atributos descrevem o vinho através de 13 atributos químicos. A classe é ternária e tem o objetivo de predizer o tipo de vinho. A distribuição da classe é de aproximadamente 33% para o tipo 1, 40% para o tipo 2 e 27% para o tipo 3. O Dataset não tem instâncias com valores em falta. Os dados estatísticos dos atributos são:

Quadro 4:7 - Dados estatísticos do Dataset Wine

Atributo	**Tipo**	**Estatística**	**Limites/Distribuição**
A1-Alcool	real	média = 13,001 +/- 0,812	[11,030 - 14,830]
A2-Malic acid	real	média = 2,336 +/- 1,117	[0,740 - 5,800]
A3-Ash	real	média = 2,367 +/- 0,274	[1,360 - 3,230]
A4-Alcalinity of ash	real	média = 19,495 +/- 3,340	[10,600 - 30,000]
A5-Magnesium	real	média = 99,742 +/- 14,282	[70,000 - 162,000]
A6-Total phenols	real	média = 2,295 +/- 0,626	[0,980 - 3,880]
A7-Flavanoids	real	média = 2,029 +/- 0,999	[0,340 - 5,080]
A8-Nonflavanoid phenols	real	média = 0,362 +/- 0,124	[0,130 - 0,660]
A9-Proanthocyanins	real	média = 1,591 +/- 0,572	[0,410 - 3,580]

A10-Color intensity	real	média = 5,058 +/- 2,318	[1,280 - 13,000]
A11-Hue	real	média = 0,957 +/- 0,229	[0,480 - 1,710]
A12-OD280/OD315 of diluted wines	real	média = 2,612 +/- 0,710	[1,270 - 4,000]
A13-Proline	real	média = 746,893 +/- 314,907	[278,000 - 1.680,000]

Este Dataset é um caso tipo de aplicação de classificação e, visto que as classes são separáveis, não é difícil de classificar, o que o comprova os resultados divulgados na literatura que se situam entre 96-100% de acurácia.

4.4.8. Zoo

O Dataset pertence ao domínio das ciências da vida, em particular da Biologia. Trata-se da classificação de animais através de alguns dos seus atributos físicos observáveis. O valor dos atributos é binário o que significa a presença ou ausência dessa observação (com exceção do atributo legs). A classe é 7-ária e tem o objetivo de predizer o tipo de vertebrado ou invertebrado a que o animal pertence. A distribuição da classe é de 4% para anfíbios, 20% para aves, 13% para peixes, 8% para insetos, 10% para invertebrados, 41% para mamíferos e 5% para répteis. O Dataset não tem instâncias com valores em falta. Os dados estatísticos dos atributos são:

Quadro 4:8 - Dados estatísticos do Dataset Zoo

Atributo	**Tipo**	**Estatística**	**Limites/Distribuição**
hair	nominal	moda = 0 (58), menor = 1 (43)	0 (58), 1 (43)
feathers	nominal	moda = 0 (81), menor = 1 (20)	0 (81), 1 (20)
eggs	nominal	moda = 1 (59), menor = 0 (42)	0 (42), 1 (59)
milk	nominal	moda = 0 (60), menor = 1 (41)	0 (60), 1 (41)
airborne	nominal	moda = 0 (77), menor = 1 (24)	0 (77), 1 (24)
aquatic	nominal	moda = 0 (65), menor = 1 (36)	0 (65), 1 (36)
predator	nominal	moda = 1 (56), menor = 0 (45)	0 (45), 1 (56)
toothed	nominal	moda = 1 (61), menor = 0 (40)	0 (40), 1 (61)
backbone	nominal	moda = 1 (83), menor = 0 (18)	0 (18), 1 (83)
breathes	nominal	moda = 1 (80), menor = 0 (21)	0 (21), 1 (80)
venomous	nominal	moda = 0 (93), menor = 1 (8)	0 (93), 1 (8)
fins	nominal	moda = 0 (84), menor = 1 (17)	0 (84), 1 (17)
legs	nominal	moda = 4 (38), menor = 5 (1)	0 (23), 2 (27), 4 (38), 5(1), 6(10), 8(2)
tail	nominal	moda = 1 (75), menor = 0 (26)	0 (26), 1 (75)
domestic	nominal	moda = 0 (88), menor = 1 (13)	0 (88), 1 (13)
catsize	nominal	moda = 0 (57), menor = 1 (44)	0 (57), 1 (44)

O Quadro 4:9 mostra a caraterização dos Datasets: o nome pelo qual são conhecidos na literatura, os tipos de variáveis existentes nos dados sendo que nominais e ordinais foram agrupados em categóricas, o número de instâncias (tuplos, casos), o número de atributos e o número de valores possíveis para cada classe alvo. Os Datasets são multivariados (várias variáveis independentes têm influência na variável dependente) e univariados (a variável dependente é somente influenciada por uma variável independente) e pertencem à tarefa de classificação visto que é nessa tarefa que se concentra este estudo.

Quadro 4:9 - Caraterização geral dos Datasets

Nome do Dataset	Tipos de variáveis	#Instâncias	#Atributos	#Classes
Adult	Categóricas, inteiros	32561	14	2
Breast-cancer	Categóricas	286	9	2
Car Evaluation	Categóricas	1728	6	4
Credit-approval	Categóricas, inteiros, reais	690	15	2
Iris	Reais	150	4	3
Lung-cancer	Inteiros	32	56	3
Wine	Inteiros, reais	178	13	3
Zoo	Categóricas, inteiros	101	16	7

Estes Datasets foram selecionados por fornecerem uma ampla variedade de possibilidades. Possuem variáveis quantitativas discretas e contínuas bem como variáveis qualitativas nominais e ordinais. O número de instâncias varia entre o mínimo de 32 e o máximo de 32561. O número de atributos varia entre 4 no mínimo e 56 no máximo. A classe alvo varia entre classificação binária e 7-ária. Alguns Datasets têm poucas instâncias, mas muitos atributos, outros com o contrário. Estas caraterísticas mostram que cada Dataset é único e que no seu conjunto, os Datasets asseguram uma boa mescla de exemplos para o teste.

4.5. Algoritmos

A terceira etapa da metodologia consiste em selecionar os algoritmos que serão usados no teste. Serão usados algoritmos das técnicas que a seguir se descrevem:

- Árvore de decisão
- Indução de regras
- Segmentação (ou clustering)
- Rede Neuronal Artificial
- Classificador Bayesiano
- Support Vector Machine

4.6. Avaliação de Performance

Existem várias formas de dividir as instâncias pelos conjuntos Trainingset e Testset, conhecido por particionamento. Neste teste serão usadas as duas mais comuns: (1) *Percentage Split* e (2) *Cross-Validation*. *Percentage Split* (também chamado de *Holdout*) corresponde a definir uma percentagem de divisão do Dataset em dois conjuntos disjuntos; é normal a divisão ser de dois terços, o que corresponde a dizer que o Trainingset terá 66% das instâncias totais e que as restantes 34% das instâncias constituirão o Testset. *Cross-Validation* é uma técnica de particionamento que divide o Dataset em *k* partes disjuntas. O algoritmo é treinado com *k-1* partes e a parte restante é usada no treino. Este procedimento é repetido *k* vezes alternando a parte para treino. A avaliação de performance é dada pelo valor médio. É comum o valor de *k* ser de 10 mas deve ser considerado o tamanho do Dataset.

Para avaliar o desempenho dos classificadores gerados pelas diferentes ferramentas é usada a medida matemática de avaliação de modelos de Data Mining mais utilizada na investigação

científica, a acurácia, que representa a percentagem de casos corretamente classificados relativamente aos casos totais:

$$Acurácia = (n.º\ de\ predições\ corretas)/(n.º\ total\ de\ predições)$$

Equação 1 – Acurácia

A acurácia é uma taxa calculada pelo número de instâncias bem classificadas sobre o valor de instâncias total. Uma instância bem classificada é aquela em que o classificador prediz a classe correta da instância do teste.

4.7. Resultados experimentais

O teste consiste em avaliar as ferramentas KNIME, Orange, RapidMiner e Weka para as técnicas de classificação Árvores de decisão, Regras de classificação, Redes Neuronais Artificiais, Classificadores Bayesianos e Clustering. Esta seção apresenta os resultados obtidos para os Datasets descritos no Quadro 4:9.

4.7.1. Preparação da experiência

Os testes serão realizados com amostragem aleatória nos modos de particionamento *Percentage Split* com 70:30 e de *Cross-Validation* com *k=5* visto que o Dataset mais pequeno contém 101 instâncias.

Relativamente à capacidade de as ferramentas executarem as técnicas de classificação, note-se que o Orange não possui nenhum *widget* para realizar Redes Neuronais Artificiais.

Na realização do teste, os parâmetros dos algoritmos serão executados com os seus valores pré-definidos pelas ferramentas, exceto os casos em que é possível selecionar que valores em falta sejam ignorados. É possível obter melhor performance mudando os parâmetros dos algoritmos, porém a diferença não será significativa para o âmbito deste teste uma vez que todos os operadores poderão ser melhorados dessa forma.

Todos os Datasets usados no teste foram salvaguardados no formato padronizado pela Weka (.arff). Todas as ferramentas têm capacidade de ler este formato nativamente.

O Orange foi programado (Figura 2) através dos *Learners widgets* disponíveis para as técnicas de classificação: *Classification Tree*, *CN2*, *naïve Bayes*, *k Nearest Neighbours* e *SVM.* O *Data widget File* e todos os *Learners* foram ligados a um *Test Learner*, parametrizado em alternância nos dois modos de particionamento dos testes. Não foi usado nenhum *widget* de pré-processamento.

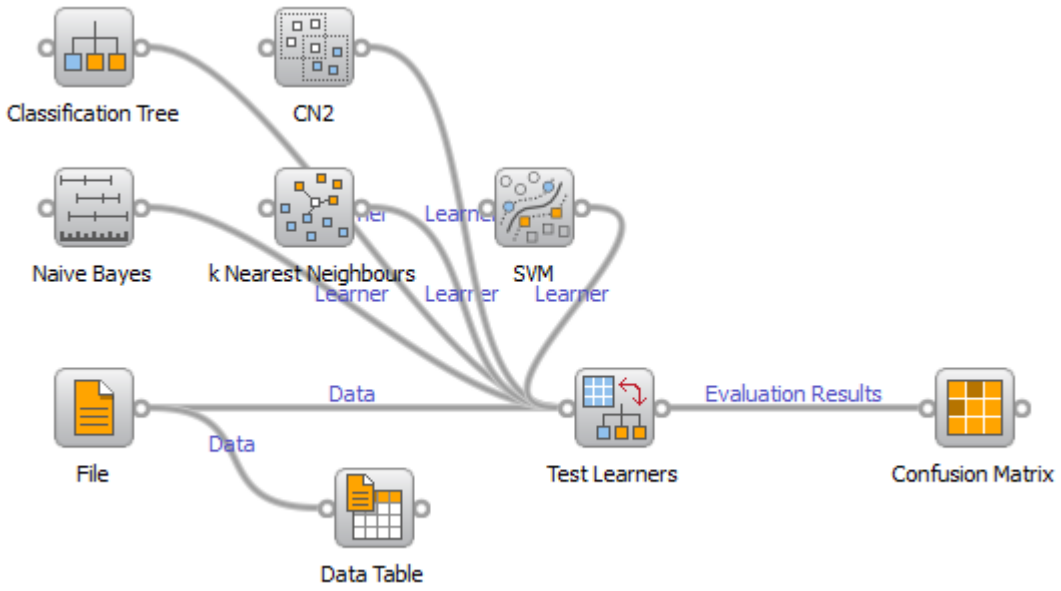

Figura 2 – Programação da experiência em Orange

O KNIME foi programado (Figura 3) com os nós disponíveis do grupo *Mining*. Para o *Decision Tree Learner* e o *Naive Bayes Learner*, os dados não foram sujeitos a pré-processamento uma vez que os nós aceitam todos os tipos de atributos. Para os restantes *Learners*, os atributos categóricos foram convertidos em valores numéricos do tipo *DoubleValue* e normalizados entre 0.0 e 1.0. Para o *RProp MLP Learner*, a classe real e a classe predita foram convertidas para o tipo *IntValue* antes da avaliação. Para o *k Nearest Neighbor* e *SVM Learner* a classe real do Trainingset e do Testset é convertida para *StringValue* antes da construção dos modelos.

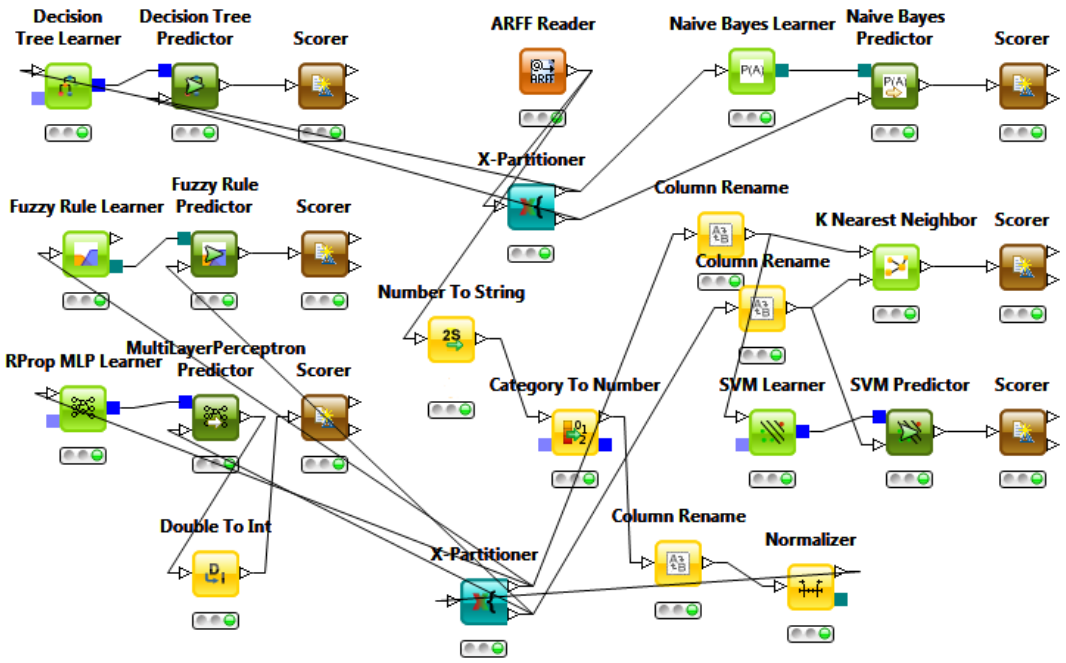

Figura 3 - Programação da experiência em KNIME

O Weka e o RapidMiner possuem vários algoritmos para cada técnica. Os testes foram exaustivos, i.e., todos os algoritmos foram testados. Os resultados mostram aquele que atingiu uma melhor acurácia ou, em caso de empate o operador que foi executado em primeiro lugar. O importante para o teste é determinar a melhor técnica e não o melhor algoritmo.

Para a programação em Weka (Figura 4) não foi realizado qualquer pré-processamento para além daquele incluído nos próprios algoritmos, e.g. discretização. A experiência consiste somente na seleção do algoritmo e modo de particionamento.

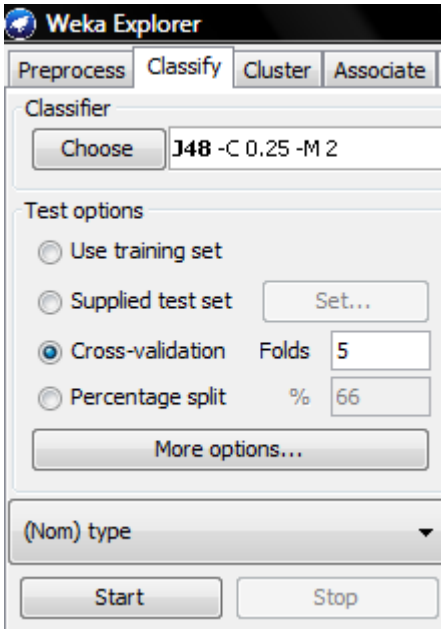

Figura 4 - Programação da experiência em Weka

O RapidMiner (Figura 5) possui alguns operadores (e.g. *NeuralNetImproved* e *LibSVMLearner*), que somente funcionam com atributos numéricos; para estes casos, os dados categóricos foram transformados em numéricos através do operador *Nominal2Numerical* e normalizados entre 0.0 e 1.0 com o operador *Normalization*.

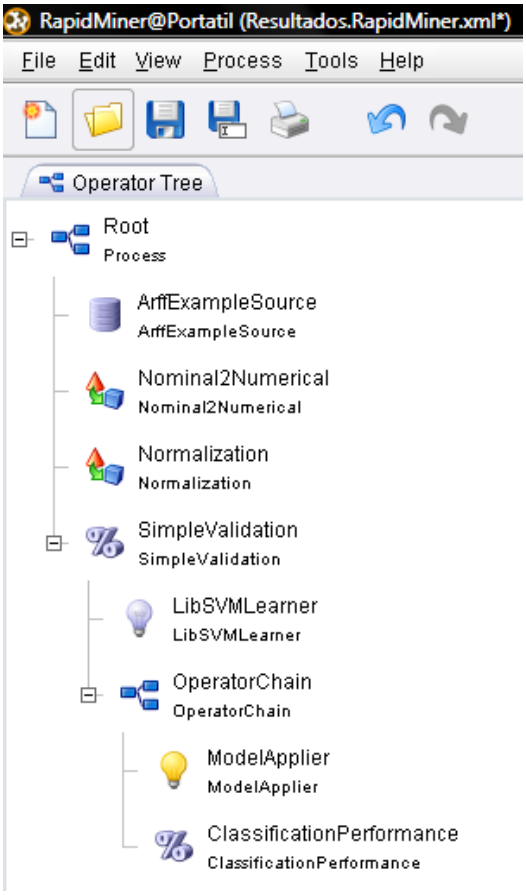

Figura 5 - Programação da experiência em RapidMiner

4.7.2. Avaliação dos resultados

As análises dos resultados podem ser realizadas segundo diferentes perspetivas. Em seguida mostram-se os resultados totais e separados por Datasets, Ferramentas, Técnicas, Particionamentos e Algoritmos (Classificadores). Os totais dos resultados agregados foram calculados através da média aritmética simples dos testes.

Não foi possível realizar alguns testes: O Orange não possui a capacidade de realizar Redes Neuronais Artificiais e para o algoritmo *CN2* da técnica Regras, o Dataset Adult resultou num erro de falta de memória disponível.

O total de todos os testes realizados resultou numa acurácia de classificação média de 81,13%. Este resultado deve ser compreendido como a medida da dificuldade que a classificação destes Datasets representa para os algoritmos apresentados. O desvio padrão indica a dispersão de resultados dos diferentes algoritmos, o que releva a importância da escolha do classificador correto.

Quadro 4:10 – Resultados do estudo de classificação

Acurácia média	Acurácia mínima	Acurácia máxima	Desvio padrão
81,13%	0,00%	100,00%	18,49%

Considerando as caraterísticas dos Datasets testados (Quadro 4:9), não é evidente a correlação entre o resultado da classificação e o tipo de variáveis, a cardinalidade de instâncias, o número de atributos ou valores para a classe alvo. Verifica-se que o Dataset com melhor acurácia (Iris) é aquele com o menor número de atributos (4) e o Dataset com a pior acurácia (Lung-cancer) é aquele com o maior número de atributos (56). No entanto, este facto não se verifica para os restantes Datasets. Apenas o Dataset Lung-cancer tem um desvio à média superior ao desvio padrão. Os restantes Datasets obtiveram uma acurácia entre o mínimo de 71,88% para o Breast-cancer e o máximo de 93,86% para o Iris.

Quadro 4:11 - Resultados do estudo de classificação por Dataset

Dataset	**Acurácia média**	**Acurácia mínima**	**Acurácia máxima**	**Desvio padrão**
Iris	93,86%	80,00%	100,00%	4,40%
Zoo	93,10%	69,24%	100,00%	5,83%
Wine	92,76%	37,74%	100,00%	12,38%
Car Evaluation	90,14%	65,44%	99,61%	7,06%
Credit Approval	83,38%	50,24%	91,30%	6,99%
Adult	80,05%	24,08%	86,93%	12,67%
Breast-cancer	71,88%	63,95%	77,92%	3,34%
Lung-cancer	43,81%	0,00%	80,00%	15,84%
Total Geral	**81,13%**	**0,00%**	**100,00%**	**18,49%**

Na perspetiva das ferramentas, observa-se que a melhor acurácia foi obtida pelo *software* da Weka com 84,28%, em concordância com o estudo de (Wahbeh, Al-Radaideh, Al-Kabi, & Al-Shawakfa, 2010), e a pior acurácia foi obtida pelo RapidMiner com 77,67%.

Quadro 4:12 – Resultados do estudo de classificação por Ferramenta

Ferramenta	**Acurácia média**	**Acurácia mínima**	**Acurácia máxima**	**Desvio padrão**
Weka	84,28%	20,00%	100,00%	17,84%
KNIME	81,40%	0,00%	100,00%	18,30%
Orange	81,19%	18,00%	98,33%	18,01%
RapidMiner	77,67%	20,00%	100,00%	19,36%
Total Geral	**81,13%**	**0,00%**	**100,00%**	**18,49%**

Na análise por técnicas podemos referir que o melhor resultado foi obtido através de modelos baseados em Árvores de Decisão com 83,10%. O pior resultado foi obtido através de Support Vector Machines com 74,72%, o que é explicado pelos fracos resultados obtidos com os Datasets Lung-cancer e Adult, em especial no RapidMiner com 25,95% e 24,14%, respetivamente. Estes Datasets são os maiores em termos de cardinalidade de instâncias e número de atributos, o que revela uma tendência para o classificador piorar com a complexidade do Dataset.

Quadro 4:13 - Resultados do estudo de classificação por Técnica

Técnica	**Acurácia média**	**Acurácia mínima**	**Acurácia máxima**	**Desvio padrão**
Árvores de Decisão	83,10%	40,00%	100,00%	15,87%
Rede Neuronal Artificial	82,85%	30,00%	100,00%	16,91%
Regras	82,76%	25,00%	100,00%	16,31%

Clustering	82,07%	39,52%	100,00%	16,17%
Classificador Bayesiano	81,76%	0,00%	100,00%	19,33%
SVM	74,72%	18,00%	100,00%	23,88%
Total Geral	**81,13%**	**0,00%**	**100,00%**	**18,49%**

Em termos de particionamento, a melhor acurácia foi obtida em *Cross-Validation k=5* com 81,41%, mas muito perto de *Percentage Split 70:30* com 80,85%. O ganho deve ser considerado tendo em conta um acréscimo de tempo de processamento, que consoante o tamanho do Dataset, pode ou não ser significativo.

Quadro 4:14 - Resultados do estudo de classificação por Modo de Particionamento

Particionamento	Acurácia média	Acurácia mínima	Acurácia máxima	Desvio padrão
X-Validation k=5	81,41%	0,00%	100,00%	18,24%
% Split 70:30	80,85%	10,00%	100,00%	18,78%
Total Geral	**81,13%**	**0,00%**	**100,00%**	**18,49%**

Os melhores algoritmos classificadores são da Weka e têm por base a técnica das Árvores de Decisão. Os melhores atingem mesmo a acurácia ótima, mas apenas num teste. Nos casos em que apenas um classificador estava disponível na ferramenta, casos do Orange e KNIME, o número máximo possível de testes nesse classificador é de 16. É por isso de salientar o resultado do classificador BayesNet da Weka que foi o melhor em 9 dos 16 testes possíveis e com uma acurácia média de 87,24%.

Quadro 4:15 - Resultados do estudo de classificação por Algoritmo/ Widget

Algoritmo/Widget	Acurácia média	Acurácia mínima	Acurácia máxima	Desvio padrão	# Testes
AODEsr	100,00%	100,00%	100,00%	0,00%	1
NBTree	100,00%	100,00%	100,00%	0,00%	1
FT	98,26%	97,78%	98,88%	0,46%	3
HNB	94,79%	92,71%	98,02%	2,32%	3
PART	94,59%	94,59%	94,59%	0,00%	1
ID3	94,36%	87,07%	98,02%	5,16%	3
SimpleCart	91,98%	86,27%	97,68%	5,71%	2
DTNB	91,66%	86,37%	100,00%	5,54%	4
BFTree	91,00%	85,99%	96,00%	5,00%	2
Jrip	90,30%	85,94%	94,67%	4,36%	2
LWL	88,78%	85,51%	95,33%	4,63%	3
LMT	87,33%	76,22%	98,44%	11,11%	2
BayesNet	87,24%	74,13%	100,00%	8,65%	9
Ridor	86,87%	77,91%	95,83%	8,96%	2
J48graft	86,24%	86,24%	86,24%	0,00%	1
RuleLearner	86,02%	53,81%	97,00%	12,03%	11
DecisionStump	85,51%	85,51%	85,51%	0,00%	2
BasicRuleLearner	85,17%	84,89%	85,51%	0,26%	3
NaiveBayes	85,02%	53,33%	100,00%	15,51%	6

MultilayerPerceptron	84,38%	30,00%	99,61%	20,02%	14
Rprop MLP Learner	84,38%	60,00%	97,11%	9,82%	16
NNge	83,86%	56,25%	98,31%	19,53%	3
KernelNaiveBayes	83,79%	60,00%	97,21%	11,47%	11
k Nearest Neighbor	83,29%	50,00%	95,24%	13,72%	16
Decision Tree Learner	83,06%	42,86%	100,00%	16,77%	16
IB1	82,62%	46,88%	100,00%	17,78%	6
IBk	82,59%	72,09%	92,42%	9,68%	4
SMO	82,40%	20,00%	100,00%	21,63%	16
Naive Bayes	82,05%	44,00%	97,76%	13,85%	16
k Nearest Neighbours	81,43%	39,52%	97,05%	16,51%	16
ADTree	81,34%	76,74%	85,94%	4,60%	2
Classification Tree	81,21%	42,86%	96,00%	15,66%	16
SVM Learner	81,06%	42,86%	95,24%	12,52%	16
NeuralNetImproved	80,93%	30,00%	100,00%	19,39%	16
CN2	80,67%	43,00%	94,67%	16,74%	16
SVM	80,53%	18,00%	98,33%	24,58%	16
NearestNeighbors	80,49%	44,76%	100,00%	16,30%	16
CHAID	80,46%	48,10%	95,56%	14,38%	11
Fuzzy Rule Learner	80,17%	25,00%	97,89%	20,57%	16
KStar	79,00%	40,00%	98,88%	27,58%	3
DecisionTable	77,50%	50,00%	95,56%	19,76%	3
Naive Bayes Learner	76,41%	0,00%	100,00%	28,30%	16
VotedPerceptron	75,14%	72,38%	77,91%	2,76%	2
LBR	74,48%	74,48%	74,48%	0,00%	1
MultiCriterionDecisionStump	73,26%	73,26%	73,26%	0,00%	1
BestRuleInduction	73,26%	73,26%	73,26%	0,00%	1
EvoSVM	70,25%	70,25%	70,25%	0,00%	1
OneR	70,00%	70,00%	70,00%	0,00%	1
LibSVMLearner	55,96%	20,00%	92,00%	21,80%	14
J48	53,13%	53,13%	53,13%	0,00%	1
AODE	48,13%	40,00%	56,25%	8,12%	2
LADTree	40,00%	40,00%	40,00%	0,00%	1
JMySVMLearner	24,19%	24,19%	24,19%	0,00%	1
Total Geral	**81,13%**	**0,00%**	**100,00%**	**18,46%**	**384**

Apesar de alguns classificadores terem obtido taxas de acurácia baixas, é de salientar que foi o melhor resultado obtido para um determinado teste.

Visto que a seleção dos Datasets considerou uma mistura de possibilidades que inclui variáveis categóricas e quantitativas com números inteiros e reais, é possível observar a melhor ferramenta e a melhor técnica de acordo com estes critérios.

Quadro 4:16 - Resultados do estudo de classificação por Tipo de Dados e Técnica

Tipo Dados / Técnica	**Acurácia média**	**Acurácia mínima**	**Acurácia máxima**	**Desvio padrão**	**# Testes**
Categóricas					
Rede Neuronal Artificial	84,18%	63,95%	99,61%	12,97%	16
Regras	83,10%	70,93%	97,89%	9,72%	16
Árvores de Decisão	82,21%	65,39%	98,44%	11,50%	16
Clustering	80,04%	65,73%	94,41%	10,18%	16
Classificador Bayesiano	79,73%	69,56%	93,63%	7,67%	16
SVM	77,60%	65,44%	95,60%	10,40%	16
Inteiros					
Árvores de Decisão	48,12%	40,00%	60,00%	6,03%	8
Clustering	47,16%	39,52%	57,14%	5,42%	8
Rede Neuronal Artificial	47,10%	30,00%	71,43%	15,01%	8
Regras	46,84%	25,00%	70,00%	13,09%	8
Classificador Bayesiano	40,33%	0,00%	60,00%	21,56%	8
SVM	34,11%	18,00%	80,00%	19,69%	8
Misto					
Regras	89,73%	82,41%	100,00%	5,00%	32
Classificador Bayesiano	89,65%	80,44%	100,00%	6,76%	32
Árvores de Decisão	89,20%	81,13%	100,00%	6,21%	32
Clustering	88,53%	72,25%	100,00%	7,91%	32
Rede Neuronal Artificial	88,37%	77,78%	97,76%	5,56%	32
SVM	79,33%	24,08%	100,00%	21,43%	32
Reais					
Classificador Bayesiano	95,75%	92,00%	100,00%	2,81%	8
Árvores de Decisão	95,44%	93,33%	97,78%	1,40%	8
Clustering	95,22%	91,11%	100,00%	2,35%	8
Rede Neuronal Artificial	93,82%	83,33%	100,00%	5,43%	8
Regras	91,85%	82,14%	95,56%	4,49%	8
SVM	91,08%	80,00%	96,67%	5,63%	8
Total Geral	**81,13%**	**0,00%**	**100,00%**	**18,46%**	**384**

Embora o número de Datasets testados não permita uma conclusão precisa, não deixa de ser um indicador observar a acurácia por tipo de dados. Desta forma é possível concluir que quando se trate de um Dataset que seja composto exclusivamente por atributos que sejam variáveis categóricas, a melhor técnica foi a RNA. Se o Dataset for composto em exclusivo por variáveis quantitativas, a melhor técnica foi a AD para números inteiros e o Classificador Bayesiano para números reais. Quando se trate de um Dataset misto, composto por variáveis categóricas e/ou Inteiros e/ou Reais, a melhor técnica foi obtida pela técnica das Regras de classificação. Note-se que os resultados médios das outras técnicas, com exceção da SVM, estão acima dos valores mínimos das melhores técnicas o que revela a importância de avaliar outros fatores para além da técnica.

Quadro 4:17 - Resultados do estudo de classificação por Tipo de Dados e Ferramenta

Tipo Dados / Ferramenta	**Acurácia média**	**Acurácia mínima**	**Acurácia máxima**	**Desvio padrão**	**#Testes**
Categóricas					
Weka	85,04%	71,33%	99,61%	10,48%	24
KNIME	81,25%	65,39%	97,89%	11,11%	24
Orange	80,57%	68,37%	95,60%	9,67%	24
RapidMiner	77,11%	63,95%	94,79%	9,55%	24
Inteiros					
RapidMiner	47,18%	20,00%	70,00%	13,61%	12
KNIME	43,55%	0,00%	80,00%	22,58%	12
Weka	43,33%	20,00%	56,25%	10,25%	12
Orange	40,63%	18,00%	59,05%	12,05%	12
Misto					
Weka	91,20%	79,31%	100,00%	6,84%	48
KNIME	88,82%	80,44%	100,00%	5,84%	48
Orange	88,67%	80,07%	98,33%	5,76%	48
RapidMiner	81,18%	24,08%	98,11%	18,04%	48
Reais					
Weka	96,04%	94,67%	97,78%	0,90%	12
RapidMiner	95,19%	86,67%	100,00%	3,65%	12
Orange	94,53%	92,00%	96,00%	1,20%	12
KNIME	89,81%	80,00%	100,00%	5,83%	12
Total Geral	**81,13%**	**0,00%**	**100,00%**	**18,46%**	**384**

A análise por ferramenta determinou que a melhor ferramenta é a Weka para Datasets compostos em exclusivo por variáveis categóricas, reais ou sempre que haja uma mistura de algum destes tipos. A ferramenta RapidMiner destacou-se na classificação de Datasets compostos em exclusivo por números inteiros. Todavia, os resultados médios de todas as ferramentas estão acima dos valores mínimos da melhor ferramenta, o que indica a importância de considerar mais do que o tipo de dados para a escolha certa.

Para determinar a ferramenta e técnica que melhor se adequa à tarefa de classificação, foi comparada a acurácia da ferramenta tendo em conta a técnica e vice-versa.

A melhor ferramenta da experiência foi a Weka visto que obteve o 1º lugar em 4 técnicas e o 2º lugar em 2 técnicas. A segunda melhor foi a KNIME que obteve o 1º lugar em 2 técnicas, o 2º lugar em 2 técnicas e o 4º lugar em 2 técnicas. A terceira melhor foi a RapidMiner que obteve o 2º lugar em 2 técnicas, o 3º lugar em 2 técnicas e o 4º lugar em 2 técnicas. A quarta melhor foi a Orange que obteve o 3º lugar em 4 técnicas e o 4º lugar em 2 técnicas.

Quadro 4:18 - Resultados do estudo de classificação por Tipo de Dados e Ferramenta

Ferramenta	**Árvore de Decisão**	**Rede Neuronal Artificial**	**Regras**	**Clustering**	**Classificador Bayesiano**	**SVM**
Weka	85,96%	83,23%	85,89%	83,09%	85,11%	82,40%

KNIME	83,06%	84,38%	80,17%	83,29%	76,41%	81,06%
Orange	81,21%	-	80,67%	81,43%	82,05%	80,53%
RapidMiner	82,15%	80,93%	84,06%	80,49%	83,48%	54,87%

A melhor técnica da experiência foi a Árvore de Decisão visto que obteve o 1º lugar na Weka e o 3º lugar nas restantes ferramentas. A segunda melhor técnica foi a Indução de Regras a par com os Classificadores Bayesianos. A Indução de Regras obteve o 1º lugar na RapidMiner, o 2º lugar na Weka, o 4º lugar na Orange e o 5º lugar na KNIME. Os Classificadores Bayesianos obtiveram o 1º lugar na Orange, o 2º lugar na RapidMiner, o 3º lugar na Weka e o 6º lugar na KNIME. A quarta melhor técnica foi o Clustering com o 2º lugar na KNIME e na Orange e o 5º lugar na Weka e em RapidMiner. A quinta melhor técnica foi as Redes Neuronais Artificiais com o 1º lugar na KNIME, o 4º lugar na Weka e na RapidMiner e o 6º lugar na Orange. A sexta melhor técnica foi o SVM com o 4º lugar na KNIME, o 5º lugar na Orange e o 6º lugar na Weka e RapidMiner.

4.8. Associação de classificadores

Uma das formas de melhorar o resultado de um modelo de classificação é através do método de associação (Ensemble). Ao invés da tradicional execução de um algoritmo para indução de um modelo, este permite executar vários algoritmos induzindo vários modelos e avaliar os diferentes resultados de forma combinada. Existem diversos métodos de combinação e de avaliação de resultados, e.g. é possível executar o mesmo algoritmo diversas vezes ou correr vários algoritmos e o resultado pode ser calculado como uma média dos resultados ou aceitando o melhor resultado obtido.

Sabendo que a associação de classificadores poderá melhorar o resultado, importa saber se essa melhoria é positivamente significativa. Com base nesta premissa foi realizado um teste com as possibilidades de associação da ferramenta Weka, tendo por base os resultados do estudo descrito anteriormente.

Os seguintes algoritmos foram testados pela ferramenta Weka:

- *Adaboost* (Freund & Schapire, 1996) – Usa um método de atribuir maior peso às instâncias mal classificadas pelo modelo induzido anteriormente como forma destas virem a ser bem classificadas no modelo induzido posteriormente. A decisão final é obtida pela soma ponderada dos modelos induzidos. A desvantagem deste método prende-se pelo facto de ser sensível ao ruído, ou seja, de tentar classificar instâncias que podem conter dados errados em detrimento das instâncias corretas.
- *Bagging* (Breiman, Bagging predictors, 1996) – Usa um método de manipulação dos dados de treino que recolhe amostras de forma aleatória, mas permite substituições (*bootstrapping*) criando possíveis omissões e duplicações de instâncias. A decisão final é por voto maioritário.
- *Dagging* (Ting & Witten, 1997) - Este algoritmo cria subconjuntos de dados de treino disjuntos pelo método de estratificação e cada um é usado para a indução de um

modelo diferente. A decisão final é por voto maioritário. A técnica é útil para classificadores de tempo quadrático ou pior.

- *Decorate* (Melville & Mooney, 2003) – Usa um método de construção de diversos modelos usando instâncias artificiais de treino especialmente construídas através dos valores dos atributos das instâncias existentes. Experiências têm demonstrado que esta técnica é consistentemente mais precisa do que o classificador base, Bagging e *Random Forests*. Esta técnica obtém maior acurácia em conjuntos de testes pequenos e desempenho comparável em conjuntos maiores.
- *MultiBoostAB* (Webb, 2000) – Esta técnica resulta da combinação do *Adaboost* (*Boosting*) com *Wagging*. Este último equivale à técnica de *Bagging*, mas ao invés de usar amostras aleatórias de *bootstrap* para formar os conjuntos de treino, o *Wagging* atribui pesos aleatórios para os casos em cada conjunto de treino.
- *MultiClassClassifier* – Esta técnica permite criar vários classificadores binários a partir de conjunto de dados de treino que tenham duas ou mais classes. Esta técnica permite usar o método de classificação de *1-1*, *1-R*, mas também *ECOC*.
- *Random Forest* (Breiman, 2001) – Envolve a aplicação de *Bagging*, o que introduz aleatoriedade na escolha dos dados de treino, à técnica de árvores de decisão, embora existam variantes para outras técnicas. A decisão final é obtida pela moda da classe obtida nos diferentes modelos induzidos.
- *RandomSubSpace* (Ho, 1998) – É similar ao algoritmo *Random Forest* visto que constrói vários modelos de árvores de decisão, mas em vez de selecionar as instâncias de forma aleatória, esta técnica seleciona subespaços (subconjuntos dos dados de treino) de forma aleatória.
- *Rotation Forest* (Rodriguez, Kuncheva, & Alonso, 2006) – O algoritmo treina diversos modelos de árvores de decisão usando conjuntos de treino obtidos através de recolha de instâncias selecionadas de aleatória com substituição (*bootstrapping*) mas sujeitos a uma análise e consequente extração de atributos não contributivos. Segundo os autores, desta forma são necessários menos modelos que em *Adaboost* (*Boosting*) ou *Bagging* para obter uma taxa de acurácia elevada.
- *Stacking* (Wolpert, 1992) – Algoritmo baseada numa técnica realizada em duas fases: a primeira envolve a indução de vários modelos com base no conjunto de dados de treino existente que, de acordo com os resultados combinados, irá dar origem a um novo conjunto de dados que generaliza as classificações obtidas. A segunda fase envolve a indução de um modelo com base no conjunto de dados criado na primeira fase.
- *Vote* – Este algoritmo permite combinar diversos modelos com base em regras de média, produto, votação maioritária, mínimo, máximo e mediana.

Os algoritmos *Stacking* e *Vote* permitem combinar modelos de diferentes técnicas enquanto os restantes somente combinam modelos similares. Por este motivo, os testes dos algoritmos *Stacking* e *Vote* foram programados com uma combinação dos algoritmos de classificação que obtiveram o melhor resultado no teste para cada uma das técnicas (Quadro 4:19).

Quadro 4:19 – Ensemble programado no estudo de classificação para os algoritmos *Stacking* e *Vote*

Dataset	**Técnica**	**Widget/Algoritmo**
Adult	Árvores de Decisão	J48graft
	Regras	DTNB
	Rede Neuronal Artificial	MultilayerPerceptron
	Classificador Bayesiano	BayesNet
	Clustering	IB1
	SVM	SMO
Breast-cancer	Árvores de Decisão	LMT
	Regras	LBR
	Rede Neuronal Artificial	VotedPerceptron
	Classificador Bayesiano	BayesNet
	Clustering	IBk
	SVM	SMO
Car Evaluation	Árvores de Decisão	LMT
	Regras	Ridor
	Rede Neuronal Artificial	MultilayerPerceptron
	Classificador Bayesiano	HNB
	Clustering	IBk
	SVM	SMO
Credit Approval	Árvores de Decisão	ADTree
	Regras	Jrip
	Rede Neuronal Artificial	MultilayerPerceptron
	Classificador Bayesiano	BayesNet
	Clustering	LWL
	SVM	SMO
Iris	Árvores de Decisão	BFTree
	Regras	Jrip
	Rede Neuronal Artificial	MultilayerPerceptron
	Classificador Bayesiano	NaiveBayes
	Clustering	LWL
	SVM	SMO
Lung-cancer	Árvores de Decisão	J48
	Regras	NNge
	Rede Neuronal Artificial	MultilayerPerceptron
	Classificador Bayesiano	AODE
	Clustering	IB1
	SVM	SMO
Wine	Árvores de Decisão	FT
	Regras	NNge
	Rede Neuronal Artificial	MultilayerPerceptron
	Classificador Bayesiano	BayesNet
	Clustering	KStar

	SVM	SMO
Zoo	Árvores de Decisão	Id3
	Regras	NNge
	Rede Neuronal Artificial	MultilayerPerceptron
	Classificador Bayesiano	HNB
	Clustering	IB1
	SVM	SMO

Todos os algoritmos são parametrizáveis embora para o teste tenham sido usados os valores pré-definidos.

Foram realizados 408 testes de Ensemble dos quais 89 revelaram uma melhoria relativamente à classificação anterior para o mesmo algoritmo sem Ensemble. A média dessa melhoria para os 89 casos foi de 1,60%. O algoritmo *SMO* (com o método de Ensemble *Bagging*, para o Dataset Lung-cancer e a técnica SVM) alcançou o maior aumento que se cifrou em 9,37%.

Dos 408 testes realizados, 28 revelaram uma melhora relativamente ao melhor resultado obtido nos testes anteriores para um determinado Dataset. A média de melhoria nesses testes foi de 0,96%. O melhor resultado deu-se com o algoritmo *NNge* (com o método de Ensemble *Rotation Forest*, para o Dataset Lung-cancer na técnica Regras) onde se alcançou uma melhoria de 6,25%.

4.9. Conclusão do estudo de classificação

Esta experiência permitiu avaliar a acurácia resultante de vários algoritmos classificadores através de diversas comparações entre 8 Datasets selecionados pelas suas diferentes caraterísticas, 4 ferramentas consideradas das melhores no mercado em utilização livre, 6 técnicas de Machine Learning para classificação e 2 modos de particionamento.

A conclusão que se pode retirar do estudo é que não existe uma ferramenta ou técnica que seja melhor que todas as outras para qualquer tarefa de classificação. Individualmente, o melhor resultado foi conseguido com a ferramenta Weka e a técnica Árvores de Decisão com 85,96%. O pior resultado foi conseguido com a ferramenta RapidMiner e a técnica SVM com 54,87%. A técnica de Ensemble permite alcançar ligeiras melhorias embora tal não se verifique na maioria dos casos.

Os testes foram realizados sem alteração dos parâmetros dos algoritmos visto que testar todos os parâmetros de todos os algoritmos revelar-se-ia numa tarefa demasiado complexa; mas esta é uma forma de conseguir melhorar um resultado.

Para concluir quanto aos valores da acurácia importa aferir o que se espera de um classificador, ou seja, quais os resultados que um classificador terá que atingir para ser considerado como aplicável ou utilizável em situação real. Uma possibilidade passa por atribuir limites aos resultados. Por exemplo, para ser classificado como razoável, o modalo tem que ter uma acurácia de, pelo menos, 70%. Claro está que não existe uma percentagem universal dados os possíveis domínios de aplicação da classificação. Desta forma, o método mais adequado seria recorrer a

um perito no domínio para definir esses valores, tal como sugerido por (Saitta, 2010). Somente dessa forma poderíamos concluir que o classificador obtido teria aplicação prática.

Neste estudo foram registados 384 testes contabilizando somente aqueles que obtiveram o melhor resultado. Pela sua dimensão, deixou de fora muitas alternativas de teste que se sugere para trabalho futuro. São elas o teste dos algoritmos com os parâmetros possíveis, testar mais ferramentas e não somente as de utilização livre, testar outros modos de particionamento do Dataset, testar mais Datasets, estender a outras tarefas de Data Mining e incluir técnicas como os algoritmos genéticos. No final, os resultados podem ainda ser usados para desenvolver um *software* que contenha os algoritmos que apresentem os melhores resultados alcançados.

5. ESTUDO

O IPC/Instituto Superior de Engenharia de Coimbra (ISEC) e a Administração Regional de Saúde do Centro (ARSC) estabeleceram um acordo de colaboração no âmbito do Mestrado em Informática e Sistemas (MIS) – Especialização em Tecnologias da Informação e do Conhecimento com vista à elaboração da tese de mestrado "Um estudo da Diabetes e Hipertensão baseado em técnicas de Data Mining aplicadas a dados da ARS Centro".

O estudo que a seguir se descreve terá as seguintes etapas, estabelecidas de acordo com a metodologia CRISP-DM:

1. Compreensão do negócio;
2. Compreensão dos dados;
3. Preparação dos dados
4. Modelação
5. Avaliação
6. Operacionalização

5.1. Compreensão do negócio

A primeira etapa da metodologia compreende a necessidade de conhecer a entidade alvo do estudo para que, posteriormente, os processos de tratamento dos dados e resultados sejam dirigidos aos intentos e objetivos da organização. As informações que se seguem podem ser consultadas em maior pormenor no Relatório de Atividade da (Administração Regional de Saúde do Centro, 2012).

A ARSC é um instituto público que tem como missão "garantir à população da respetiva área geográfica de intervenção o acesso à prestação de cuidados de saúde de qualidade, adequando os recursos disponíveis às necessidades em saúde, respeitando as regras de equidade, cumprindo e fazendo cumprir o PNS e as leis e regulamentos em vigor"

Para além dos serviços centrais, a ARSC integra serviços desconcentrados – os Agrupamentos de Centros de Saúde (ACeS). Após uma restruturação a vigorar desde 30 de novembro de 2012, os ACeS passaram a ser os seguintes: Baixo Mondego, Baixo Vouga, Cova da Beira, Dão-Lafões, Pinhal Interior Norte e Pinhal Litoral. A ARSC integra ainda duas Unidades Locais de Saúde (ULS): Guarda e Castelo Branco. Os ACeS agregam 64 Centros de Saúde (CS) e as ULS 20. Da ARSC fazem ainda parte os Centros Hospitalares e Hospitais localizados nessas áreas geográficas, aqui não descritas por não fazerem parte do estudo.

A ARSC abrange uma área geográfica de 23.274 Km^2, que integra 77 concelhos e a que corresponde 26% do território de Portugal Continental. Serve 1.737.216 residentes, de acordo com os Censos 2011, a que corresponde 17% da população residente no Continente e a uma densidade populacional de 75 habitantes/Km^2. Os CS escolhidos pela variedade geográfica e demográfica para fornecerem os dados deste estudo são:

Quadro 5:1 - Caraterização dos Centros de Saúde da ARS Centro

CS	ACeS	Localização	Densidade populacional
Arnaldo Sampaio	Pinhal Litoral II	Litoral	184 Habitantes/Km^2
Eiras	Baixo Mondego I	Litoral	259 Habitantes/Km^2
Fundão	Cova da Beira	Interior	64 Habitantes/Km^2
Tábua	Pinhal Interior Norte I	Interior	52 Habitantes/Km^2

Visto que as doenças crónicas como a Diabetes Mellitus e a Hipertensão tem uma relação estreita com a idade da população, importa conhecer a estrutura etária da população na região estudada. Em 2011, 23% da população residente na Região de Saúde do Centro tinha mais de 65 anos, 54% tinha entre 25 e 64 anos, 10% tinha entre 16 e 24 anos, e apenas 13% tinha menos de 15 anos de idade. A estrutura etária nos CS em estudo é semelhante à da região, embora a prevalência dos idosos (mais de 65 anos) seja maior no interior. Estes números confirmam o envelhecimento populacional no período de 2001-2011, pelo decréscimo de 12% de crianças e jovens com idade inferior a 15 anos, pelo decréscimo de 27% de adultos jovens entre os 15 e os 24 anos e um aumento de 14% de idosos com mais de 65 anos.

A Diabetes Mellitus foi responsável pelo aumento de 2,2% da taxa de internamento por diagnóstico ocorrido em 2012 por comparação ao período homólogo. Em 2012, a idade média de diagnóstico de Diabetes Mellitus foi de 73 anos; a esperança média de vida à nascença na Região Centro é atualmente de 79,9 anos.

Em relação aos Cuidados de Saúde Primários (CSP), a maioria dos utentes inscritos tem médico de família atribuído (92%) sendo que 0,2% dos não inscritos não têm médico de família por opção. Em 2012 existia na rede de CSP um total de 26.976 utentes diabéticos que apresentavam pelo menos 3 registos de HbA1C nos últimos 12 meses na Região de Saúde do Centro. Este indicador apresenta um discreto aumento (0,8%) relativamente ao ano anterior. Verificou-se também um aumento de 7,7% dos utentes com compromisso de vigilância no programa de diabetes, comparativamente a 2011.

Quadro 5:2 - Registos de controlo de Diabéticos por ACeS

ACeS	2011			2012		
	(1)	(2)	(3)	(1)	(2)	(3)
Pinhal Litoral II	2.387	5.853	40,8%	2.598	6.706	38,7%
Baixo Mondego I	2.859	6.532	43,8%	2.926	6.753	43,3%
Cova da Beira	202	2.018	10,0%	377	2.435	15,5%
Pinhal Interior Norte I	1.352	3.459	39,1%	1.409	3.675	38,3%
Região de Saúde do Centro	**24.563**	**61.422**	**40,0%**	**26.976**	**66.164**	**40,8%**

(1) N.º de utentes com pelo menos 3 registos de HbA1C
(2) N.º de utentes com compromisso de vigilância no programa de diabetes
(3) % de diabéticos com pelo menos 3 registos de HbA1C

A ARSC tem um programa de Prevenção e Controlo da Diabetes. Dada a frequente associação da diabetes com a hipertensão arterial e o colesterol elevado, o controlo destes dois fatores de

risco faz parte integrante do controlo da diabetes. O consumo de medicamentos para a diabetes em Portugal tem vindo a aumentar significativamente na última década - cerca de 24%, em termos da dose diária definida/1.000 habitantes/dia. As razões apontadas são, para além do aumento da prevalência da doença, o aumento do número e da proporção de pessoas tratadas, bem como as dosagens médias utilizadas nos tratamentos.

A prevalência da diabetes em 2011 foi de 12,7% na população portuguesa com idades compreendidas entre os 20 e os 79 anos, o que corresponde a um total de aproximadamente 1.003.000 indivíduos. Em 7,2% da população portuguesa referida, esta já havia sido diagnosticada e em 5,5% ainda não tinha sido diagnosticada. Existe ainda 26,4% da população com Hiperglicemia intermédia, o que eleva a população afetada a cerca de 3.000.000 de portugueses.

Em 2012, na rede de CSP da ARSC encontravam-se registados 129.589 utentes com diabetes, sendo 101.925 nas Unidades de Cuidados de Saúde Primários (UCSP) e 27.664 nas Unidades de Saúde Familiar (USF), num universo de 1.907.299 utentes registados. Verificaram-se em 2012 os seguintes indicadores relativamente ao período homólogo:

- Um aumento de 0,5% de inscritos com diagnóstico de diabetes;
- Um aumento de 1,5% de utentes diabéticos com compromisso de vigilância. Esta variação é observada pelo aumento do número de utentes com compromisso de vigilância (9,9%) e de utentes diagnosticados (7,7%);
- Um aumento de 3,4% de utentes com diabetes com pelo menos um exame dos pés, resultante de um maior aumento do número de utentes vigiados com pelo menos um registo de exame aos pés (16,3%) do que do número de utentes com compromisso de vigilância (9,9%);
- Um aumento de 2,7% de diabéticos com pelo menos 2 registos de HbA1c nos últimos 12 meses.

Estes números preocupantes em termos de saúde pública, embora em linha com a tendência mundial de aumento da prevalência da doença, levaram a orientações específicas do Conselho Diretivo da ARSC para implementação de algumas medidas nas consultas autónomas e multidisciplinares da diabetes: (1) os médicos de família deverão dispor de um período do seu horário para consultas de diabetes; autónomas, em equipa de pelo menos médico e enfermeiro e, se possível, com outros profissionais necessários ao tratamento e gestão da diabetes tipo 2 (caso de nutricionistas e psicólogos) e (2) incentivo da consulta de "pé diabético" a nível dos CSP. Estas Consultas existem em praticamente todos os hospitais da Região Centro havendo alguns que têm em funcionamento hospital de dia de diabetes.

A prevalência de retinopatia diabética em diabéticos tipo 1 é de cerca 40%, enquanto em diabéticos tipo 2 é de 20%. A faixa etária mais atingida situa-se entre 30-65 anos, sendo o sexo feminino mais afetado. Estão envolvidos no rastreio (diagnóstico) sistemático da retinopatia diabética todos os centros de saúde do âmbito territorial da ARSC (à exceção do ACeS Cova da Beira e outro) e o tratamento com laser é realizado em alguns dos serviços de oftalmologia. Em 2012 foram realizadas 18.496 retinografias.

Em termos de Recursos Humanos, em todos os ACeS, USF, UCSP e Hospitais, as respetivas direções designaram um profissional ou equipa de profissionais para serem interlocutores e responsáveis pela implementação de matérias relacionadas com o Programa de Prevenção e Controlo da Diabetes. Estes profissionais interagem periodicamente com a Equipa Coordenadora Regional.

5.2. Compreensão dos dados

A segunda etapa da metodologia compreende a recolha dos dados e a realização de uma análise qualitativa dos mesmos com vista a formular hipóteses e em consequência definir as tarefas e as técnicas utilizáveis.

A fase de recolha de dados compreende duas tarefas: (1) A seleção dos dados e (2) a extração dos dados.

A ARSC armazena muitos dados no decorrer da sua atividade; para este estudo são importantes somente os relacionados com o âmbito do estudo da Diabetes Mellitus e Hipertensão. Nesta tarefa é fundamental selecionar aqueles necessários para o estudo de entre todos os metadados disponíveis.

Da revisão da literatura realizada aos trabalhos de Data Mining aplicados à Diabetes Mellitus e Hipertensão foi recolhida a seguinte lista de atributos dos conjuntos de dados testados nos diversos trabalhos:

- Caraterísticas demográficas e gerais
  - Idade
  - Género sexual
  - Estado civil
  - Habilitações académicas
  - Óbito
  - Ocupação (e.g. profissão)
  - Rendimento mensal/anual
- Histórico da doença
  - Evolução do diagnóstico (Tipo II para Tipo I, Gestacional para Tipo I, etc.)
  - Nº de consultas
  - Nº de internamentos
  - Duração de diagnóstico (Há quantos anos a doença foi diagnosticada)
  - Familiares com doença
  - Nº de gravidezes (com e sem diabetes gestacional)
- Sintomas
  - Retinopatia - Lesão da retina que pode levar à perda de visão
  - Nefropatia - Lesão renal que pode levar à necessidade de hemodiálise
  - Neuropatia - Lesão nos nervos do organismo que pode levar a dores, sensação de calor ou frio, ou torpor nos pés e pernas
  - Hipoglicemia

- Hiperglicemia
- Doenças cardio e cérebrovasculares (angina de peito, ataques cardíacos e acidentes vasculares cerebrais)
- Úlceras/Pé diabético - Alterações nos pés que, pela doença arterial, neuropatia e infeções mais difíceis de combater, leva ao aparecimento de úlceras de difícil e prolongado tratamento que podem terminar pela necessidade de amputação
- Disfunção e impotência sexual masculina

- Sinais de alarme
  - Polidipsia - Sede constante e intensa
  - Poliúria - Urinar em maior quantidade e com mais regularidade
  - Polifagia - Muita fome e dificuldade em saciá-la
  - Xerostomia - Sensação de boca seca
  - Fadiga
  - Comichão no corpo (em especial nos genitais)
  - Visão turva
  - Enjoo e vómitos
  - Perda de peso
  - Infeções que demoram a curarem
  - Edema
  - Pieira
- Medidas antropométricas
  - Índice de massa corporal (IMC)
  - Peso
  - Altura
  - Outras: Perímetro da cintura, altura da perna, largura da coxa, etc.
- Testes
  - Tensão Arterial
    - Sistólica
    - Diastólica
  - Glicémia/Concentração de glicose
    - Glicose Instantânea ou Aleatória – *Random Blood Sugar (RBS) ou Instant Blood Sugar* (IBS) – Concentração de glicose em plasma medida sem levar em conta a última ingestão de comida ou bebida
    - Glicose em Jejum – *Fasting Blood Sugar* (FBS) – Concentração de glicose em plasma medido em jejum de pelo menos 8 horas
    - Glicose Pós-prandial – *2-hour Postprandial Blood Sugar* – Concentração de glicose em plasma medido 2 horas após a última refeição
    - Glicose Oral – *Oral Glucose Tolerance Level* – Concentração de glicose medido usando um teste oral (ingestão de bebida adocicada) de tolerância à glicose de 2 horas – usado tipicamente para grávidas

        - HbA1c (Hemoglobina glicada) – Mede a quantidade de glicose presa aos glóbulos vermelhos. Este teste permite o diagnóstico e a evolução da doença durante aproximadamente os últimos 3 meses
        - Hiperlipidemia - níveis anormalmente elevados de algum ou todos os lípidos e/ou lipoproteínas no sangue
        - Glicosuria – nível de glicose na urina
        - Proteinuria geral – nível de presença em excesso de proteínas
    - Colesterol
        - Lipoproteínas de alta densidade (HDL)
        - Lipoproteínas de baixa densidade (LDL)
    - Triglicéridos
    - Nível de potássio
    - Nível de sódio
- Fatores de risco relacionados com o estilo de vida
    - Tabaco
    - Álcool
    - Chá ou café
    - Alimentação (Tipo de Dieta)
        - Tipo de carne ou peixe
        - Preferência por comida doce ou salgada
    - *Stress*
    - Prática de exercício físico
    - Duração do sono
- Prescrições
    - Terapia anti-hipertensiva
    - Medicamentos
    - Dieta
    - Redução de peso
    - Cessação tabágica
    - Exercício físico
    - Insulina
        - Monoterapia (medicação com 1 princípio ativo)
        - Biterapia (medicação com 2 princípios ativos)
        - Triterapia (medicação com 3 princípios ativos)

A lista de atributos foi passada para a ARSC que se encarregou da extração dos dados do seu Data Warehouse, concluída em 5 de março de 2014. Os dados foram extraídos por Centro de Saúde e distribuídos por diversos ficheiros devido à elevada cardinalidade, conforme se mostra no Quadro 5:3 seguinte.

Quadro 5:3 - Resumo global da distribuição da extração de dados por diversos ficheiros

Dados CS - Arnaldo Sampaio	15 ficheiros	448.059.609 bytes

Dados CS - Eiras	54 ficheiros	322.751.200 bytes
Dados CS - Fundão	247 ficheiros	1.955.862.872 bytes
Dados CS - Tábua	46 ficheiros	266.508.615 bytes

Os dados existentes foram analisados e, devido à sua elevada dimensionalidade e cardinalidade, divididos em 5 partes:

1. Utentes;
2. Consultas;
3. MCDT;
4. Prescrições;
5. Diabetes.

Nos quadros que se seguem são mostrados os mapeamentos entre a informação solicitada e a extraída, com descrição sucinta do seu significado e a indicação do nome do campo, em cada parte respetiva.

O Quadro 5:4 diz respeito a dados gerais sobre utentes. Contudo, verifica-se a existência de dados estreitamente ligados às patologias em estudo como os testes de tensão arterial e vários sintomas. Observa-se ainda a existência de alguns campos extraídos que não foram solicitados que se devem à estruturação do modelo dimensional do Data Warehouse e das respetivas tabelas de armazenamento dos dados bem como de atributos julgados pertinentes pela ARSC.

Quadro 5:4 - Metadados extraídos de caraterização geral dos utentes

Utentes			
Atributos desejados		**Atributos extraídos**	
Categoria	**Atributo**	**Nome**	**Descrição**
		COD_D_UTENTE	Chave Sintética da Dimensão Utente
Histórico da doença	Nº de consultas	DATA_CONSULTA	Data da consulta
		NUM_EPISODIO	Número do Episódio
		NUMERO_INSCR_CENTRO_SAUDE	Número da Inscrição no Centro de Saúde
		CODIFICACAO_CENTRO_SAUDE	Código do Centro de Saúde
		NOME_UTENTE	Nome do Utente
	Género sexual	GENERO_SEXUAL	Género Sexual do Utente
Caraterísticas demográficas e gerais	Óbito	OBITO	Óbito
	Ocupação (profissão)	SITUACAO_PROFISSIONAL	Situação Profissional do Utente
	Idade	IDADE_DATA	Idade à data da consulta
	Idade	IDADE_ACTUAL	Idade atual (à data da extração da informação)
	Índice de massa corporal (IMC)	IMC	Índice de massa corporal
Medidas antropométricas	Peso	PESO	Peso
	Altura	ALTURA	Altura
	Perímetro da cintura	PERIMETRO_CINTURA	Perímetro da cintura

	Nefropatia	DSC_NEFROPATIA	Nefropatia
Sintomas	Retinopatia	DSC_RETINOPATIA	Retinopatia
	Neuropatia	DSC_NEUROPATIA	Neuropatia
	Doenças cárdeo e cérebro vasculares	ANO_ACI_VASC_CEREB	Ano do AVC
	Doenças cárdeo e cérebro vasculares Retinopatia	ANO_ENFT_MIOCARDIO	Ano de Enfarte do Miocárdio
		ANO_CEGUEIRA	Ano de Cegueira
	Nefropatia	ANO_INSUF_RENAL	Ano de Insuficiência Renal
	Úlceras/Pé diabético	ANO_AMPUT_ABAIXO_TORNEZELO	Ano de Amputação, abaixo do tornozelo
	Úlceras/Pé diabético Tensão Arterial Diastólica	ANO_AMPUT_ACIMA_TORNEZELO	Ano de Amputação, acima do tornozelo
		PRESSAO_ARTERIAL_DIASTOLICA	Valor da Tensão Arterial Diastólica
Testes	Tensão Arterial Sistólica	PRESSAO_ARTERIAL_SISTOLICA	Valor da Tensão Arterial Sistólica
		COD_PROG_SAUDE	Código de programa de saúde associado
		DSC_PROG_SAUDE	Descrição do programa de saúde associado
	Nº de gravidezes	N_GRAVIDEZES	Número de gravidezes
Histórico da doença			

O Quadro 5:5 é a caraterização da consulta de acordo com as diferentes fases (Motivo, Avaliação e Procedimento). Inclui ainda a caraterização de acordo com a Classificação Internacional de Cuidados de Saúde Primários (ICPC).

Quadro 5:5 - Metadados extraídos de caraterização geral das consultas

Consultas			
Atributos desejados		**Atributos extraídos**	
Categoria	**Atributo**	**Nome**	**Descrição**
		COD_D_UTENTE	Chave Sintética da Dimensão Utente
		NUM_EPISODIO	Número do Episódio
Histórico da doença	Nº de consultas	DATA_CONSULTA	Data da consulta
		NUMERO_INSCR_CENTRO_SAUDE	Número da Inscrição no Centro de Saúde
		CODIFICACAO_CENTRO_SAUDE	Código do Centro de Saúde
		NOME_UTENTE	Nome do Utente
Caraterísticas demográficas e gerais	Género sexual	GENERO_SEXUAL	Género Sexual do Utente
	Óbito	OBITO	Óbito
		COD_SOAP	Código das Fases da Consulta

		DSC_SOAP	Descrição das Fases da Consulta (Motivo; Diagnóstico e Procedimento)
		COD_ICPC_MASTER	Código ICPC Geral
		DSC_ICPC_MASTER	Descrição ICPC Geral
		COD_ICPC_DETAIL	Código ICPC Detalhe
		DSC_ICPC_DETAIL	Descrição ICPC Detalhe

O Quadro 5:6 contém dados referentes aos meios complementares de diagnóstico e terapêutica, vulgo exames, que engloba exames laboratoriais, imagiológicos, colheita de amostras por meios mais ou menos invasivos, e atos de tratamento variados, realizados em regime ambulatório ou em internamento hospitalar.

Quadro 5:6 - Metadados extraídos de caraterização geral dos MCDT

MCDT			
Atributos desejados		**Atributos extraídos**	
Categoria	**Atributo**	**Nome**	**Descrição**
		Chave Sintética da Dimensão Utente	COD_D_UTENTE
		Número do Episódio	NUM_EPISODIO
Histórico da doença	Nº de consultas	Data da consulta	DATA_CONSULTA
		Número da Inscrição no Centro de Saúde	NUMERO_INSCR_CENTRO_SAUDE
		Código do Centro de Saúde	CODIFICACAO_CENTRO_SAUDE
		Nome do Utente	NOME_UTENTE
Caraterísticas demográficas e gerais	Género sexual	Género Sexual do Utente	GENERO_SEXUAL
	Óbito	Óbito	OBITO
Testes	vários	Código do exame realizado	COD_MCDT_RESULTADO
		Descrição do exame realizado	DSC_MCDT_RESULTADO
		Descrição ao detalhe do exame realizado	DSC_ITEM_MCDT_RESULTADO
		Resultado do exame realizado	VAL_RESULTADO

O Quadro 5:7 contém dados referentes às prescrições medicamentosas realizadas. Os medicamentos estão descritos pelo nome comercial e pela Denominação Comum Internacional (DCI). Está ainda incluída a forma de embalagem e quantidade de embalagens prescritas.

Quadro 5:7 - Metadados extraídos de caraterização geral das prescrições

Prescrições			
Atributos desejados		**Atributos extraídos**	
Categoria	**Atributo**	**Nome**	**Descrição**
		COD_D_UTENTE	Chave Sintética da Dimensão Utente

		NUM_EPISODIO	Número do Episódio
Histórico da doença	Nº de consultas	DATA_CONSULTA	Data da consulta
		NUMERO_INSCR_CENTRO_SAUDE	Número da Inscrição no Centro de Saúde
		CODIFICACAO_CENTRO_SAUDE	Código do Centro de Saúde
		NOME_UTENTE	Nome do Utente
Caraterísticas demográficas e gerais	Género sexual	GENERO_SEXUAL	Género Sexual do Utente
	Óbito	OBITO	Óbito
Prescrições	vários	COD_RECEITA	Código da receita
		COD_MEDICAMENTO	Código do medicamento
		DSC_MEDICAMENTO	Descrição do medicamento
		COD_MEDICAMENTO_DCI	Código DCI do medicamento
		DSC_MEDICAMENTO_DCI	Descrição DCI do medicamento
		COD_EMBALAGEM	Código da embalagem
		DSC_EMBALAGEM	Descrição da embalagem
		VAL_QTD	Quantidade na receita

O Quadro 5:8 inclui dados sobre o diagnóstico e a evolução da Diabetes Mellitus.

Quadro 5:8 - Metadados extraídos de caraterização geral da Diabetes Mellitus

Diabetes			
Atributos desejados		**Atributos extraídos**	
Categoria	**Atributo**	**Nome**	**Descrição**
		COD_D_UTENTE	Chave Sintética da Dimensão Utente
Histórico da doença	Nº de consultas	DATA_CONSULTA	Data da consulta
		NUM_EPISODIO	Número do Episódio
		NUMERO_INSCR_CENTRO_SAUDE	Número da Inscrição no Centro de Saúde
		CODIFICACAO_CENTRO_SAUDE	Código do Centro de Saúde
		NOME_UTENTE	Nome do Utente
Caraterísticas demográficas e gerais	Género sexual	GENERO_SEXUAL	Género Sexual do Utente
	Óbito	OBITO	Óbito
Histórico da doença	Evolução do diagnóstico	TIPO_DIABETES	Tipo de Diabetes (I, II, Gestacional)
		DTA_INICIO_PROB	Data de início dos Diabetes
		DTA_FIM_PROB	Data de fim dos Diabetes

Histórico da doença	Duração do diagnóstico	N_ANOS_DIAGNOSTICADO	Duração de diagnóstico em anos face à data de início e de fim

Após a indicação dos metadados extraídos, é possível verificar no Quadro 5:9 a inexistência de uma quantidade muito significativa de dados usados na literatura e que, por inexistência no Data Warehouse, limita de forma crítica a elaboração do estudo:

- O rendimento e o nível académico têm sido associados a um cuidado maior tanto na prevenção como na longevidade de pacientes;
- O conhecimento da história familiar da doença é de natureza fundamental no diagnóstico visto que a mesma está normalmente relacionada com aspetos hereditários;
- Os sinais de alarme podem ajudar no diagnóstico precoce; seria de importância crucial conhecer os sinais que mais estão associados à doença para divulgação e prevenção futura;
- Muitos estudos provam que a adoção de estilos de vida correta favorece a prevenção e a longevidade.

Quadro 5:9 - Metadados não extraídos por inexistência no Data Warehouse

Atributos desejados inexistentes para análise	
Caraterísticas demográficas e gerais	Estado civil
	Habilitações académicas
	Rendimento mensal
Histórico da doença	N° de internamentos
	Familiares com doença
Sintomas	Hipoglicemia
	Hiperglicemia
	Disfunção e impotência sexual masculina
Sinais de alarme	Polidipsia
	Poliúria
	Polifagia
	Xerostomia
	Fadiga
	Comichão no corpo (em especial nos genitais)
	Visão turva
	Enjoo e vómitos
	Perda de peso
	Infeções que demoram a curar
	Edema
	Pieira
Medidas antropométricas	Altura da perna, largura da coxa, etc.
Fatores de risco relacionados com o estilo de vida	Tabaco
	Álcool

	Chá ou café
	Alimentação - Tipo de carne ou peixe
	Alimentação - Preferência por comida doce ou salgada
	Stress
	Prática de exercício físico
	Duração do sono

A realização de uma análise qualitativa dos dados extraídos permite compreender em primeira mão os valores existentes através de técnicas de visualização e sumarização. A importância de conhecer os dados deriva da sua relação com as hipóteses a serem formuladas.

Os dados revelam registos de utentes com dados compreendidos em períodos diferentes para cada Centro de Saúde, a saber:

- Arnaldo Sampaio – 01/Jan/2011 a 21/Jan/2013;
- Eiras – 01/Jan/2011 a 18/Dez/2013;
- Fundão – 01/Jan/2011 a 05/Jul/2013;
- Tábua – 01/Jan/2011 a 19/Fev/2013.

Embora os CS tenham dados diferentes, são estruturalmente idênticos e os dados partilham o significado e forma, pelo que a análise dos dados de um CS seja suficiente na identificação dos problemas e correções a realizar. Em seguida apresenta-se a análise realizada aos registos de utentes, consultas, prescrições e diabetes do CS Arnaldo Sampaio e ao registo de MCDT do CS Eiras por terem um tamanho de dados que permite a análise através do Microsoft Excel 2010. O CS com maior volume de dados é o Fundão cujos registos não cabem no tamanho máximo de uma folha de cálculo.

No registo de utentes do CS Arnaldo Sampaio constam 371.089 registos, cujos atributos são caraterizados conforme se indica de seguida:

- Código de Utente (COD_D_UTENTE):
  - 53.679 registos unívocos (n.º de utentes);
  - Série de valores [210100000055836-210100910109101].
- Data de Consulta (DATA_CONSULTA):
  - 666 registos unívocos (n.º de dias com consultas);
  - 01/Jan/2011 a 21/Jan/2013.
- Episódio/consulta (NUM_EPISODIO):
  - 340.180 registos unívocos (n.º de consultas);
  - Série de valores [111-21841611].
- Número de inscrição no CS (NUMERO_INSCR_CENTRO_SAUDE):
  - 53.679 registos unívocos (n.º de utentes);
  - Série de valores [55836-910109101].
- Código de CS (CODIFICACAO_CENTRO_SAUDE):
  - 31.806 registos unívocos (códigos internos);
  - Série de valores [0-2553101].

- Nome de Utente (NOME_UTENTE):
  - 52.813 registos unívocos (nomes diferentes);
  - 15 nomes estão precedidos de um espaço;
  - Alguns registos contêm somente 1 nome e 1 apelido, o que explica a diferença de 866 entre o n.º de utentes e o n.º de nomes.
- Género Sexual (GENERO_SEXUAL):
  - 53.683 registos unívocos, agrupados por utente;
  - A distribuição verifica 29.667 registos de género Feminino e 24.016 Masculino.
  - A diferença de 4 registos com o n.º de utentes deve-se a que alguns utentes contenham registos com ambos os géneros: 210100120101054, 210100120101055, 210100070024702 e 210100060043875.
- Óbito (OBITO):
  - 53.874 registos unívocos, agrupados por utente;
  - 354 utentes registados com valor "1" (óbito), os restantes estão em branco;
  - A diferença de 195 registos com o n.º de utentes deve-se a haver utentes que possivelmente faleceram no período de registos de consultas, existindo com ambos os valores possíveis.
- Ocupação (SITUACAO_PROFISSIONAL):
  - 54.030 registos unívocos, agrupados por utente;
  - Os registos distribuem-se da seguinte forma:

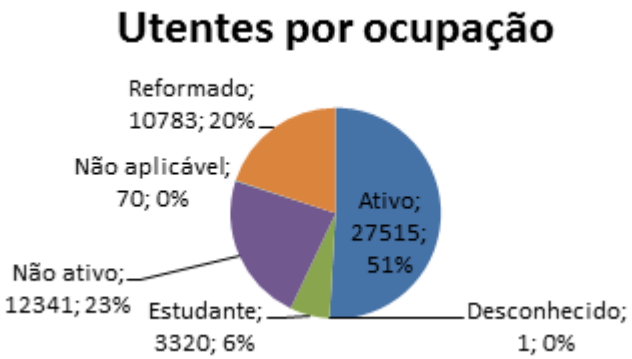

Figura 6 – Distribuição de utentes do CS Arnaldo Sampaio por ocupação

A diferença de 351 registos entre o n.º de utentes e registos de ocupação é devida a utentes com registos em mais que uma ocupação em consultas diferentes.

- Idade à data de consulta (IDADE_DATA)
  - 106.854 registos unívocos, agrupados por utente;
  - Série de valores [0-103];

- Importa notar que um utente que tenha sido consultado em diferentes alturas pode surgir com valores de idade diferentes, tanto em anos diferentes como no mesmo ano, mas em tempos diferentes.

- Idade Atual (IDADE_ATUAL)
  - 53.681 registos unívocos, agrupados por utente;
  - Série de valores [1-105];
  - A diferença entre os registos unívocos e o n.º de utentes deve-se a existirem 2 utentes com idades diferentes no mesmo período. São eles o n.º 210100110007602 e o n.º 210100160092636;
  - Os registos distribuem-se da seguinte forma:

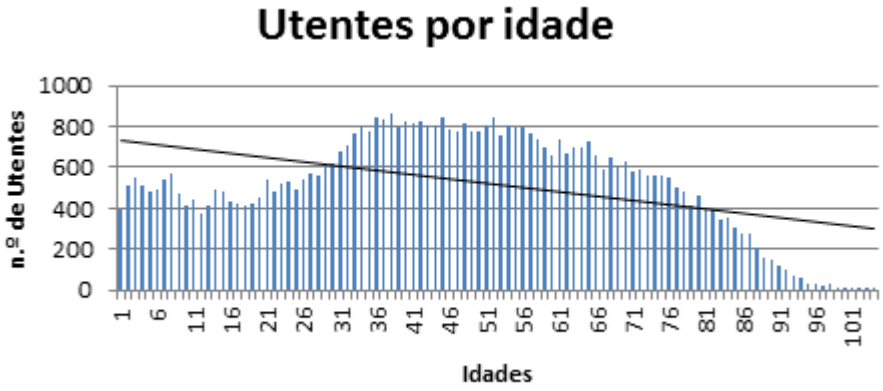

Figura 7 - Distribuição de utentes do CS Arnaldo Sampaio por idade

- Índice de Massa Corporal (IMC):
  - 81.665 registos unívocos, agrupados por utente;
  - 52.012 registos estão em branco;
  - Série de valores [0-660.000];
  - Apesar de não haver um limite inferior e superior, de acordo com a (World Health Organization, 2014), o IMC situa-se: (1) do nascimento aos 5 anos entre 11 e 20, (2) entre os 6 e 19 anos entre os 12 e 32 e (3) acima de 20 anos entre 17 (extrema magreza) e 40 (extrema obesidade). Os valores não compreendidos dentro destes intervalos devem ser vistos com cuidado pois podem representar dados incorretos.
  - Esta coluna resulta dos valores de peso e altura pelo que os dados erróneos devem ser vistos em combinação com estes atributos.
- Peso (PESO):
  - 89.196 registos unívocos, agrupados por utente;
  - 51.734 registos estão em branco e 3 estão com valor 0;
  - Série de valores [0-12.990];
  - Os dados de peso devem ser vistos em relação à idade, por exemplo valores de 1,50 Kg podem ser corretos para recém-nascidos, mas certamente não serão para outras idades;

- Observa-se que a maioria dos dados está registada em Kg embora os valores superiores pareçam ser erros de confusão de unidade de medida com g.;

- Altura (ALTURA):
  - 72.416 registos unívocos, agrupados por utente;
  - 51.991 registos estão em branco e 2 estão com valor 0;
  - Série de valores [0-16.114];
  - Observa-se que a maioria dos dados está registada em cm embora em 75 registos com valores inferiores a 1,83 inclusive, pareçam ser erros de confusão de unidade de medida com m.;
  - Os dados com valores acima de 505 inclusive parecem ser casos de falta de introdução da casa decimal embora acima de 2011 sejam inexplicáveis;
  - Os valores entre 2,73 e 33,5 contêm erros, em alguns casos o valor da altura é igual ao peso;
  - Existem registos de valores baixos de altura relativamente à idade, mas que podem não ser dados incorretos. Veja-se, por exemplo, o caso extremo de 2 utentes de 29 e 68 anos de idade, ambos com 70 cm de altura. Recorde-se que o nanismo extremo ocorre a partir dos 60 cm, pelo que estes valores podem estar corretos.
- Perímetro abdominal (PERIMETRO_CINTURA):
  - 58.974 registos unívocos, agrupados por utente;
  - 53.658 registos estão em branco;
  - A relação entre o perímetro abdominal e o peso pode ser observada para detetar erros nos dados. Vejamos alguns casos verificados nos dados: (1) perímetro abdominal de 212 cm não encontra correspondência num peso de 72 Kg e (2) perímetro abdominal de 155 cm não corresponde a um peso de 71 Kg, mas o mesmo perímetro faz todo o sentido para o peso de 127 Kg;
  - O perímetro abdominal à nascença é de ligeiramente acima de 30 cm, logo alguém com um perímetro de 31 cm, 25 anos de idade e valores de peso e altura normais não pode estar correto;
  - Os valores deste atributo abaixo de 31 inclusive e acima de 180 não estão corretos na correspondência com outros dados e os restantes carecem de análise cuidada especialmente nas caudas da distribuição.
- Nefropatia (DSC_NEFROPATIA):
  - 53.707 registos unívocos, agrupados por utente;
  - 53.679 registos estão em branco;
  - 28 apresentam a classe “Ativa”;
  - A diferença em relação ao n.º de utentes deve-se a que 28 utentes contenham registos com o atributo em ambas as classes: em branco e “Ativa”.
- Retinopatia (DSC_RETINOPATIA):
  - 54.646 registos unívicos, agrupados por utente;
  - 53.676 registos estão em branco;
  - 618 apresentam a classe “Consulta” e 352 a classe “Referenciação”;

    - A diferença em relação ao n.º de utentes deve-se a que 967 utentes contenham registos com o atributo em mais que uma classe.
- Neuropatia (DSC_NEUROPATIA):
    - 55.427 registos unívocos, agrupados por utente;
    - 53.672 registos estão em branco;
    - 1.298 apresentam a classe "Sim" e 457 a classe "Não";
    - A diferença em relação ao n.º de utentes deve-se a que 1.748 utentes contenham registos com o atributo em mais que uma classe.
- AVC (ANO_ACI_VASC_CEREB):
    - 53.698 registos unívocos, agrupados por utente;
    - 53.679 registos estão em branco;
    - Série de valores [1992-2012];
    - A diferença em relação ao n.º de utentes deve-se a que 19 utentes contenham registos com o atributo em mais que uma classe.
- Enfarte do miocárdio (ANO_ENFT_MIOCARDIO):
    - 53.700 registos unívocos, agrupados por utente;
    - 53.679 registos estão em branco;
    - Série de valores [1-2012];
    - O registo com classe "1" diz respeito ao utente 210100060010608 e significa a ocorrência e não o ano do enfarte do miocárdio. Sem este a série de valores seria [1988-2012];
    - A diferença em relação ao n.º de utentes deve-se a que 21 utentes contenham registos com o atributo em mais que uma classe.
- Cegueira (ANO_CEGUEIRA):
    - 53.680 registos unívocos, agrupados por utente;
    - 53.679 registos estão em branco;
    - 1 registo apresenta a classe "1999";
    - O utente do registo anterior apresenta também um registo em branco, o que explica a diferença em relação ao n.º de utentes.
- Insuficiência renal (ANO_INSUF_RENAL):
    - 53.681 registos unívocos, agrupados por utente;
    - 53.679 registos estão em branco;
    - 1 utente apresenta o ano 2009 e outro o ano 2011;
    - Os 2 utentes dos registos anteriores apresentam também um registo em branco, o que explica a diferença em relação ao n.º de utentes.
- Amputação abaixo do tornozelo (ANO_AMPUT_ABAIXO_TORNEZELO):
    - 53.681 registos unívocos, agrupados por utente;
    - 53.679 registos estão em branco;
    - 1 utente apresenta o ano 2000 e outro o ano 2003;
    - Os 2 utentes dos registos anteriores apresentam também um registo em branco, o que explica a diferença em relação ao n.º de utentes.

- Amputação acima do tornozelo (ANO_AMPUT_ACIMA_TORNEZELO):
  - 53.679 registos unívocos, agrupados por utente;
  - 53.679 registos estão em branco;
- TAD (PRESSAO_ARTERIAL_DIASTOLICA):
  - 88.339 registos unívocos, agrupados por utente;
  - 52.274 registos estão em branco;
  - Série de valores [0-186] mm Hg;
  - Os recém-nascidos de nascimento prematuro e/ou com baixo peso à nascença apresentam valores de TAD acima de 20 e que sobe acima de 30 após os primeiros dias de vida. A análise destes valores em combinação com a idade pode ajudar a detetar erros de introdução dos dados;
  - A diferença em relação ao n.º de utentes deve-se a que 34.660 utentes contenham mais que um registo em que o valor da medição seja diferente. O facto de existirem várias medições para cada utente permite analisar o progresso.
- TAS (PRESSAO_ARTERIAL_SISTOLICA):
  - 89.443 registos unívocos, agrupados por utente;
  - 52.272 registos estão em branco;
  - Série de valores [1-250] mm Hg;
  - Os recém-nascidos de nascimento prematuro e/ou com baixo peso à nascença apresentam valores de TAS acima de 45 e que sobe acima de 55 após os primeiros dias de vida. A análise destes valores em combinação com a idade pode ajudar a detetar erros de introdução dos dados;
  - A diferença em relação ao n.º de utentes deve-se a que 35.764 utentes contenham mais que um registo em que o valor da medição seja diferente. O facto de existirem várias medições para cada utente permite analisar o progresso.
- Programa de Saúde (COD_PROG_SAUDE e DSC_PROG_SAUDE):
  - Os atributos correspondem ao código e à descrição do programa de saúde pelo que serão aqui tratados em conjunto;
  - 81.632 registos unívocos, agrupados por utente;

- Os registos distribuem-se da seguinte forma:

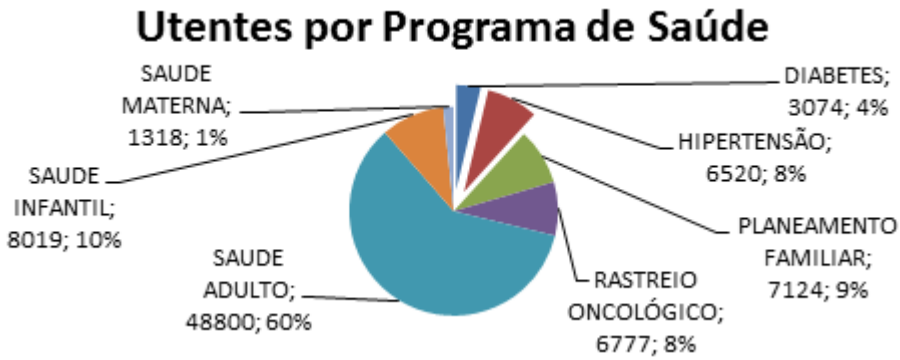

Figura 8 - Distribuição de utentes do CS Arnaldo Sampaio por programa de saúde

- Gravidez (N_GRAVIDEZES):
  - 54.404 registos unívocos, agrupados por utente;
  - 53.662 registos mostram a classe "0" e 742 a classe "1";
  - A diferença em relação ao n.º de utentes deve-se a que 725 utentes contenham registos com o atributo em mais que uma classe.

No registo de Consultas do CS Arnaldo Sampaio constam 571.681 registos, cujos atributos são caraterizados conforme se indica de seguida:

- Código de Utente (COD_D_UTENTE):
  - 38.797 registos unívocos (n.º de utentes);
  - Série de valores [210100000055836- 210100910109101];
  - Em comparação com o mesmo atributo no registo de Utentes, verifica-se que 14.882 utentes não têm registos de consultas.
- Episódio/consulta (NUM_EPISODIO):
  - 158.552 registos unívocos (n.º de consultas);
  - Série de valores [112-21841011];
  - Em comparação com o mesmo atributo no registo de Utentes, verifica-se que 181.628 episódios não têm registo de consultas.
- Data de Consulta (DATA_CONSULTA):
  - 629 registos unívocos (n.º de dias com consultas);
  - 02/Jan/2011 a 21/Jan/2013;
  - Em comparação com o mesmo atributo no registo de Utentes, verifica-se que existem 37 dias que não constam do registo de consultas.
- Número de inscrição no CS (NUMERO_INSCR_CENTRO_SAUDE):
  - 38.797 registos unívocos (n.º de utentes);
  - Série de valores [55836-910109101];
  - Em comparação com o mesmo atributo no registo de Utentes, verifica-se que 14.882 números de inscrição (o mesmo que código de utente) não constam do registo de consultas.

- Código de CS (CODIFICACAO_CENTRO_SAUDE):
  - 26.626 registos unívocos (códigos internos);
  - Série de valores [101-2553101];
  - Em comparação com o mesmo atributo no registo de Utentes, verifica-se que 5.180 códigos não constam do registo de consultas, e.g. entre 0 e 100;
- Nome de Utente (NOME_UTENTE):
  - 38.858 registos unívocos (nomes diferentes);
  - Alguns nomes estão precedidos de um espaço;
  - Alguns registos contêm somente 1 nome e 1 apelido, o que explica a diferença de 61 entre o n.º de utentes e o n.º de nomes;
  - Em comparação com o mesmo atributo no registo de Utentes, verifica-se que 13.954 registos não constam do registo de consultas.
- Género Sexual (GENERO_SEXUAL):
  - 38.800 registos unívocos, agrupados por utente;
  - A distribuição verifica 22.392 registos de género Feminino e 16.408 Masculino;
  - A diferença de 3 registos com o n.º de utentes deve-se a que alguns utentes contenham registos com ambos os géneros: 210100120101054, 210100120101055 e 210100070024702;
  - Em comparação com o mesmo atributo no registo de Utentes, verifica-se que 14.883 registos não constam do registo de consultas.
- Óbito (OBITO):
  - 38.875 registos unívocos, agrupados por utente;
  - 161 utentes registados com valor "1" (óbito), os restantes estão em branco;
  - A diferença de 78 registos com o n.º de utentes deve-se a haver utentes que faleceram no período de registos de consultas, existindo com ambos os valores possíveis;
  - Em comparação com o mesmo atributo no registo de Utentes, verifica-se que 14.999 registos não constam do registo de consultas.
- Fases da consulta (COD_SOAP e DSC_SOAP):
  - Os atributos correspondem ao código e à descrição da fase da consulta pelo que serão aqui tratados em conjunto;
  - 96.230 registos unívocos, agrupados por utente;

- Os registos distribuem-se da seguinte forma:

Figura 9 - Distribuição de consultas do CS Arnaldo Sampaio por fases

- ICPC Classe (COD_ICPC_MASTER e DSC_ ICPC_MASTER):
  - Os atributos correspondem ao código e à descrição da Classificação Internacional de Cuidados de Saúde Primários pelo que serão aqui tratados em conjunto;
  - 107.607 registos unívocos, agrupados por utente;
  - Os registos distribuem-se da seguinte forma:

Quadro 5:10 - Distribuição de utentes por classe ICPC

Descrição da classe ICPC	#Utentes	%
APARELHO CIRCULATORIO (3)	9.682	9%
APARELHO DIGESTIVO	7.234	7%
APARELHO GENITAL FEMININO (INCLUINDO MAMA) (6)	3.164	3%
APARELHO GENITAL MASCULINO	1.634	2%
APARELHO RESPIRATORIO	9.907	9%
APARELHO URINARIO	3.385	3%
ENDOCRINO,METABOLICO E NUTRICIONAL (1)	9.089	8%
GERAL E INESPECIFICO	25.774	24%
GRAVIDEZ E PLANEAMENTO FAMILIAR (2)	6.606	6%
OLHOS	1.815	2%
OUVIDOS	2.186	2%
PELE	6.058	6%
PROBLEMAS SOCIAIS	730	1%
PSICOLOGICO	5.973	6%
SANGUE,ORGAOS HEMATOPOIETICOS E LINFATICOS (BACO,MEDULA OSSEA)	1.406	1%
SISTEMA MUSCULO-ESQUELETICO	10.193	9%
SISTEMA NERVOSO (5)	2.771	3%
Total Geral	**107.607**	**100%**

Legenda:
(1) Inclui: Diabetes insulino-dependente e diabetes não insulino-dependente;
(2) Inclui: Diabetes gestacional
(3) Inclui: Hipertensão sem complicações, tensão arterial elevada e hipotensão postural
(4) Inclui: Sensação de tensão,
(5) Inclui: Cefaleia de tensão
(6) Inclui: Síndrome de tensão pré-menstrual

- ICPC Subclasse (COD_ICPC_MASTER e DSC_ ICPC_MASTER):
  - Os atributos correspondem ao código e à descrição da Classificação Internacional de Cuidados de Saúde Primários pelo que serão aqui tratados em conjunto;
  - 298.252 registos unívocos, agrupados por utente;
  - Os registos distribuem-se por 681 subclassificações. Em seguida apresentam-se as 3 maiores e as 3 menores, bem como as que incluem explicitamente o termo Diabetes e relacionam-se com tensão arterial:

Quadro 5:11 - Distribuição de utentes por subclasse ICPC

Pos.	**Descrição da subclasse ICPC**	**#Utentes**	**%**
1	MEDICAÇÃO-PRESCRIÇÃO / PEDIDO / RENOVAÇÃO / INJECÇÃO	29.348	10%
2	MEDICINA PREVENTIVA / DE ACOMPANHAMENTO GERAL	19.336	6%
3	ANALISES DE SANGUE	18.525	6%
(...)			
13	HIPERTENSÃO SEM COMPLICAÇÕES	5.984	2%
17	DIABETES NÃO INSULINO-DEPENDENTE	2.869	1%
39	SENSAÇÃO DE ANSIEDADE / NERVOSISMO / TENSÃO	1.070	0%
49	TENSÃO ARTERIAL ELEVADA	881	0%
137	DIABETES INSULINO-DEPENDENTE	270	0%
290	CEFALEIA DE TENSÃO	76	0%
358	HIPOTENSÃO POSTURAL	50	0%
520	DIABETES GESTATIONAL	13	0%
581	SINDROME DE TENSÃO PRE-MENSTRUAL	6	0%
			(...)
679	DESIDRATAÇÃO	1	0%
680	CORPO ESTRANHO NO NARIZ / LARINGE / BRONQUIOS	1	0%
681	CONTRACEPÇÃO POS-COITAL	1	0%
Total Geral		**298.252**	**100%**
Nota: A relação classe com subclasse está indicada no Quadro 5:10.			

No registo de MCDT do CS Eiras constam 518.560 registos, cujos atributos são caraterizados conforme se indica de seguida:

- Código de Utente (COD_D_UTENTE):
  - 16.371 registos unívocos (n.º de utentes);
  - Série de valores [206270000000001-206270910028568].
- Episódio/consulta (NUM_EPISODIO):
  - 148.993 registos unívocos (n.º de consultas);
  - Série de valores [111-6701213].
- Data de Consulta (DATA_CONSULTA):
  - 721 registos unívocos (n.º de dias com consultas);

- 03/Jan/2011 a 18/Dez/2013.

- Número de inscrição no CS (NUMERO_INSCR_CENTRO_SAUDE):
  - 16.371 registos unívocos (n.º de utentes);
  - Série de valores [1-910028568].
- Código de CS (CODIFICACAO_CENTRO_SAUDE):
  - 14.678 registos unívocos (códigos internos);
  - Série de valores [101-9900028801].
- Nome de Utente (NOME_UTENTE):
  - 17.313 registos unívocos (nomes diferentes);
  - 7 nomes estão precedidos de um espaço;
  - Alguns registos contêm somente 1 nome e 1 apelido, o que explica a diferença de 942 entre o n.º de utentes e o n.º de nomes.
- Género Sexual (GENERO_SEXUAL):
  - 16.378 registos unívocos, agrupados por utente;
  - A distribuição verifica 9.004 registos de género Feminino e 7.374 Masculino.
  - A diferença de 7 registos com o n.º de utentes deve-se a que alguns utentes contenham registos com ambos os géneros: 206270040006083, 206270040008129, 206270040007270, 206270000027695, 206270010017330, 206270040026771 e 206270000024617.
- Óbito (OBITO):
  - 16.482 registos unívocos, agrupados por utente;
  - 121 utentes registados com valor "1" (óbito), os restantes estão em branco;
  - A diferença de 111 registos com o n.º de utentes deve-se a haver utentes que possivelmente faleceram no período de registos de consultas, existindo com ambos os valores possíveis.
  - O registo de utentes deste CS apresenta todos os registos deste atributo em branco, o que neste registo não se verifica. Sugere assim que a criação de um conjunto de dados único deve ser feita combinando os atributos de todos os registos;
- Código do MCDT realizado (COD_MCDT_RESULTADO) e Descrição (DSC_MCDT_RESULTADO):
  - Os atributos correspondem ao código e à descrição do MCDT pelo que serão aqui tratados em conjunto;
  - 424 registos unívocos (n.º de MCDT);
  - Série de valores do código [4674-9318];
  - 198.164 registos unívocos, agrupados por utente;
  - 16.291 registos estão em branco;
  - Foram realizados uma média de 11,11 MCDT a cada utente.
- Descrição dos itens do MCDT realizado (DSC_ITEM_MCDT_RESULTADO):
  - 407 registos unívocos (n.º de detalhe de MCDT);
  - 195.402 registos unívocos, agrupados por utente;

- Embora existam MCDT com apenas um item examinado, e.g. HGB A1C, existem outros onde vários itens são avaliados, e.g. um hemograma.

- Resultado do item do MCDT realizado (VAL_RESULTADO):
  - O resultado está associado com o item do MCDT e não com o MCDT que pode incluir vários itens;
  - Este campo inclui valores numéricos e descrições alfanuméricas, de acordo com o tipo de MCDT realizado;
  - 143.842 registos de 16.286 utentes estão em branco.

No registo de Prescrições do CS Arnaldo Sampaio constam 806.874 registos, cujos atributos são caraterizados conforme se indica de seguida:

- Código de Utente (COD_D_UTENTE), Episódio/consulta (NUM_EPISODIO), Data de Consulta (DATA_CONSULTA), Número de inscrição no CS (NUMERO_INSCR_CENTRO_SAUDE), Código de CS (CODIFICACAO_CENTRO_SAUDE), Nome de Utente (NOME_UTENTE), Género Sexual (GENERO_SEXUAL) e Óbito (OBITO):
  - Em tudo igual aos atributos com os mesmos nomes no registo de utentes.
- Código de Receita (COD_RECEITA):
  - 318.197 registos unívocos (n.º de prescrições);
  - 363.208 registos unívocos, agrupados por utente;
  - 45.011 registos estão em branco;
  - Série de valores [7573210-7994075];
  - Foram passadas prescrições a 41.984 utentes;
  - 11.695 utentes consultados não têm prescrições;
  - Média de prescrições: 5,93/utente existente e 7,58/utente consultado;
  - 2.250 utentes tiveram mais que uma prescrição por mês, sendo que o máximo de prescrições foi de 104 a um único utente;
  - 9.291 prescrições têm código de receita, mas o código de medicamento está em branco, têm apenas embalagem “999999999-OUTRO MEDICAMENTO”.
- Medicamento (COD_MEDICAMENTO e DSC_MEDICAMENTO):
  - Os atributos correspondem ao código e à descrição comercial proprietária do medicamento pelo que serão aqui tratados em conjunto;
  - 5.798 registos unívocos (n.º de medicamentos);
  - 331.246 registos unívocos, agrupados por utente;
  - 45.691 utentes têm o código de medicamento em branco;
  - Série de valores [1-55.787];
  - O medicamento “52817” não tem descrição, apenas código;
  - O medicamento “38279-Ben-U-Ron” foi o mais prescrito, a 2.939 utentes.
- DCI (COD_MEDICAMENTO_DCI e DSC_MEDICAMENTO_DCI):
  - Os atributos correspondem ao código e à descrição da DCI do medicamento pelo que serão aqui tratados em conjunto;
  - 957 registos unívocos (n.º de princípios ativos);
  - 301.859 registos unívocos, agrupados por utente;

- 45.691 utentes têm o código de DCI em branco, o que respeita a ligação entre a designação proprietária do medicamento e a DCI;
- Série de valores [0-3420];
- 2.439 utentes e 68 medicamentos têm registos com a DCI a "0", sendo que os mais prescritos são medidores e fitas de teste de controlo de glicose;
- O DCI "95-Paracetamol" foi o mais receitado, a 7.521 utentes.

- Tipo de embalagem (COD_EMBALAGEM e DSC_EMBALAGEM):
  - Os atributos correspondem ao código e à descrição da embalagem do medicamento pelo que serão aqui tratados em conjunto;
  - 7.413 registos unívocos (n.º de embalagens diferentes);
  - 337.302 registos unívocos, agrupados por utente;
  - 45.011 utentes contêm registos de embalagem em branco;
  - Série de valores [3- 999999999];
  - 5.040 utentes têm registos com a embalagem "999999999-OUTRO MEDICAMENTO"; excluindo esta, a segunda embalagem mais registada foi a "86556-Blister - 18 unidade(s)" com 2.929 registos.
- Quantidade de embalagens (VAL_QTD):
  - 106.774 registos unívocos, agrupados por utente;
  - 45.011 utentes têm a quantidade prescrita em branco, o que respeita a ligação entre o tipo e a quantidade de embalagem;
  - Série de valores [0-4];
  - 23 registos têm a quantidade errada de "0" pois dizem respeito a prescrições de "164853-UNIDOSE" ou "16201-Blister - 1 unidade(s)"; este valor deveria ser "1";
  - A quantidade deve ser vista relativamente à embalagem; a quantidade 1 de uma embalagem de 20 blisters é maior que a quantidade 4 de uma embalagem de 1 blister;
  - A quantidade de embalagem mais prescrita está distribuída da seguinte forma: "1" a 40.847 utentes, "2" a 20.790 utentes, "3" a 59 utentes e "4" a 47 utentes.

No registo de Diabetes do CS Arnaldo Sampaio constam 1.242.225 registos, cujos atributos são caraterizados conforme se indica de seguida:

- Código de Utente (COD_D_UTENTE), Episódio/consulta (NUM_EPISODIO), Data de Consulta (DATA_CONSULTA), Número de inscrição no CS (NUMERO_INSCR_CENTRO_SAUDE), Código de CS (CODIFICACAO_CENTRO_SAUDE), Nome de Utente (NOME_UTENTE), Género Sexual (GENERO_SEXUAL) e Óbito (OBITO):
  - Em tudo igual aos atributos com os mesmos nomes no registo de utentes.
- Tipo de Diabetes (TIPO_DIABETES):
  - 53.982 registos unívocos, agrupados por utentes;
  - A diferença em relação ao n.º de utentes deve-se a que 303 utentes contenham dois ou mais tipos de diabetes, e.g. os utentes 210100090039523, 210100140086015 e 210100070041765 têm 3 tipos diferentes;

- 53.662 utentes têm o tipo de diabetes “N/A”;
- Os registos dos utentes diabéticos distribuem-se da seguinte forma:

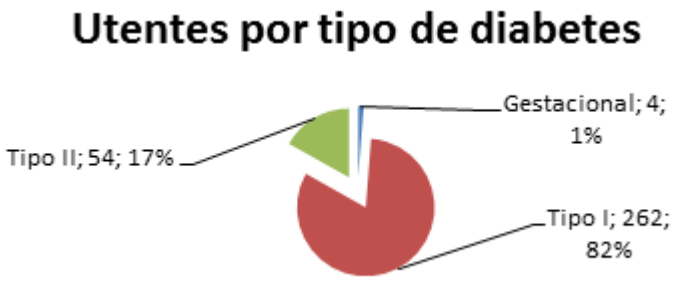

Figura 10 - Distribuição de utentes do CS Arnaldo Sampaio por tipo de diabetes

- Existem registos de 320 pacientes diabéticos, o que em comparação com o n.º de utentes (53.679) corresponde uma taxa de 0,6%. Este valor não corresponde à taxa de prevalência verificada em Portugal que se situa nos 12,84%, segundo a (IDF - International Diabetes Federation, 2012);
- Podemos também confirmar que os dados estão incorretos pela comparação dos utentes com Diabetes Tipo I e Tipo II. Segundo os dados da (Sociedade Portuguesa de Diabetologia, 2013), os pacientes de Tipo II são 9 vezes superiores aos de Tipo I o que contraria os dados da ARSC que afirmam o contrário.

- Data de diagnóstico da doença (DTA_INICIO_PROB):
  - 149.685 registos unívocos, agrupados por utente;
  - Série de valores [3/Abr/1950 a 24/Jan/2043];
  - Os utentes 210100070042799, 210100090019832 e 210100070041765 têm datas de início de diagnóstico com anos superiores ao atual;
  - A validação deste campo pode ser feita analisando se a data não é superior à idade do paciente;
  - A diferença em relação ao n.º de utentes deve-se a que muitos utentes tenham várias datas de início registadas, embora seja possível usar a data mais antiga;
  - 30.430 utentes apresentam uma data de diagnóstico que contrasta com o atributo de tipo de diabetes definido como “N/A”.
- Data de término da doença (DTA_FIM_PROB):
  - 56.531 registos unívocos, agrupados por utente;
  - Série de valores [20/Abr/1952 a 14/Mai/2020];
  - 53.609 registos estão em branco, o que neste caso não significa em erro de entrada de dados, mas sim que não existe término da doença;
  - Existem datas com anos entrados com valores de 2 dígitos, mas aceites como 4 dígitos no intervalo [0007 a 1012];
  - Os utentes 210100140070615 e 210100910087488 têm datas no futuro;

  - A diferença em relação ao n.º de utentes deve-se a que muitos utentes tenham várias datas de término registadas, embora seja possível usar a data mais recente;
  - 2.963 utentes apresentam uma data de término que contrasta com o atributo de tipo de diabetes definido como "N/A".
- N.º de anos desde o diagnóstico (N_ANOS_DIAGNOSTICADO):
  - Este campo é um campo calculado pela diferença entre o atributo Data de diagnóstico da doença (DTA_INICIO_PROB) e a data de extração dos dados, expressa em anos. A correção dos dados é, portanto, dependente do atributo.

Tendo em conta a análise de dados realizada, é possível idealizar algumas das hipóteses de tarefas a concretizar. Através dos dados dos Utentes é possível elaborar modelos de classificação de utentes diabéticos (ou não) e hipertensos (ou não). Através de modelos de segmentação induzidos com base nos mesmos dados é possível descrever os utentes. Nos dados sobre os MCDT e prescrições é possível elaborar modelos de associação que determine aqueles que são realizados em conjunto com mais regularidade.

5.3. Preparação dos dados

Esta etapa, também chamada de pré-processamento, deriva da necessidade de preparar os dados para os algoritmos de indução dos modelos. Uma vez que existem diversos algoritmos, os dados são preparados tendo em conta o formato requerido.

A fase de preparação de dados compreende a seleção de registos como o processamento de dados em falta relevantes para as tarefas em causa. As técnicas de preenchimento automático através de funções matemáticas, como a média dos registos vizinhos, não se aplicam visto existir uma grande variabilidade entre utentes. De forma genérica, visto que os conjuntos de dados são de cardinalidade elevada, a eliminação de registos com dados em falta é uma opção válida, embora elimine registos com valores em outros atributos.

As ações de pré-processamento são consequentes da análise realizada e são idênticas para todos os CS.

No caso dos dados para a tarefa de classificação, nos registos de utentes foram executadas as seguintes ações:

- O atributo código de utente identifica cada utente de forma unívoca. O registo de utentes contempla vários registos por utente; os registos foram agregados para haver apenas um registo para cada utente;
- Foi criado um novo atributo "n.º de consultas" que agrega através da função de contagem os diversos valores de episódios registados;
- Os atributos data de consulta, n.º de inscrição e codificação do CS bem como o nome do utente foram descartados;
- Os registos que não contêm dados de peso, altura, PAD e PAS foram eliminados; esta decisão pode ser limitativa pois muitos pacientes não são medidos por não apresentarem sintomas de algo que não tenha a ver com diabetes e hipertensão.

Contudo, inclui-los obrigaria a por o valor zero ou uma medida padrão o que iria distorcer os dados.

- No CS Arnaldo Sampaio, o atributo género sexual do paciente 210100070024702 contém as duas formas. Embora possa tratar-se de um caso de mudança de género, a manutenção do nome é indicativa da provável ocorrência de um erro. Visto que estes casos são raros, cada caso foi analisado e convertido consoante o nome do utente.
- Os utentes com registos nos dois estados do atributo Óbito foram alterados para "1". Os dados com o atributo em branco foram modificados para “0”.
- Para o atributo Ocupação foi mantida a última situação profissional encontrada.
- O atributo idade foi calculado com base na média dos valores registados nas consultas, arredondado para números inteiros.
- O atributo IMC foi calculado com base na média dos valores registados nas consultas, arredondado para números inteiros. De acordo com os valores de referência da OMS e especificados na análise de dados, foram eliminados valores abaixo de 8 e acima de 23 para idades até aos 5 anos, abaixo de 9 e acima de 35 para idades entre os 6 e 19 anos e abaixo de 14 e acima de 50 para idades acima de 20 anos. Os valores de referência abrangem 96% dos casos; por isso foi adicionada uma margem de 3 pontos em cada extremo para incluir os casos raros de extrema magreza e extrema obesidade. Ainda assim, esta margem não é suficiente para os casos de adultos superobesos e por isso a margem foi aumentada até aos 50; embora sejam casos extremamente raros eles existem nos dados e estão diretamente ligados a casos diabéticos.
- O IMC é calculado em função do peso e altura, pelo que torna estes dois atributos redundantes e por isso foram descartados;
- No CS Arnaldo Sampaio, cerca de 87% dos utentes têm o atributo Perímetro Abdominal em branco. Este facto deve-se a que sejam recolhidas medidas somente quando se trate de determinadas situações clínicas. Se os registos com este atributo em branco fossem mantidos a proporção entre pacientes diabéticos e não diabéticos inverter-se-ia e deixaria de representar a taxa de prevalência da doença. Por este motivo, o atributo foi descartado.
- Os atributos Nefropatia, Retinopatia, Neuropatia, Ano de AVC, Ano de Enfarte do Miocárdio, Ano de Cegueira, Ano de Insuficiência Renal e Ano de Amputação Abaixo e Acima do Tornozelo foram descartados por não conterem registos com dados suficientes. Embora tenha sido considerada a possibilidade de uso destes atributos considerando que os valores em branco fossem indicação da não presença desse sintoma, a proporção entre os registos com e sem a presença dos sintomas não seria real e enviesaria os dados.
- Os atributos PAS e PAD foram tratados de igual forma. Os registos com valores em branco foram eliminados assim como os registos que apresentam valores de PAD inferiores a 20 e de PAS inferiores a 45. Após as eliminações, ambos os atributos foram calculados com base na média dos valores registados nas consultas, arredondado para números inteiros.

- A classe Diagnóstico foi criada através da existência de registos que demonstrem uma associação entre o utente e o programa de saúde correspondente. Os utentes com registos associados ao programa de saúde Diabetes foram sujeitos a uma comparação com o atributo Gravidez na deteção da classe "Diabetes Gestacional". Caso o utente tenha associação com o programa de saúde Diabetes e Hipertensão foi classificado como "Diabetes com Hipertensão". Caso o utente tenha associação ao programa de saúde Diabetes, mas sem o atributo Gravidez e sem associação ao programa de saúde Hipertensão foram classificados como "Diabetes". Os utentes associados com o programa de saúde Hipertensão, mas sem associação a Diabetes foram classificados como "Hipertensão". Todos os demais casos foram classificados como "Normal".

No caso dos dados para as tarefas de associação, nos registos de MCDT e Prescrições foram executadas as seguintes ações:

- Durante a extração dos dados alguns dos registos foram duplicados. Por isso, os registos foram agrupados por todos os atributos para eliminar redundâncias;
- Visto estarmos interessados em associações referentes às patologias de Diabetes Mellitus e Hipertensão, os registos foram classificados por diagnóstico utilizando o ficheiro criado para classificação e foram excluídos os registos de utentes com Diagnóstico "Normal";
- Foram excluídos os registos em que todos os campos específicos das prescrições, referentes ao medicamento, substância ativa e embalagem, se encontram em branco. No caso dos MCDT, foram excluídos os registos em que todos os campos específicos referentes ao exame, detalhes e valores resultantes, se encontram em branco. Por erro na extração dos dados das prescrições do CS do Fundão, as descrições contêm uma mistura de diacríticos em vários códigos. Tanto quanto possível, as descrições foram corrigidas. Se tal não fosse realizado, poderia haver associações que não seriam extraídas por serem considerados vários produtos e não um só.
- Foram excluídos os registos com a embalagem "999999999-OUTRO MEDICAMENTO" referentes a prescrições não especificadas e os registos com MCDT "3681-OUTRO MCDT". Devido à sua generalidade, não constitui valor qualquer associação que pudesse ser encontrada;
- Os registos de MCDT foram agrupados pelo atributo "DSC_MCDT_RESULTADO" e foi renomeado para "MCDT"; este indica o teste realizado. No caso das prescrições foi criado o atributo "Prescrição" que indica a substância ativa prescrita ou a descrição comercial do produto, caso se trate de prescrições de equipamentos como medidores de glicose ou consumíveis como as tiras de teste;
- Os registos foram convertidos numa tabela em que as linhas listam os utentes, as colunas os MCDT ou as prescrições, conforme o caso, e o seu cruzamento a indicação da existência desta relação a este utente, indicado pelo termo "yes", ou em branco indicando a inexistência. Este formato é o requerido pelo algoritmo de associação.
- A conversão em tabela agrupou as transações de cada utente; este facto é importante pois a análise da associação deixa de estar em cada registo, ou seja, a análise passará de encontrar as associações dos MCDT e prescrições realizados em cada consulta ou

receita para ser em cada utente. Esta decisão teve como fator preponderante o conjunto de dados ser esparso.

- Uma vez que o interesse está em associações passíveis de utilização em outros casos, os registos com poucas prescrições foram suprimidos; a tabela resultante contém, no máximo, as 100 mais frequentes. Poderão existir mais que 100 atributos em cada conjunto de dados, pois optou-se pela manutenção de todos os atributos caso estes correspondam em valor ao centésimo valor mais frequente;
- As descrições de MCDT muito extensas foram reduzidas acrescentado um sinal de reticências no lugar do texto suprimido, por exemplo o teste "HEMOGRAMA COM FÓRMULA LEUCOCITÁRIA (ERITROGRAMA, CONTAGEM DE LEUCÓCITOS, CONTAGEM DE PLAQUETAS, FÓRMULA LEUCOCITÁRIA E MORFOLOGIA), S" foi reduzido para "HEMOGRAMA COM FÓRMULA LEUCOCITÁRIA...".
- Foi gerado um conjunto de dados para cada tipo de diagnóstico: Diabetes, Diabetes com Hipertensão, Diabetes Gestacional e Hipertensão.
- Devido à inexistência de dados, não foi possível criar um conjunto de dados para análise de associação dos MCDT no CS Arnaldo Sampaio e no CS do Fundão. Pela mesma razão não foi possível criar um conjunto de dados para a "Diabetes Gestacional" no CS de Tábua.
- Os conjuntos de dados criados para a "Diabetes Gestacional" revelaram a existência de poucos registos: 5 para prescrições no CS Arnaldo Sampaio, 3 para MCDT e prescrições no CS de Eiras e 1 para prescrições no CS do Fundão. Por esse motivo optou-se por descartar a análise de associação a este tipo de diagnóstico.

Os dados de todos os conjuntos de dados foram anonimizados e convertidos para ficheiros com o formato ".arff". O código de utente foi convertido para um valor numérico inteiro sequencial. O mesmo utente tem agora um código sem relação direta com o valor original e diferente em cada ficheiro.

5.4. Modelação

Esta etapa da metodologia consiste na execução dos algoritmos aplicáveis à tarefa de Data Mining em causa, tendo por base os conjuntos de dados preparados na etapa anterior. Os passos a executar diferem entre a indução de modelos de classificação e modelos de associação. Por esse motivo, descrevem-se de seguida em separado.

5.4.1. Classificação

Tendo por base as conclusões retiradas no estudo do capítulo anterior, a indução do modelo final passará por diversas fases. Será criado um primeiro modelo tendo por base o conjunto original de dados. De seguida far-se-á a indução de outro modelo tendo por base o conjunto de dados sujeito a uma prévia seleção de atributos. Numa fase posterior, o conjunto de dados será discretizado com vista a obter um modelo mais compacto sem comprometer o resultado. O modelo final será aquele que apresentar melhores resultados.

O estudo do capítulo anterior demonstrou que não existe uma combinação de ferramenta, tipo de particionamento e algoritmo que produza sempre o melhor resultado, embora alguns se tenham destacado mais vezes que outros. Por esse motivo, os conjuntos de dados serão sujeitos aos melhores do teste, i.e., processamento através da ferramenta Weka, particionamento do tipo *cross-validation* com *5-folds* e os algoritmos da técnica de árvores de decisão. O Quadro 5:12 apresenta os resultados obtidos com os conjuntos de dados originais:

Quadro 5:12 - Resultados dos modelos de classificação obtidos a partir dos conjuntos de dados originais

CS	**Algoritmo**	**Acurác.**	**VP Rácio**	**FP Rácio**	**Precisão**	**Sensib.**	**Medida-F**	**AUC**
Arnaldo Sampaio	BFTree	Err	Err	Err	Err	Err	Err	Err
	DecisionStump	71,98	0,72	0,177	0,684	0,72	0,69	0,809
	FT	73,33	0,733	0,242	0,702	0,733	0,715	0,798
	J48	72,19	0,722	0,23	0,701	0,722	0,71	0,777
	J48graft	72,33	0,723	0,231	0,701	0,723	0,711	0,776
	LADTree	74,64	0,746	0,217	0,673	0,746	0,704	0,897
	LMT	75,26	0,753	0,238	0,695	0,753	0,709	0,907
	NBTree	73,84	0,738	0,178	0,736	0,738	0,725	0,898
	RandomTree	67,71	0,677	0,225	0,682	0,677	0,679	0,727
	REPTree	73,63	0,736	0,238	0,702	0,736	0,712	0,876
	SimpleCart	74,34	0,743	0,241	0,663	0,743	0,699	0,866
Eiras	BFTree	73,79	0,738	0,137	0,723	0,738	0,726	0,849
	DecisionStump	67,99	0,68	0,156	0,626	0,68	0,634	0,815
	FT	74,25	0,742	0,14	0,728	0,742	0,734	0,844
	J48	72,13	0,721	0,146	0,709	0,721	0,715	0,818
	J48graft	72,51	0,725	0,145	0,712	0,725	0,718	0,819
	LADTree	74,13	0,741	0,14	0,721	0,741	0,728	0,908
	LMT	76,06	0,761	0,129	0,746	0,761	0,749	0,921
	NBTree	73,49	0,735	0,128	0,728	0,735	0,729	0,909
	RandomTree	68,33	0,683	0,155	0,684	0,683	0,684	0,764
	REPTree	73,64	0,736	0,148	0,712	0,736	0,723	0,892
	SimpleCart	75,04	0,75	0,139	0,729	0,75	0,738	0,892
Fundão	BFTree	74,16	0,742	0,174	0,719	0,742	0,723	0,845
	DecisionStump	70,68	0,707	0,127	0,702	0,707	0,685	0,822
	FT	73,94	0,739	0,215	0,716	0,739	0,727	0,831
	J48	72,61	0,726	0,221	0,704	0,726	0,714	0,782
	J48graft	72,8	0,728	0,22	0,705	0,728	0,716	0,783
	LADTree	74,41	0,744	0,237	0,709	0,744	0,719	0,908
	LMT	76,25	0,763	0,216	0,727	0,763	0,737	0,922
	NBTree	74,7	0,747	0,163	0,744	0,747	0,74	0,906
	RandomTree	68,3	0,683	0,215	0,688	0,683	0,685	0,736
	REPTree	74,73	0,747	0,231	0,719	0,747	0,726	0,864
	SimpleCart	74,26	0,743	0,252	0,703	0,743	0,715	0,861

Tábua	BFTree	71,63	0,716	0,163	0,709	0,716	0,702	0,833
	DecisionStump	66,01	0,66	0,164	0,648	0,66	0,626	0,785
	FT	73,21	0,732	0,156	0,72	0,732	0,725	0,833
	J48	70,75	0,707	0,174	0,695	0,707	0,701	0,8
	J48graft	70,85	0,709	0,175	0,696	0,709	0,701	0,8
	LADTree	72,45	0,725	0,168	0,707	0,725	0,715	0,893
	LMT	74,73	0,747	0,162	0,726	0,747	0,734	0,906
	NBTree	72,39	0,724	0,131	0,74	0,724	0,726	0,899
	RandomTree	67,81	0,678	0,179	0,682	0,678	0,68	0,75
	REPTree	71,81	0,718	0,17	0,704	0,718	0,708	0,867
	SimpleCart	73,4	0,734	0,171	0,723	0,734	0,722	0,867

O algoritmo *LMT* (*Logistics Model Tree*) obteve o melhor resultado em todos os conjuntos de dados: 75,26% no CS Arnaldo Sampaio, 76,06% no CS de Eiras, 76,25% no CS de Fundão e 74,73% no CS de Tábua, alcançando uma acurácia média de 75,58%. Este resultado indica uma falha de classificação em cada quatro casos.

A inclusão de atributos que não contribuem para melhorar a classificação irá contribuir para piorar o tempo de indução do modelo, ter como resultado um modelo *overfitted* sensível a dados com ruído e acrescenta complexidade de interpretação do modelo induzido. Por este motivo importa determinar se o resultado do modelo melhorará ao serem considerados somente os atributos determinantes do conjunto de dados.

A Weka possui diversos algoritmos de seleção de atributos que se distribuem em duas classes: *Subset* e *Attribute*. Os algoritmos da classe *Subset* avaliam os atributos através da análise de redundância e correlação entre eles e da contribuição coletiva para a classificação final, resultando num subconjunto de atributos mais significativos. Os algoritmos da classe *Attribute* avaliam os atributos através de medidas de correlação e dependência mútua e resultam numa lista de atributos ordenada por importância para a classificação final.

Testar todos os algoritmos de seleção de atributos com todos os conjuntos de dados e com todas as possibilidades de parametrização resulta numa combinação com um número elevado de possibilidades, pelo que importa simplificar. Os conjuntos de dados são estruturalmente idênticos e existe uma semelhança muito forte nos dados, o que torna possível testar os algoritmos de seleção de atributos somente num conjunto de dados e aplicar as conclusões a todos os outros. Não foi realizada qualquer tentativa de melhorar o resultado pela alteração dos parâmetros pré-definidos.

A Weka possui o algoritmo *AttributeSelectedClassifier* que permite testar um algoritmo de classificação sujeitando o conjunto de dados a um prévio algoritmo de seleção de atributos. Usando o algoritmo *LMT*, que melhores resultados obteve no teste anterior, foram testados todos os algoritmos de seleção de atributos, sem alteração dos parâmetros pré-definidos, com o conjunto de dados de classificação do CS Arnaldo Sampaio. O Quadro 5:13 apresenta os resultados obtidos para a classe Subset e os algoritmos com erro ou que resultaram sem atributos foram descartados:

Quadro 5:13 - Resultados dos algoritmos de seleção de atributos da classe *Subset*

Evaluator	**Search**	**Acurácia**	**Atributos considerados relevantes**
Cfs Subset Eval	BestFirst	74,8524	Idade, PAS, Ocupação
	ExhaustiveSearch	74,8524	Idade, PAS, Ocupação
	GeneticSearch	74,8524	Idade, PAS, Ocupação
	GreedyStepwise	74,8524	Idade, PAS, Ocupação
	LinearForwardSelection	74,8524	Idade, PAS, Ocupação
	RandomSearch	75,1214	Ocupação, Idade, IMC, PAS, PAD
	RankSearch	74,8524	Ocupação, Idade, PAS
	ScatterSearchV1	74,8674	Ocupação, Idade, IMC, PAS,
	SubsetSizeForwardSelection	74,8524	Ocupação, Idade, PAS
Classifier Subset Eval	GeneticSearch	71,1916	PAS
	RandomSearch	73,8887	Consultas, GéneroSexual, Óbito, Idade
	RankSearch	73,6795	Idade
Consistency Subset Eval	BestFirst	75,226	Utente, GéneroSexual, Ocupação, Idade, IMC, PAS, PAD
	ExhaustiveSearch	75,226	Utente, GéneroSexual, Ocupação, Idade, IMC, PAS, PAD
	GeneticSearch	75,226	Utente, GéneroSexual, Ocupação, Idade, IMC, PAS, PAD
	GreedyStepwise	75,226	Utente, GéneroSexual, Ocupação, Idade, IMC, PAS, PAD
	LinearForwardSelection	75,226	Utente, GéneroSexual, Ocupação, Idade, IMC, PAS, PAD
	RandomSearch	75,2559	Utente, Consultas, GéneroSexual, Óbito, Ocupação, Idade, IMC, PAS, PAD
	RankSearch	75,226	Utente, GéneroSexual, Ocupação, Idade, IMC, PAS, PAD
	ScatterSearchV1	74,8076	Utente, Ocupação, Idade, PAS
	SubsetSizeForwardSelection	75,226	Utente, GéneroSexual, Ocupação, Idade, IMC, PAS, PAD
Filtered Subset Eval	BestFirst	74,7927	Ocupação, Idade, PAS
	ExhaustiveSearch	74,7927	Ocupação, Idade, PAS
	GeneticSearch	74,7927	Ocupação, Idade, PAS
	GreedyStepwise	74,7927	Ocupação, Idade, PAS
	LinearForwardSelection	74,7927	Ocupação, Idade, PAS
	RandomSearch	75,084	Ocupação, Idade, IMC, PAS, PAD
	RankSearch	74,7927	Ocupação, Idade, PAS
	ScatterSearchV1	74,7703	Ocupação, Idade, IMC, PAS
	SubsetSizeForwardSelection	74,7927	Ocupação, Idade, PAS
Wrapper Subset Eval	GeneticSearch	71,1916	PAS
	RandomSearch	73,8887	Consultas, GéneroSexual, Óbito, Idade
	RankSearch	73,6795	Idade

Conforme se pode observar nos resultados apresentados, vários algoritmos apontam resultados idênticos, ou seja, determinam os mesmos atributos como os determinantes do conjunto de dados. O melhor algoritmo do teste foi o *ConsistencysubsetEval* com *Random Search*. Note-se que o algoritmo determinou que todos os atributos são relevantes e sem surpresa o resultado obtido é igual ao obtido com o algoritmo *LMT* sem remoção de atributos.

Outra possibilidade de seleção de atributos é ordenar os mesmos por relevância para a classificação. Esta é a função dos algoritmos de seleção de atributos da classe *Attribute*. Cada algoritmo computa de acordo com um critério de avaliação de atributos. De acordo com os mesmos, a ordem de importância dos atributos é a seguinte:

Quadro 5:14 - Determinação da importância dos atributos

		Critério de avaliação de atributos					
		Chi-Squared	**Gain Ratio**	**Info Gain**	**One-R**	**Relief-F**	**Symmetrical Uncertainty**
Atributos	Idade	6813,376	0,1252	0,4227	73,4778	0,10811	0,1779
	PAS	4831,353	0,1013	0,2922	71,0347	0,03693	0,1372
	Ocupação	3577,442	0,1213	0,1975	71,2215	0,0358	0,1315
	IMC	2735,829	0,0599	0,1742	66,5820	0,04715	0,0813
	PAD	2098,344	0,0439	0,1287	66,9929	0,0331	0,0598
	Utente	1887,799	0,0239	0,0970	64,6918	0,04011	0,0356
	Género Sexual	363,339	0,0208	0,0193	66,8211	0,00256	0,0167
	Consultas	0	0	0	66,8211	0,01881	0
	Óbito	0	0	0	66,8211	0,00241	0

Com base nestes resultados, os conjuntos de dados foram sujeitos a nova indução do modelo de classificação. Através de diferentes iterações, foi executado o algoritmo *LMT* com todos os atributos e em seguida removido aquele com menor contribuição dos restantes.

Quadro 5:15 - Resultados dos algoritmos de seleção de atributos da classe *Attribute*

Atributos usados	**Acurácia**
Idade, PAS, Ocupação, IMC, PAD, Utente, Género Sexual, Consultas, Óbito	75,2559
Idade, PAS, Ocupação, IMC, PAD, Utente, Género Sexual, Consultas	75,1886
Idade, PAS, Ocupação, IMC, PAD, Utente, Género Sexual	75,226
Idade, PAS, Ocupação, IMC, PAD, Utente	75,0616
Idade, PAS, Ocupação, IMC, PAD	75,084
Idade, PAS, Ocupação, IMC	75,1065
Idade, PAS, Ocupação	74,7927
Idade, PAS	74,7479
Idade	73,6795

Os resultados mostram que a remoção de atributos não contribuiu para a descoberta de um modelo com resultados mais positivos que o induzido através do algoritmo *LMT* e validou os resultados anteriores que demonstravam a obtenção do melhor resultado com o conjunto de atributos total dos dados.

O conjunto de dados contém diversos atributos que são variáveis contínuas. Esta caraterística pode contribuir para modelos complexos de alta dimensão que se revelam ineficientes. O objetivo da discretização é reduzir os valores para um dado número de intervalos. Duas questões que se levantam nesta técnica são a determinação do número de intervalos e dos limites dos intervalos pois podem ter a mesma ou diferentes proporções. Existem dois algoritmos da Weka para discretização, um supervisionado que usa a informação da classe e outro não supervisionado que não usa.

Quadro 5:16 - Resultados dos algoritmos de discretização

Algoritmo	**Acurácia**	**VP Rácio**	**FP Rácio**	**Precisão**	**Sensib.**	**Medida-F**	**AUC**
Supervisionado	75,7041	0,757	0,225	0,735	0,757	0,727	0,911
Não supervisionado	75,1363	0,751	0,236	0,72	0,751	0,718	0,905

Os resultados mostram que o algoritmo supervisionado desta técnica melhorou em 44 décimas o modelo induzido.

Quadro 5:17 - Resultados da discretização do conjunto de dados do CS de Arnaldo Sampaio pelo algoritmo supervisionado

Atributo	**Intervalo**	**Contagem**
Utente	(-inf.-138.5]	138
	(138.5-564.5]	426
	(564.5-912.5]	348
	(912.5-1467.5]	555
	(1467.5-1597.5]	130
	(1597.5-1656.5]	59
	(1656.5-1711.5]	55
	(1711.5-1860.5]	149
	(1860.5-2983.5]	1123
	(2983.5-3967.5]	984
	(3967.5-4268.5]	301
	(4268.5-4460.5]	192
	(4460.5-5763.5]	1303
	(5763.5-7340.5]	1577
	(7340.5-8568.5]	1228
	(8568.5-8970.5]	402
	(8970.5-9446.5]	476
	(9446.5-10414.5]	968
	(10414.5-10673.5]	259
	(10673.5-11661.5]	988
	(11661.5-11879.5]	218
	(11879.5-12148.5]	269
	(12148.5-sup.)	1237
Consultas	Todos	13385
Género Sexual	Feminino	8808
	Masculino	4577
Óbito	Todos	13385
Ocupação	Não Activo	3075

	Activo	7180
	Estudante	560
	Reformado	2554
	Não Aplicável	15
	Desconhecido	1
Idade	(-inf.-16.5]	2043
	(16.5-27.5]	1261
	(27.5-40.5]	2621
	(40.5-47.5]	1496
	(47.5-52.5]	1079
	(52.5-57.5]	1043
	(57.5-65.5]	1455
	(65.5-70.5]	822
	(70.5-sup.)	1565
IMC	(-inf.-17.5]	1193
	(17.5-19.5]	737
	(19.5-21.5]	1116
	(21.5-23.5]	1658
	(23.5-25.5]	1939
	(25.5-27.5]	1941
	(27.5-30.5]	2387
	(30.5-sup.)	2414
PAS	(-inf.-103.5]	1575
	(103.5-117.5]	2425
	(117.5-120.5]	1073
	(120.5-126.5]	1490
	(126.5-132.5]	1898
	(132.5-140.5]	2112
	(140.5-sup.)	2812
PAD	(-inf.-53.5]	675
	(53.5-60.5]	962
	(60.5-70.5]	2570
	(70.5-75.5]	1831
	(75.5-81.5]	3026
	(81.5-90.5]	3120
	(90.5-sup.)	1201
Diagnóstico	Normal	8944
	Hipertensão	2627
	Diabetes	733
	Diabetes com Hipertensão	1076
	Diabetes Gestacional	5

Embora as árvores de decisão tenham sido a técnica que melhores resultados obteve no teste, existiram outras técnicas e algoritmos que se destacaram em situações específicas. Por esse motivo, e sempre com o objetivo de obter o melhor modelo classificador, o conjunto de dados do CS Arnaldo Sampaio foi sujeito aos algoritmos disponíveis em outras técnicas para comparação.

Quadro 5:18 - Resultados de classificação do conjunto de dados do CS de Arnaldo Sampaio com outras classes de algoritmos de classificação que não árvores de decisão

Classe	Algoritmo	Acurácia	VP Rácio	FP Rácio	Precisão	Sensi-bilidade	Medida-F	AUC
Regras	ConjunctiveRule	71,16	0,712	0,167	0,687	0,712	0,685	0,817
	DecisionTable	74,49	0,745	0,248	0,68	0,745	0,7	0,894
	DTNB	74,67	0,747	0,199	0,736	0,747	0,71	0,896
	JRip	72,41	0,724	0,427	0,679	0,724	0,665	0,67
	OneR	73,6	0,736	0,293	0,642	0,736	0,686	0,721
	PART	70,86	0,709	0,225	0,694	0,709	0,701	0,788
	Ridor	72,1	0,721	0,267	0,689	0,721	0,701	0,727
	ZeroR	66,82	0,668	0,668	0,447	0,668	0,535	0,5
Bayes	BayesNet	72,93	0,729	0,143	0,752	0,729	0,723	0,899
	NaiveBayes	73,07	0,731	0,145	0,739	0,731	0,714	0,898
	NaiveBayesUpdateable	73,07	0,731	0,145	0,739	0,731	0,714	0,898
SVM	SMO	75,33	0,753	0,233	0,67	0,753	0,707	0,785
Regres-são	Logistic	75,2	0,752	0,231	0,705	0,752	0,712	0,907
	SimpleLogistic	75,23	0,752	0,239	0,692	0,752	0,708	0,907
ANN	MultilayerPerceptron	74,99	0,75	0,245	0,71	0,75	0,71	0,903
Cluster	RBFNetwork	74,37	0,744	0,209	0,714	0,744	0,706	0,897
	IB1	67,68	0,677	0,233	0,678	0,677	0,677	0,722
	IBk	67,68	0,677	0,233	0,678	0,677	0,677	0,722
	Kstar	71	0,71	0,251	0,689	0,71	0,699	0,874
	LWL	72,04	0,72	0,177	0,684	0,72	0,69	0,874
	HyperPipes	66,99	0,67	0,663	0,572	0,67	0,54	0,589

Observa-se assim a confirmação dos resultados do estudo mostrado no capítulo anterior que revela a técnica das Árvores de Decisão como aquela em que se obteve o melhor resultado. Uma vez que o melhor modelo foi obtido através do algoritmo *LMT* com os dados discretizados pelo algoritmo supervisionado, em seguida apresentam-se os resultados para todos os conjuntos de dados.

Quadro 5:19 - Resultados do algoritmo *LMT* para os conjuntos de dados de todos os CS discretizados pelo algoritmo supervisionado

CS	Algoritmo LMT com dados discretizados pelo algoritmo supervisionado						
	Acurácia	VP Rácio	FP Rácio	Precisão	Sensib.	Medida-F	AUC
Arnaldo Sampaio	75,7041	0,757	0,225	0,735	0,757	0,727	0,911
Eiras	76,1817	0,762	0,130	0,749	0,762	0,749	0,924
Fundão	76,3775	0,764	0,211	0,727	0,764	0,740	0,922
Tábua	74,6457	0,746	0,158	0,737	0,746	0,736	0,910

Os resultados obtidos mostram uma ligeira melhoria em todos os conjuntos de dados, mas suportam a afirmação que serão somente capazes de acertar três em cada quatro casos. Os modelos de classificação obtidos através do algoritmo *LMT* discretizados pelo algoritmo supervisionado podem ser consultados no capítulo 8.1 (Anexos).

Considerando os resultados de acurácia para cada classe, o diagnóstico preciso é difícil embora este nem sempre seja o desejado. Num sistema de prevenção e deteção precoce da doença, o importante é sinalizar a possível existência de uma patologia e não o diagnóstico preciso, que ocorre posteriormente por um médico. Por esse motivo, os dados de todos os CS foram unidos num conjunto de dados e o atributo classe foi reduzido a um valor binário: normal e doença, sinalizando a ausência ou presença de patologia, respetivamente. Neste cenário não se verifica a necessidade de incluir os atributos código de utente e óbito e por isso foram eliminados na indução do modelo. Os dados foram discretizados pelo algoritmo supervisionado e o modelo induzido pelo algoritmo de classificação *LMT* que pode ser consultado no no capítulo 8.1 (Anexos).

Quadro 5:20 - Resultados do algoritmo *LMT* para o conjunto de dados de todos os CS com classe de diagnóstico binária discretizados pelo algoritmo supervisionado

Algoritmo LMT com dados discretizados pelo algoritmo supervisionado						
Acurácia	**VP Rácio**	**FP Rácio**	**Precisão**	**Sensib.**	**Medida-F**	**AUC**
87,0742	0,871	0,14	0,872	0,871	0,871	0,942

Neste último modelo o valor de acurácia subiu de uma média de 76% para 87%, e os valores de acurácia de classificação de cada classe são aproximados.

5.4.2. Associação

O software Weka possui quatro algoritmos de indução de modelos de associação cujo resultado pode ser utilizado pelos ficheiros criados no ponto anterior. O algoritmo *A-priori* pode ser usado quando se deseja dar importância diferenciada entre os parâmetros tradicionais de suporte e confiança. O algoritmo *PredictiveApriori* combina os parâmetros suporte e confiança num parâmetro chamado acurácia para extrair as melhores regras. O algoritmo *Tertius* descobre regras através de representação lógica de primeira ordem e contém vários parâmetros. O algoritmo *FilteredAssociator* permite correr qualquer um dos algoritmos anteriores em dados previamente filtrados.

De acordo com as conclusões do estudo de (Aher & Lobo, 2012) que compara os algoritmos do *software* Weka, o algoritmo *A-priori* foi aquele que obteve o melhor resultado. Contudo, os ficheiros de dados de Associação da ARSC são de baixa cardinalidade e os dados estão esparsos pelo que trabalhar somente com o parâmetro de suporte afigura-se como não ser o ideal. Por estas razões, as regras induzidas para o resultado deste estudo foram geradas através do algoritmo *PredictiveApriori*. As regras induzidas podem ser consultadas no capítulo 8.2 (Anexos).

5.5. Avaliação

O passo seguinte na metodologia refere que os modelos obtidos devem ser sujeitos a avaliação. Mesmo que existam poucos dados para que uma correta avaliação seja realizada, é de valorizar qualquer esforço que obtenha o diagnóstico precoce e melhore o prognóstico para os pacientes de Diabetes Mellitus e Hipertensão.

Relativamente aos modelos de classificação, os resultados são bastante interessantes num sistema de triagem precoce, fundamental em Diabetes Mellitus e Hipertensão, aproximando-se de um erro por cada 10 testes.

As regras de associação obtidas revelam que os MCDT não apresentam variações significativas entre CS embora existam entre as diferentes classes de diagnóstico. As regras encontradas estão associadas aos MCDT mais frequentes, i.e., testes laboratoriais de despistagem de ocorrência anormal através de análise bioquímica. Relativamente às prescrições, as regras sobre Diabetes Mellitus são dominadas pelo princípio ativo Metformina (antidiabético oral). As prescrições a pacientes hipertensos revelam muita frequência de Paracetamol (analgésico), Amoxicilina (antibiótico) e Acetilcisteína (agente mucolítico) que não têm uma relação direta com a patologia.

A correta avaliação dos modelos induzidos requer que seja feita por um especialista do domínio que analise as conclusões verificadas e valide os modelos como passíveis de serem utilizados. Visto que nesta altura não existe esta possibilidade, fica sugerido para trabalho futuro.

5.6. Operacionalização

Após a avaliação dos modelos obtidos, o passo de operacionalização consiste em apresentar resultados e implementar medidas de alteração aos processos médicos de rastreio e a programação informática de sistemas na organização que utilizem o conhecimento obtido. A primeira parte é satisfeita pela redação desta dissertação. Para as restantes, foi desenvolvida uma aplicação em Windows Forms com C#.NET que implementa os classificadores obtidos (Figura 11). O utilizador tem a opção de escolher entre um dos Centros de Saúde estudados e o classificador geral que engloba todos os Centros de Saúde. O maior valor determina a classe.

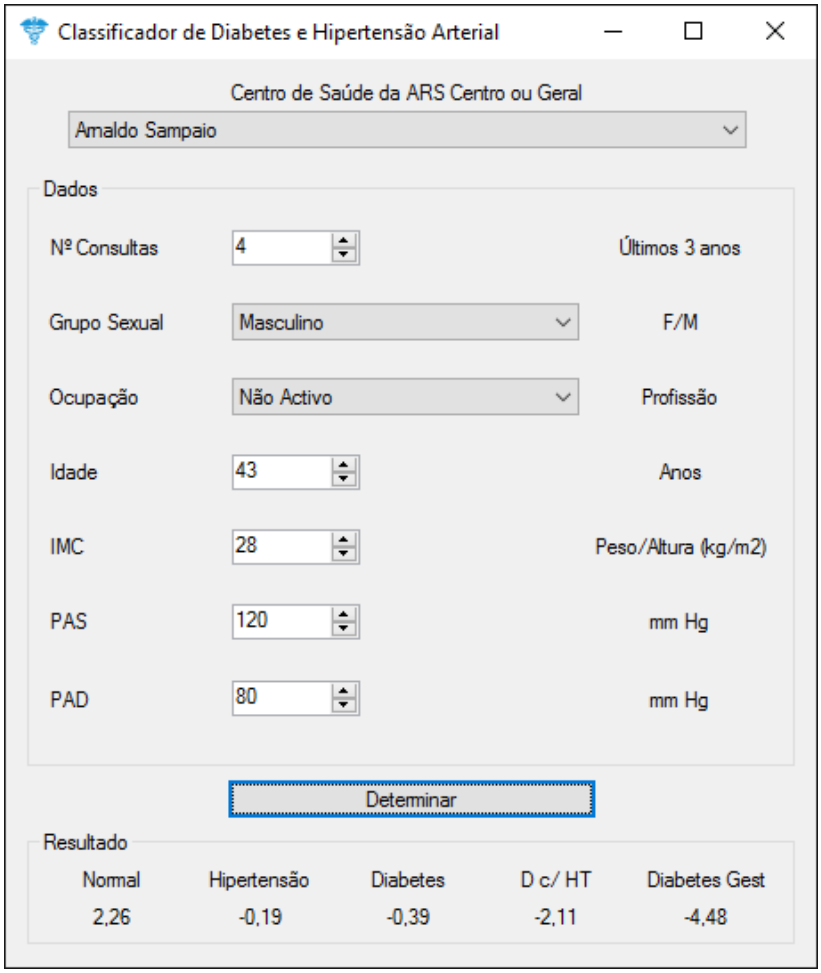

Figura 11 – Classificador de utentes desenvolvido em função do resultado deste estudo

Fica a sugestão para que a sua implementação seja feita pela ARSC para rastreio não invasivo como medida de deteção precoce a baixo custo.

6. CONCLUSÃO

6.1. Conclusões

De uma forma geral, pode dizer-se que o objetivo de estudar os dados sobre as doenças Diabetes Mellitus e Hipertensão, aplicando métodos e técnicas de Tecnologias da Informação e do Conhecimento, integrados na área científica de Business Intelligence e Data Mining, para extrair conhecimento dos dados existentes no Data Warehouse da ARSC, foi alcançado. Com efeito, foi possível descrever e compreender os dados, usando técnicas de Descrição e Visualização bem como Sumarização, e obter um modelo matemático classificador, com uma acurácia de 87%, e modelos de associação de terapêuticas, que revelam prescrições medicamentosas com maior incidência para Metformina, desconhecidos até então.

A realização de um teste determinante da melhor ferramenta, tipo de particionamento e técnica de modelação para classificação concluiu que não existe tal combinação que resulte sempre no melhor resultado. Neste trabalho foi utilizado o *software* Weka e os algoritmos testados foram aqueles que consistentemente obtiveram os melhores resultados dos testes, particionamento *cross-validation* com *5-folds* e árvores de decisão incluindo técnicas de Ensemble.

No âmbito deste estudo estiveram envolvidos 4 Centros de Saúde escolhidos por serem representativos de duas dimensões: A localização geográfica no litoral e interior do país e a densidade da população por ser de maior ou menor dimensão. A conclusão do estudo revela que entre a amostra com a mellhor e a pior classificação se encontra uma diferença de apenas 1,52%. Este facto indica uma população homogénea no que às conclusões de este estudo diz respeito.

Os dados da ARSC confirmam o estado da arte referido pelos estudos mencionados na revisão literária. No domínio da Saúde, os dados continuam a ser poucos e de fraca qualidade, embora aos poucos se comece a entender a sua importância para obter melhores resultados. Através da revisão efetuada aos estudos literários sobre esta matéria foi possível identificar um total de 84 atributos relacionados com estas patologias, distribuídos por: Caraterísticas demográficas e gerais (7), histórico da doença (6), sintomas (11), sinais de alarme (12), medidas antropométricas (6), testes (25), fatores de risco relacionados com o estilo de vida (8) e prescrições (9). Não foi possível extrair dados do Data Warehouse da ARSC para 29 destes atributos. A inexistência de uma quantidade significativa de dados, usados noutros trabalhos, limita de forma crítica a elaboração de conclusões e comparações, tais como posição socio-económica, habilitações literárias ou história familiar são determinantes no processo de diagnótisco sem recurso a testes invasivos bio-químicos.

Os registos extraídos do Data Warehouse para os Centros de Saúde estudados encontram-se no período compreendido entre 1 de janeiro de 2011 e uma data de término que varia entre 21 de janeiro de 2013 e 18 de dezembro de 2013. Na fase de pré-processamento dos dados foi necessário proceder a muitas alterações de limpeza e integridade; dos 3.510.429 registos extraídos do Data Warehouse resultaram diversos conjuntos de dados que no seu total perfazem 28.416 registos de classificação e 15.421 registos de associação. Considerando esta redução para

1,25% do total de registos, a diferença temporal registada entre Centros de Saúde não teve expressão nos resultados finais.

Importa referir que os dados foram também anonimizados devido à sua natureza e com respeito pela necessidade de tornar impossível a sua rastreabilidade ao utente em questão.

A modelação dos dados para obtenção de um classificador revelou que o algoritmo *LMT* obteve o melhor resultado em todos os conjuntos de dados alcançando uma acurácia média de 75,58%. Os dados foram sujeitos a algoritmos de seleção de atributos que validaram que a remoção de atributos contribuiria negativamente para o resultado obtido. Os dados foram também sujeitos a algoritmos de discretização dos atributos que são variáveis contínuas, que melhorou em 0,44% décimas a acurácia do modelo. Os resultados de acurácia finais dos modelos obtidos foram 75,70% para o CS Arnaldo Sampaio, 76,18% para o CS Eiras, 76,38% para o CS Fundão e 74,65% para o CS Tábua. Para que o modelo possa ser utilizado como sinalizador da existência de uma patologia sem precisar qual, os conjuntos de dados foram unidos para classificar apenas em duas classes: Normal e Doença. Com este conjunto de dados foi possível obter uma acurácia de 87,07%.

A modelação dos dados para obtenção de um regras de associação através do algoritmo *PredictiveApriori* revelou que os MCDT não apresentam variações significativas entre CS e que as regras mais frequentes estão associadas aos testes laboratoriais de despistagem de ocorrência anormal através de análise bioquímica. Relativamente às prescrições, as regras são dominadas pelo princípio ativo Metformina (antidiabético oral) associado com Paracetamol (analgésico), Amoxicilina (antibiótico) e Acetilcisteína (agente mucolítico) que não têm uma relação direta com as patologias em estudo.

Com base nos modelos de classificação obtidos foi criada uma aplicação de *software* que permite sinalizar a possível presença ou ausência de patologia. Esta ferramenta pode ser usada como MCDT que acarreta as vantagens de não ser um teste invasivo e ser de baixo custo.

6.2. Propostas para trabalho futuro

Este estudo confirma a tese de que é possível usar técnicas de Data Mining nos dados da ARSC que se podem traduzir em importantes melhorias na saúde dos utentes dos Centros de Saúde. Por esse motivo, recomenda-se que os dados armazenados sejam melhorados em qualidade e quantidade. Com mais e melhores dados recomenda-se a revisão dos modelos obtidos e através de análise pericial.

Tal como ficou demonstrado, a combinação de muitas técnicas, muitos algoritmos, muitas ferramentas de software, muitos parâmetros e muitas formas de tratamento dos dados inviabiliza a determinação do melhor resultado num curto espaço de tempo. Por este motivo, em Data Mining recorre-se muitas vezes a executar o algoritmo que oferece melhores resultados diversas vezes alterando os seus parâmetros. Para obter modelos com resultados mais precisos recomenda-se a reparametrização dos algoritmos.

Relativamente à aplicação de *software* desenvolvida com base nos resultados deste estudo, a mesma pode ser implementada em diversas plataformas, mais próximas dos utentes, como as plataformas móveis para dispositivos como o *Smartphone* e para *Kiosks* instalados nos Serviços de Saúde e na Internet.

7. REFERÊNCIAS BIBLIOGRÁFICAS

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *Proceeding of the 20th VLDB Conference* (pp. pp.487-499). Santiago, Chile: Very Large Data Base Endowment.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining Association Rules between Sets of Items in Large Databases. *Proceedings of the 1993 SCM SIGMOD Conference.* Washington, DC, USA: ACM.

Akinci, F., Coyne, J., Healey, B., & Minear, J. (2004). National Performance Measures for Diabetes Mellitus Care. *Disease Manage & Health Outcomes, 5*, pp. 285-298.

AlJarullah, A. A. (2011). Decision Tree Discovery for the Diagnosis of Type II Diabetes. *International Conference on Innovations in Information Technology* (pp. 303-307). IEEE.

Aljumah, A. A., Ahmad, M. G., & Siddiqui, M. K. (July de 2013). Application of data mining: Diabetes health care in young and old patients. (K. S. University, Ed.) *Computer and Information Sciences, 25*, pp. 127-136.

Apte, C., Liu, B., Pednault, E. P., & Smyth, P. (August de 2002). Business Applications of Data Mining. *Communications of the ACM, 45*.

Ban, H.-J., Heo, J. Y., Oh, K.-S., & Park, K.-J. (2010). Identification of Type 2 Diabetes-associated combination of SNPs using Support Vector Machine. *BMC Genetics, 26*.

Bayu, A. T., Rodiyatul, F. S., & Hermansyah. (August de 2011). An Early Detection Method of Type-2 Diabetes Mellitus in Public Hospital. *Telkomnika, 9*, pp. 287-294.

Borges, L. C., Marques, V. M., & Bernardino, J. (2013). Comparison of Data Mining techniques and tools for data classification. *Proceedings of the International C* Conference on Computer Science and Software Engineering (C3S2E'13)* (pp. 113-116). Porto, Portugal: ACM, New York, NY, USA.

Breault, J. L. (2001). Data Mining Diabetic Databases: Are Rough Sets a Useful Addition? *In Proc. 33rd Symposium on the Interface, Computing Science and Statistics.* Costa Mesa, Orange County, California.

Breiman, L. (Outubro de 2001). Random Forests. (R. E. Schapire, Ed.) *Machine Learning, 45*, pp. 5-32.

Brown, N., Critchley, J., Bogowicz, P., Mayige, M., & Unwin, N. (2012). Risk scores based on self-reported or available clinical data to detect undiagnosed Type 2 Diabetes: A systematic review. *Diabetes Research and Clinical Practice 98*, pp. 369-385.

Castellani, B., & Castellani, J. (Setembro de 2003). Data Mining: Qualitative Analysis With Health Informatics Data. *Qualitative Health Research, 13 (No. 7)*, 1005-1018.

Chae, Y. M., Kim, H. S., Tark, K. C., Park, H. J., & Ho, S. H. (2003). Analysis of healthcare quality indicator using data mining and decision support system. (Pergamon, Ed.) *Expert Systems with Applications, 24*, 167-172.

Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial Intelligence in Medicine, 26*, 1-24.

Clark, P., & Boswell, R. (1991). Rule Induction with CN2: Some Recent Improvements. Em Y. Kodratoff (Ed.), *Proceedings of the Fifth European Conference - Machine Learning (EWSL-91)* (pp. pp. 151-163). Berlin: Springer-Verlag.

Cohen, W. W. (1995). Fast Effective Rule Induction. *Proceedings of the Twelfth International Conference on Machine Learning* (pp. pp. 115-123). Morgan Kaufmann.

Concaro, S., Sacchi, L., Cerra, C., & Bellazzi, R. (2009). Mining Administrative and Clinical Diabetes Data with Temporal Association Rules. Em K.-P. A. (Eds.), *Medical Informatics in a United and Healthy Europe* (pp. 574-578). IOS Press.

Couvreur, C. (1996). The EM algorithm: A Guided Tour. *Proceedings of the 2nd European Workshop on Computationaly Intensive Methods in Control and Signal Processing.* Prague.

Defays, D. (1976). An efficient algorithm for a complete link method. *The Computer Journal, Volume 20 Number 4*, pp.364-366.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Methodological), Volume 39, No. 1*, pp. 1-38.

Dietterich, T. G., & Bakiri, G. (1995). Solving Multiclass Learning Problems via Error-Correcting Output Codes. *Journal of Artificial Intelligence Research, 2*, pp. 263-286.

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of computer and system sciences, 55*, pp. 119-139.

Ganji, M. F., & Abadeh, M. S. (2011). A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis. *Expert Systems with Applications*, 14650-14659.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., . . . Lander, E. S. (15 de Oct de 1999). Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science, 286*, pp. 531-537.

Hamming, R. W. (1950). Error Detecting and Error Correcting Codes. *The Bell System Technical Journal, XXIX*, pp. 147-160.

Han, J., Pei, J., & Yin, Y. (2000). Mining Frequent Patterns without Candidate Generation. *Proceedings of the Conference on the Management of Data (SIGMOD'00).* Dallas, Texas, USA: ACM Press, New York, NY, USA .

Huang, M.-J., Chen, M.-Y., & Lee, S.-C. (2007). Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. *Expert Systems with Applications, 32*, 856-867.

Huang, Y., McCullagh, P., Black, N., & Harper, R. (2007). Feature selection and classification model construction on type 2 diabetic patients' data. *Artificial Intelligence in Medicine*, 251-262.

IDF - International Diabetes Federation. (2012). *IDF Diabetes Atlas, 5th Edition.* Belgium: Internation Diabetes Federation.

Kaplan, E. A. (February de 2008). Using Data Mining to deal with Diabetes. *Employee Benefit News*, pp. 27 - cover story.

Karthikeyani, V., Begum, I. P., Tajudin, K., & Begam, I. S. (December de 2012). Comparative of Data Mining Classification Algorithm (CDMCA) in Diabetes Disease Prediction. *International Journal of Computer Applications, 60*, pp. 0975-8887.

Khajehei, M., & Etemady, F. (2010). Data Mining and Medical Research Studies. *Second International Conference on Computational Intelligence, Modelling and Simulation* (pp. 119-122). IEEE Computer Society.

Kira, K., & Rendell, L. A. (1992). The Feature Selection Problem: Traditional Methods and a New Algorithm. *Tenth National Conference on Artificial Intelligence (AAAI-92)* (pp. 129-134). San Jose Convention Center, San Jose, California: Association for the Advancement of Artificial Intelligence.

Koh, H. C., & Tan, G. (2005). Data Mining Applications in Healthcare. *Journal of Healthcare Information Management, Vol. 19, No. 2*, pp. 64-72.

Lloyd, S. P. (March de 1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, Volume IT-28 (2)*, pp. pp. 129-137.

Marling, C., Wiley, M., Bunescu, R., Shubrook, J., & Schwartz, F. (2012). Emerging Applications for Intelligent Diabetes Management. *AI Magazine, Summer 2012*, pp. 67-78.

Martin, E., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96).* Portland, Oregon, USA: Association for the Advancement of Artificial Intelligence.

McCarthy, J. (August de 2000). Phenomenal Data Mining. *Communication of the ACM, 43*, pp. 75-79.

Meng, X.-H., Huang, Y.-X., Rao, D.-P., & Liu, Q. (2012). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *Kaohsiung Journal of Medical Sciences*, 1-7.

Meyer, D., Leisch, F., & Hornik, K. (2003). The support vector machine under test. *Neurocomputing, 55*, pp. pp.169-186.

National Institute of Diabetes and Digestive and Kidney Diseases. (2013). *Pima Indians Diabetes Data Set*. (I. S. University of California, Editor, V. Sigillito , P. Turney, Produtores, & Applied Physics Laboratory, The Johns Hopkins University) Obtido em 27 de Janeiro de 2014, de UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

Novakovic, J. (23-25 de November de 2010). The Impact of Feature Selection on the Accuracy of Naive Bayes Classifier. *18th Telecommunications forum TELFOR 2010*, (pp. 1113-1116). Serbia, Belgrade.

Nuwangi, S. M., Oruthotaarachchi, C. R., Tilakaratna, J. M., & Caldera, H. A. (2010). Usage of association rules and classification techniques in knowledge extraction of diabetes. *Advanced Information Management and Service (IMS), 2010 6th International Conference on* (pp. 372-377). Seoul: IEEE.

Nuwangi, S. M., Oruthotaarachchi, C. R., Tilakaratna, J. M., & Caldera, H. A. (2010). Utilization of Data Mining Techniques in Knowledge Extraction for Diminution of Diabetes. *Second Vaagdevi International Conference on Information Technology for Real World Problems* (pp. 3-8). Warangal, India: IEEE.

Oracle . (02 de 03 de 2014). Obtido de Oracle: http://www.oracle.com/pt/index.html

Otero, F. E., Freitas, A. A., & Johnson, C. G. (2012). Inducing decision trees with an ant colony optimization algorithm. *Applied Soft Computing, 12*, 3615-3626.

Parthiban, G., Rajesh, A., & Srivatsa, S. K. (June de 2011). Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method. *International Journal of Computer Applications, 24*, pp. 7-11.

Patil, B., Joshi, R. C., & Toshniwal, D. (2010). Hybrid prediction model for Type-2 diabetic patients. *Expert Systems with Applications, 37*, 8102-8108.

Pawlak, Z. (1 de 10 de 1982). Rough Sets. *International Journal of Computer and Information Sciences, 11*, pp. 341-356.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning.* Amsterdam: Morgan Kaufmann.

Santos, F. M., & Azevedo, C. S. (2005). *Data Mining - Descoberta de conhecimento em bases de dados* (Vol. 1). Lisboa, Portugal: FCA - Editora de Informática.

Santos, M. Y., & Ramos, I. (2009). *Business Intelligence - Tecnologias da Informação na Gestão de Conhecimento* (2ª edição ed., Vol. 1). Lisboa, Portugal: FCA - Editora de Informática.

Sapna, S., Tamilarasi, A., & Kumar, M. P. (January de 2012). Implementation of Genetic Algorithm in Predicting Diabetes. *International Journal of Computer Science Issues, 9*.

SAS Enterprise Miner. (2013). (SAS Institute Inc.) Obtido em 21 de 08 de 2013, de http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html

Schank, R. C. (1982). *Dynamic Memory: A Theory of Learning in Computers and People.* New York: Cambridge University Press.

Shearer, C. (Fall de 2000). The CRISP-DM Model: The New Blueprint for Data Mining. (H. J. Watson, Ed.) *Journal of Data Warehousing, vol. 5, n.4*, 13-22.

Sibson, R. (1972). SLINK: An optimally efficient algorithm for the single-link method. *The Computer Journal, Volume 16 Number 1*, pp. 30-34.

Sneath, P. H., & Sokal, R. R. (1973). *Numerical Taxonomy: The principles and Practice of Numerical Classification.* San Francisco, USA: W. H. Freeman and Company.

Sociedade Portuguesa de Diabetologia. (2013). *Diabetes: Factos e Números 2012 - Relatório Anual do Observatório Nacional da Diabetes.* Lisboa: Letra Solúvel – Publicidade e Marketing, Lda.

Sokal, R., & Michener, C. D. (20 de Março de 1958). A Statistical Method for Evaluating Systematic Relationships. *Science Bulletin, Volume XXXVIII Pt. II No. 22*.

Su, C.-T., Yang, C.-H., Hsu, K.-H., & Chiu, W.-K. (2006). Data Mining for the Diagnosis of Type II Diabetes from Three-Dimensional Body Surface Anthropometrical Scanning Data. *Computers and Mathematics with Applications, 51*, 1075-1092.

Suh, S. C., & Vudumula, G. P. (2011). The role of conceptual hierarchies in the diagnosis and prevention of diabetes. Em Y. Cho, S. Kawala, & F. Ko (Ed.), *7th International Conference on Networked Computing and Advanced Information Management (NCM), 2011* (pp. 267-275). Gyeongju, Korea: IEEE.

Temurtas, H., Yumusak, N., & Temurtas, F. (2009). A comparative study on diabetes disease diagnosis using neural networks. *Expert Systems with Applications, 36*, 8610-8615.

Toussi, M., Lamy, J.-B., Toumelin, P., & Venot, A. (2009). Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes. *BMC Medical Informatics and Decision Making, 9*.

Ward, J. H. (March de 1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association, Volume 58, Issue 301*, pp.236-244.

World Health Organization. (04 de 03 de 2014). Obtido de WHO | World Health Organization: http://www.who.int/en/

Wright, A., Chen, E. S., & Maloney, F. L. (2010). An automated technique for identifying associations between medications, laboratory results and problems. *Journal of Biomedical Informatics, 43*, 891-901.

Xuezhong, Z., Shibo, C., Baoyan, L., Runsun, Z., Yinghui, W., Ping, L., . . . Xiufeng, Y. (2010). Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support. *Artificial Intelligence in Medicine, 48*, 139-152.

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.-F., & Hua, L. (2012). Data Mining in Healthcare and Biomedicine: A Survey of the Literature. *J Med Syst, 36*, pp. 2431-2448.

Zadeh, L. A. (Junho de 1965). Fuzzy sets. *Information and Control, 8*, pp. 338-353.

8. ANEXOS

8.1. Modelos de Classificação

8.1.1. Algoritmo LMT+Discretize-Rfirst-last para o CS Arnaldo Sampaio

=== Run information ===

Scheme:weka.classifiers.trees.LMT -I -1 -M 15 -W 0.0

Relation: Class_AS-weka.filters.supervised.attribute.Discretize-Rfirst-last

Instances: 13385

Attributes: 10

Utente

Consultas

Género Sexual

Óbito

Ocupação

Idade

IMC

PAS

PAD

Diagnóstico

Test mode:5-fold cross-validation

=== Classifier model (full training set) ===

Logistic model tree

: LM_1:48/48 (13385)

Number of Leaves : 1

Size of the Tree : 1

LM_1:

Class 0 :

1.78 +

[Utente='(564.5-912.5]'] * 0.21 +

[Utente='(1860.5-2983.5]'] * -0.12 +

[Utente='(2983.5-3967.5]'] * -0.21 +

[Utente='(12148.5-inf)'] * 0.25 +

[Género Sexual] * -0.35 +

[Ocupação=Reformado] * -0.12 +

[Idade='(-inf-16.5]'] * 2.38 +

[Idade='(16.5-27.5]'] * 1.67 +

[Idade='(27.5-40.5]'] * 1.16 +

[Idade='(40.5-47.5]'] * 0.55 +

[Idade='(47.5-52.5]'] * 0.16 +

[Idade='(57.5-65.5]'] * -0.48 +

[Idade='(65.5-70.5]'] * -1.08 +

[Idade='(70.5-inf)'] * -1.97 +

[IMC='(17.5-19.5]'] * 0.36 +

[IMC='(19.5-21.5]'] * 0.56 +

[IMC='(21.5-23.5]'] * 0.27 +

[IMC='(23.5-25.5]'] * 0.2 +

[IMC='(27.5-30.5]'] * -0.06 +

[IMC='(30.5-inf)'] * -0.56 +

[PAS='(103.5-117.5]'] * 0.78 +

[PAS='(126.5-132.5]'] * -0.14 +

[PAS='(140.5-inf)'] * -0.66 +

[PAD='(60.5-70.5]'] * 0.4

Class 1 :

-0.11 +

[Utente='(-inf-138.5]'] * 0.31 +

[Utente='(138.5-564.5]'] * -1.29 +

[Utente='(1467.5-1597.5]'] * -0.5 +

[Utente='(1656.5-1711.5]'] * -1.6 +

[Utente='(1711.5-1860.5]'] * -0.46 +

[Utente='(1860.5-2983.5]'] * 0.11 +

[Utente='(2983.5-3967.5]'] * 0.24 +

[Utente='(4268.5-4460.5]'] * 0.3 +

[Utente='(8568.5-8970.5]'] * -0.7 +

[Utente='(8970.5-9446.5]'] * 0.15 +

[Utente='(10414.5-10673.5]'] * -0.7 +

[Utente='(10673.5-11661.5]'] * 0.36 +

[Utente='(11661.5-11879.5]'] * 0.46 +

[Utente='(11879.5-12148.5]'] * -0.77 +

[Género Sexual] * 0.04 +

[Ocupação=Não Aplicável] * -1.39 +

[Idade='(-inf-16.5]'] * -1.89 +

[Idade='(16.5-27.5]'] * -1.53 +

[Idade='(27.5-40.5]'] * -0.64 +

[Idade='(40.5-47.5]'] * -0.32 +

[Idade='(47.5-52.5]'] * -0.32 +

[IMC='(-inf-17.5]'] * -0.61 +

[IMC='(17.5-19.5]'] * -0.38 +

[IMC='(27.5-30.5]'] * 0.23 +

[PAS='(-inf-103.5]'] * -1.17 +

[PAS='(103.5-117.5]'] * 0.28 +

[PAS='(120.5-126.5]'] * 0.39 +

[PAS='(126.5-132.5]'] * 0.58 +

[PAS='(132.5-140.5]'] * 1.22 +

[PAS='(140.5-inf)'] * 1.3 +

[PAD='(60.5-70.5]'] * -0.09 +

[PAD='(70.5-75.5]'] * -0.07 +

[PAD='(75.5-81.5]'] * -0.07 +

[PAD='(81.5-90.5]'] * 0.14 +

[PAD='(90.5-inf)'] * 0.16

Class 2 :

-1.26 +

[Utente='(912.5-1467.5]'] * 0.18 +

[Utente='(1597.5-1656.5]'] * 3.2 +

[Utente='(1656.5-1711.5]'] * 1.42 +

[Utente='(1711.5-1860.5]'] * -0.41 +

[Utente='(1860.5-2983.5]'] * 0.84 +

[Utente='(2983.5-3967.5]'] * 0.52 +

[Utente='(3967.5-4268.5]'] * 0.32 +

[Utente='(4460.5-5763.5]'] * 0.38 +

[Utente='(7340.5-8568.5]'] * -0.43 +

[Utente='(9446.5-10414.5]'] * -0.76 +

[Utente='(10414.5-10673.5]'] * -0.26 +

[Utente='(10673.5-11661.5]'] * -0.63 +

[Utente='(11879.5-12148.5]'] * 1.02 +

[Utente='(12148.5-inf)'] * -0.44 +

[Género Sexual] * 0.47 +

[Ocupação=Activo] * 0.33 +

[Ocupação=Reformado] * 0.46 +

[Ocupação=Não Aplicável] * 1.83 +

[Idade='(-inf-16.5]'] * -2.46 +

[Idade='(16.5-27.5]'] * -0.37 +

[Idade='(27.5-40.5]'] * -0.63 +

[Idade='(40.5-47.5]'] * -0.47 +

[Idade='(52.5-57.5]'] * 0.21 +

[Idade='(57.5-65.5]'] * 0.18 +

[IMC='(19.5-21.5]'] * -0.59 +

[IMC='(27.5-30.5]'] * 0.1 +

[IMC='(30.5-inf)'] * 0.13 +

[PAS='(117.5-120.5]'] * -0.31 +

[PAS='(132.5-140.5]'] * 0.13 +

[PAD='(-inf-53.5]'] * -1.82 +

[PAD='(60.5-70.5]'] * -0.25 +

[PAD='(75.5-81.5]'] * 0.36 +

[PAD='(81.5-90.5]'] * 0.07 +

[PAD='(90.5-inf)'] * -0.41

Class 3 :

-1.21 +

[Utente='(564.5-912.5]'] * -0.66 +

[Utente='(912.5-1467.5]'] * 0.35 +

[Utente='(1656.5-1711.5]'] * -0.96 +

[Utente='(5763.5-7340.5]'] * -0.24 +

[Utente='(8970.5-9446.5]'] * 0.66 +

[Utente='(9446.5-10414.5]'] * 0.18 +

[Utente='(10673.5-11661.5]'] * 0.61 +

[Utente='(11661.5-11879.5]'] * 0.4 +

[Utente='(12148.5-inf)'] * -0.42 +

[Género Sexual] * 0.25 +

[Ocupação=Activo] * 0.35 +

[Ocupação=Reformado] * 0.47 +

[Idade='(-inf-16.5]'] * -0.9 +

[Idade='(16.5-27.5]'] * -2.48 +

[Idade='(27.5-40.5]'] * -2.39 +

[Idade='(40.5-47.5]'] * -1.19 +

[Idade='(47.5-52.5]'] * -0.8 +

[Idade='(65.5-70.5]'] * 0.17 +

[Idade='(70.5-inf)'] * 0.22 +

[IMC='(17.5-19.5]'] * -0.66 +

[IMC='(19.5-21.5]'] * -0.48 +

[IMC='(21.5-23.5]'] * -0.41 +

[IMC='(23.5-25.5]'] * -0.14 +

[IMC='(27.5-30.5]'] * 0.35 +

[IMC='(30.5-inf)'] * 0.57 +

[PAS='(-inf-103.5]'] * -2.53 +

[PAS='(117.5-120.5]'] * -0.4 +

[PAS='(132.5-140.5]'] * 0.41 +

[PAS='(140.5-inf)'] * 0.95 +

[PAD='(70.5-75.5]'] * 0.14 +

[PAD='(75.5-81.5]'] * 0.12

Class 4 :

-8.93 +

[Utente='(564.5-912.5]'] * 1.51 +

[Utente='(4268.5-4460.5]'] * 3.45 +

[Utente='(4460.5-5763.5]'] * -0.83 +

[Utente='(5763.5-7340.5]'] * 0.75 +

[Utente='(7340.5-8568.5]'] * 0.67 +

[Utente='(12148.5-inf)'] * 1.19 +

[Género Sexual] * -3.5 +

[Ocupação=Activo] * 3.5 +

[Ocupação=Reformado] * -0.05 +

[Idade='(16.5-27.5]'] * 3.57 +

[Idade='(27.5-40.5]'] * 1.63 +

[Idade='(40.5-47.5]'] * 0.68 +

[Idade='(47.5-52.5]'] * -0.86 +

[Idade='(52.5-57.5]'] * -0.84 +

[Idade='(57.5-65.5]'] * -0.84 +

[IMC='(19.5-21.5]'] * -0.84 +

[IMC='(21.5-23.5]'] * -1.71 +

[IMC='(23.5-25.5]'] * 0.87 +

[IMC='(27.5-30.5]'] * -1.73 +

[PAS='(-inf-103.5]'] * -0.82 +

[PAS='(117.5-120.5]'] * -1.73 +

[PAS='(126.5-132.5]'] * -2.7 +

[PAD='(60.5-70.5]'] * -1.73 +

[PAD='(81.5-90.5]'] * 0.31 +

[PAD='(90.5-inf)'] * 0.46

Time taken to build model: 862.39 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	10133	75.7041 %
Incorrectly Classified Instances	3252	24.2959 %
Kappa statistic	0.4824	
Mean absolute error	0.1268	
Root mean squared error	0.2516	
Relative absolute error	62.6872 %	
Root relative squared error	79.1343 %	
Total Number of Instances	13385	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.929	0.287	0.867	0.929	0.897	0.937	Normal
	0.632	0.164	0.485	0.632	0.548	0.857	Hipertensão
	0.089	0.005	0.533	0.089	0.152	0.838	Diabetes
	0.094	0.013	0.395	0.094	0.152	0.883	Diabetes com Hipertensão
	0	0	0	0	0	0.607	Diabetes Gestacional
Weighted Avg.	0.757	0.225	0.735	0.757	0.727	0.911	

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as

8308 596 17 23 0 | a = Normal

843 1659 23 102 0 | b = Hipertensão

254 384 65 30 0 | c = Diabetes

173 785 17 101 0 | d = Diabetes com Hipertensão

5 0 0 0 0 | e = Diabetes Gestacional

8.1.2. Algoritmo LMT+Discretize-Rfirst-last para o CS Eiras

=== Run information ===

Scheme:weka.classifiers.trees.LMT -I -1 -M 15 -W 0.0

Relation: Class_Eiras-weka.filters.supervised.attribute.Discretize-Rfirst-last

Instances: 5945

Attributes: 10

Utente

Consultas

Género Sexual

Óbito

Ocupação

Idade

IMC

PAS

PAD

Diagnóstico

Test mode:5-fold cross-validation

=== Classifier model (full training set) ===

Logistic model tree

: LM_1:67/67 (5945)

Number of Leaves : 1

Size of the Tree : 1

LM_1:

Class 0 :

-1.09 +

[Utente='(449.5-1271.5]'] * -0.28 +

[Utente='(1271.5-1760.5]'] * -0.23 +

[Utente='(1760.5-3478.5]'] * -0.16 +

[Utente='(3478.5-4002.5]'] * 0.4 +

[Utente='(4140.5-4590.5]'] * 0.31 +

[Utente='(4706.5-5026.5]'] * 0.65 +

[Utente='(5026.5-5452.5]'] * 0.74 +

[Utente='(5452.5-5560.5]'] * 0.47 +

[Utente='(5560.5-5807.5]'] * 0.66 +

[Consultas='(-inf-5.5]'] * -1.83 +

[Consultas='(5.5-7.5]'] * -0.81 +

[Consultas='(7.5-10.5]'] * -0.76 +

[Consultas='(26.5-inf)'] * 0.53 +

[Género Sexual] * 0.7 +

[Ocupação=Não Activo] * -0.15 +

[Ocupação=Não Aplicável] * -0.94 +

[Idade='(-inf-22.5]'] * -1.62 +

[Idade='(22.5-36.5]'] * -5.01 +

[Idade='(36.5-42.5]'] * -0.25 +

[Idade='(42.5-48.5]'] * 0.28 +

[Idade='(48.5-57.5]'] * 0.34 +

[Idade='(57.5-63.5]'] * 0.7 +

[Idade='(63.5-inf)'] * 1.37 +

[IMC='(-inf-18.5]'] * -1.86 +

[IMC='(18.5-21.5]'] * -0.19 +

[IMC='(21.5-24.5]'] * -0.49 +

[IMC='(25.5-28.5]'] * 0.09 +

[IMC='(28.5-inf)'] * 0.94 +

[PAS='(-inf-112.5]'] * -1.7 +

[PAS='(112.5-125.5]'] * -0.27 +

[PAS='(125.5-132.5]'] * -0.05 +

[PAS='(140.5-158.5]'] * 0.4 +

[PAS='(158.5-inf)'] * 0.19 +

[PAD='(-inf-56.5]'] * 0.13 +

[PAD='(56.5-63.5]'] * 0.52 +

[PAD='(91.5-inf)'] * -0.2

Class 1 :

2.52 +

[Utente='(-inf-449.5]'] * 0.14 +

[Utente='(449.5-1271.5]'] * 0.07 +

[Utente='(1271.5-1760.5]'] * 0.37 +

[Utente='(3478.5-4002.5]'] * -0.83 +

[Utente='(4140.5-4590.5]'] * -0.37 +

[Utente='(4590.5-4706.5]'] * -0.39 +

[Utente='(4706.5-5026.5]'] * -0.28 +

[Utente='(5026.5-5452.5]'] * -0.25 +

[Utente='(5560.5-5807.5]'] * -0.34 +

[Consultas='(-inf-5.5]'] * 1.09 +

[Consultas='(5.5-7.5]'] * 1.09 +

[Consultas='(7.5-10.5]'] * 0.46 +

[Consultas='(17.5-26.5]'] * -0.46 +

[Consultas='(26.5-inf)'] * -0.94 +

[Género Sexual] * -0.92 +

[Ocupação=Reformado] * -0.24 +

[Ocupação=Não Activo] * 0.05 +

[Ocupação=Estudante] * -0.12 +

[Idade='(-inf-22.5]'] * 1.53 +

[Idade='(22.5-36.5]'] * 0.05 +

[Idade='(42.5-48.5]'] * -0.26 +

[Idade='(48.5-57.5]'] * -1.23 +

[Idade='(57.5-63.5]'] * -1.7 +

[Idade='(63.5-inf)'] * -2.55 +

[IMC='(-inf-18.5]'] * 1.51 +

[IMC='(21.5-24.5]'] * 0.09 +

[IMC='(25.5-28.5]'] * -0.3 +

[IMC='(28.5-inf)'] * -0.36 +

[PAS='(-inf-112.5]'] * 0.41 +

[PAS='(112.5-125.5]'] * 0.84 +

[PAS='(125.5-132.5]'] * 0.3 +

[PAS='(140.5-158.5]'] * -0.18 +

[PAS='(158.5-inf)'] * -0.81 +

[PAD='(56.5-63.5]'] * -0.1 +

[PAD='(76.5-82.5]'] * -0.1 +

[PAD='(82.5-91.5]'] * -0.5 +

[PAD='(91.5-inf)'] * -0.75

Class 2 :

1.02 +

[Utente='(-inf-449.5]'] * 0.19 +

[Utente='(1271.5-1760.5]'] * 0.08 +

[Utente='(3478.5-4002.5]'] * 0.9 +

[Utente='(4002.5-4140.5]'] * -0.19 +

[Utente='(4590.5-4706.5]'] * -0.48 +

[Utente='(4706.5-5026.5]'] * -0.35 +

[Utente='(5026.5-5452.5]'] * -0.3 +

[Utente='(5807.5-inf)'] * -0.92 +

[Consultas='(-inf-5.5]'] * -0.59 +

[Consultas='(17.5-26.5]'] * -0.09 +

[Consultas='(26.5-inf)'] * 0.04 +

[Ocupação=Não Activo] * -0.14 +

[Ocupação=Estudante] * 0.49 +

[Ocupação=Não Aplicável] * 0.93 +

[Ocupação=Desconhecido] * -1.96 +

[Idade='(-inf-22.5]'] * -1.96 +

[Idade='(22.5-36.5]'] * -1.27 +

[Idade='(36.5-42.5]'] * -0.23 +

[Idade='(42.5-48.5]'] * 0.46 +

[Idade='(48.5-57.5]'] * -0.03 +

[Idade='(63.5-inf)'] * 0.45 +

[IMC='(18.5-21.5]'] * -0.03 +

[IMC='(24.5-25.5]'] * 0.04 +

[IMC='(28.5-inf)'] * 0.2 +

[PAS='(-inf-112.5]'] * -1.57 +

[PAS='(112.5-125.5]'] * -0.28 +

[PAS='(125.5-132.5]'] * -0.32 +

[PAS='(132.5-140.5]'] * -0.22 +

[PAD='(-inf-56.5]'] * -0.3 +

[PAD='(63.5-76.5]'] * -0.05 +

[PAD='(76.5-82.5]'] * 0.09 +

[PAD='(82.5-91.5]'] * 0.26 +

[PAD='(91.5-inf)'] * 0.43

Class 3 :

-1.37 +

[Utente='(-inf-449.5]'] * -1.44 +

[Utente='(1271.5-1760.5]'] * -0.38 +

[Utente='(1760.5-3478.5]'] * -0.1 +

[Utente='(3478.5-4002.5]'] * 0.51 +

[Utente='(4002.5-4140.5]'] * 0.95 +

[Utente='(4140.5-4590.5]'] * 0.99 +

[Utente='(4590.5-4706.5]'] * 0.43 +

[Utente='(4706.5-5026.5]'] * 0.26 +

[Utente='(5026.5-5452.5]'] * 0.24 +

[Utente='(5452.5-5560.5]'] * -1.66 +

[Utente='(5560.5-5807.5]'] * -0.55 +

[Utente='(5807.5-inf)'] * -3.29 +

[Consultas='(-inf-5.5]'] * -0.41 +

[Consultas='(5.5-7.5]'] * -0.53 +

[Consultas='(7.5-10.5]'] * -0.08 +

[Consultas='(10.5-17.5]'] * 0.1 +

[Consultas='(26.5-inf)'] * 0.31 +

[Género Sexual] * 0.66 +

[Ocupação=Activo] * -0.31 +

[Ocupação=Estudante] * 0.36 +

[Ocupação=Não Aplicável] * -0.81 +

[Idade='(-inf-22.5]'] * -0.52 +

[Idade='(22.5-36.5]'] * -0.78 +

[Idade='(36.5-42.5]'] * 0.4 +

[Idade='(42.5-48.5]'] * 0.74 +

[Idade='(57.5-63.5]'] * -0.27 +

[IMC='(-inf-18.5]'] * -0.81 +

[IMC='(18.5-21.5]'] * 0.58 +

[IMC='(24.5-25.5]'] * -0.32 +

[IMC='(25.5-28.5]'] * -0.07 +

[IMC='(28.5-inf)'] * 0.17 +

[PAS='(-inf-112.5]'] * -0.47 +

[PAS='(112.5-125.5]'] * 0.36 +

[PAS='(125.5-132.5]'] * 0.63 +

[PAS='(140.5-158.5]'] * -0.22 +

[PAS='(158.5-inf)'] * -1.8 +

[PAD='(-inf-56.5]'] * -0.3 +

[PAD='(56.5-63.5]'] * -0.58 +

[PAD='(82.5-91.5]'] * -0.11

Class 4 :

-11.21 +

[Utente='(449.5-1271.5]'] * 1.75 +

[Utente='(1760.5-3478.5]'] * 2.23 +

[Utente='(4706.5-5026.5]'] * -0.85 +

[Consultas='(-inf-5.5]'] * -2.67 +

[Consultas='(5.5-7.5]'] * -1.68 +

[Consultas='(17.5-26.5]'] * 0.36 +

[Consultas='(26.5-inf)'] * -2.57 +

[Género Sexual] * -2.72 +

[Ocupação=Activo] * 3.26 +

[Idade='(22.5-36.5]'] * 1.82 +

[Idade='(36.5-42.5]'] * 2.65 +

[Idade='(48.5-57.5]'] * -1.71 +

[Idade='(57.5-63.5]'] * -0.86 +

[IMC='(18.5-21.5]'] * -2.58 +

[IMC='(21.5-24.5]'] * -2.62 +

[IMC='(24.5-25.5]'] * -1.7 +

[IMC='(28.5-inf)'] * 0.92 +

[PAS='(-inf-112.5]'] * 2.65 +

[PAS='(112.5-125.5]'] * -0.84 +

[PAS='(132.5-140.5]'] * 2.1 +

[PAS='(140.5-158.5]'] * -2.56 +

[PAS='(158.5-inf)'] * -0.82 +

[PAD='(56.5-63.5]'] * 0.69 +

[PAD='(76.5-82.5]'] * -1.69 +

[PAD='(82.5-91.5]'] * -1.71 +

[PAD='(91.5-inf)'] * 2.57

Time taken to build model: 370.91 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	4529	76.1817 %
Incorrectly Classified Instances	1416	23.8183 %
Kappa statistic	0.5928	
Mean absolute error	0.1248	
Root mean squared error	0.2506	
Relative absolute error	51.9941 %	
Root relative squared error	72.3431 %	
Total Number of Instances	5945	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.466	0.059	0.582	0.466	0.517	0.901	Diabetes com Hipertensão
	0.92	0.135	0.893	0.92	0.907	0.966	Normal
	0.683	0.174	0.593	0.683	0.635	0.868	Hipertensão
	0.006	0	0.333	0.006	0.012	0.797	Diabetes
	0	0	0	0	0	0.443	Diabetes Gestacional
Weighted Avg.	0.762	0.13	0.749	0.762	0.749	0.924	

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
415	41	434	1	0	a = Diabetes com Hipertensão
16	3015	244	0	1	b = Normal

244 264 1098 1 0 | c = Hipertensão

38 52 77 1 0 | d = Diabetes

0 3 0 0 0 | e = Diabetes Gestacional

8.1.3. Algoritmo LMT+Discretize-Rfirst-last para o CS Fundão

=== Run information ===

Scheme:weka.classifiers.trees.LMT -I -1 -M 15 -W 0.0

Relation: Class_Fundão-weka.filters.supervised.attribute.Discretize-Rfirst-last

Instances: 3158

Attributes: 10

Utente

Consultas

Género Sexual

Óbito

Ocupação

Idade

IMC

PAS

PAD

Diagnóstico

Test mode:5-fold cross-validation

=== Classifier model (full training set) ===

Logistic model tree

: LM_1:51/51 (3158)

Number of Leaves : 1

Size of the Tree : 1

LM_1:

Class 0 :

0.09 +

[Utente='(-inf-863.5]'] * 0.01 +

[Utente='(863.5-1127.5]'] * -0.09 +

[Utente='(1127.5-1787.5]'] * -0.49 +

[Utente='(1787.5-2331.5]'] * -0.76 +

[Utente='(2331.5-inf)'] * 0.06 +

[Consultas='(-inf-4.5]'] * -0.37 +

[Consultas='(4.5-8.5]'] * -0.03 +

[Consultas='(8.5-13.5]'] * 0.08 +

[Consultas='(13.5-inf)'] * 0.49 +

[Género Sexual] * -0.48 +

[Ocupação=Activo] * -0.02 +

[Ocupação=Estudante] * 0.8 +

[Ocupação=Não Activo] * 0.01 +

[Ocupação=Não Aplicável] * -2.4 +

[Idade='(-inf-35.5]'] * -1.72 +

[Idade='(35.5-48.5]'] * -0.72 +

[Idade='(48.5-58.5]'] * -0.56 +

[IMC='(-inf-19.5]'] * -0.43 +

[IMC='(19.5-24.5]'] * 0.08 +

[IMC='(24.5-27.5]'] * -0.05 +

[IMC='(27.5-inf)'] * 0.28 +

[PAS='(-inf-110.5]'] * -0.03 +

[PAS='(110.5-126.5]'] * 0.6 +

[PAS='(126.5-140.5]'] * 0.49 +

[PAD='(-inf-60.5]'] * 0.05 +

[PAD='(69.5-74.5]'] * 0.17 +

[PAD='(82.5-inf)'] * -0.09

Class 1 :

1.51 +

[Utente='(863.5-1127.5]'] * -0.03 +

[Utente='(1127.5-1787.5]'] * 0.06 +

[Utente='(1787.5-2331.5]'] * 0.03 +

[Utente='(2331.5-inf)'] * -0.25 +

[Consultas='(-inf-4.5]'] * 1.25 +

[Consultas='(4.5-8.5]'] * 0.62 +

[Consultas='(13.5-inf)'] * -0.05 +

[Género Sexual] * 0.53 +

[Ocupação=Estudante] * -0.22 +

[Ocupação=Não Aplicável] * 2.43 +

[Idade='(-inf-35.5]'] * 1.4 +

[Idade='(35.5-48.5]'] * 1.14 +

[Idade='(48.5-58.5]'] * 0.14 +

[Idade='(58.5-68.5]'] * -0.51 +

[Idade='(68.5-inf)'] * -0.9 +

[IMC='(-inf-19.5]'] * 0.38 +

[IMC='(24.5-27.5]'] * -0.18 +

[IMC='(27.5-inf)'] * -0.32 +

[PAS='(126.5-140.5]'] * -0.89 +

[PAS='(140.5-inf)'] * -1.91 +

[PAD='(60.5-69.5]'] * 0.3 +

[PAD='(69.5-74.5]'] * 0.07 +

[PAD='(74.5-82.5]'] * -0.46 +

[PAD='(82.5-inf)'] * -0.12

Class 2 :

0.4 +

[Utente='(-inf-863.5]'] * -0.02 +

[Utente='(863.5-1127.5]'] * 0.5 +

[Utente='(1127.5-1787.5]'] * 0.02 +

[Utente='(2331.5-inf)'] * -0.04 +

[Consultas='(13.5-inf)'] * 0.08 +

[Ocupação=Não Activo] * 0.24 +

[Ocupação=Não Aplicável] * -2.41 +

[Idade='(-inf-35.5]'] * -1.67 +

[Idade='(35.5-48.5]'] * -0.51 +

[Idade='(48.5-58.5]'] * -0.3 +

[Idade='(68.5-inf)'] * 0.21 +

[IMC='(19.5-24.5]'] * -0.05 +

[IMC='(27.5-inf)'] * 0.41 +

[PAS='(-inf-110.5]'] * -0.68 +

[PAS='(110.5-126.5]'] * 0.29 +

[PAS='(140.5-inf)'] * -0.3 +

[PAD='(-inf-60.5]'] * -0.44 +

[PAD='(60.5-69.5]'] * -0.03 +

[PAD='(74.5-82.5]'] * 0.08 +

[PAD='(82.5-inf)'] * 0.82

Class 3 :

-1.16 +

[Utente='(863.5-1127.5]'] * 0.18 +

[Utente='(1127.5-1787.5]'] * -0.7 +

[Utente='(2331.5-inf)'] * 0.14 +

[Consultas='(-inf-4.5]'] * -1.64 +

[Consultas='(4.5-8.5]'] * -0.39 +

[Consultas='(8.5-13.5]'] * 0.03 +

[Consultas='(13.5-inf)'] * 1.12 +

[Género Sexual] * -0.53 +

[Ocupação=Não Activo] * -0.21 +

[Idade='(-inf-35.5]'] * -5.82 +

[Idade='(35.5-48.5]'] * -0.24 +

[Idade='(58.5-68.5]'] * 0.61 +

[Idade='(68.5-inf)'] * 1.3 +

[IMC='(-inf-19.5]'] * -4.36 +

[IMC='(19.5-24.5]'] * -0.59 +

[IMC='(24.5-27.5]'] * 0.05 +

[IMC='(27.5-inf)'] * 0.62 +

[PAS='(-inf-110.5]'] * -0.06 +

[PAS='(126.5-140.5]'] * 0.23 +

[PAD='(-inf-60.5]'] * 0.45 +

[PAD='(69.5-74.5]'] * -0.05 +

[PAD='(82.5-inf)'] * 0.5

Class 4 :

-41.23 +

[Utente='(2331.5-inf)'] * 4.1 +

[Consultas='(8.5-13.5]'] * 4.8 +

[Ocupação=Não Activo] * 5.84 +

[Idade='(35.5-48.5]'] * 8.56 +

[IMC='(27.5-inf)'] * 8.91 +

[PAS='(110.5-126.5]'] * 2.8 +

[PAS='(140.5-inf)'] * -0.06 +

[PAD='(60.5-69.5]'] * 9.3

Time taken to build model: 109.68 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	2412	76.3775 %
Incorrectly Classified Instances	746	23.6225 %
Kappa statistic	0.4877	
Mean absolute error	0.1196	
Root mean squared error	0.2456	
Relative absolute error	60.6198 %	
Root relative squared error	78.2585 %	
Total Number of Instances	3158	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class

0.037	0.013	0.196	0.037	0.062	0.822	Diabetes
0.943	0.274	0.882	0.943	0.912	0.948	Normal
0.545	0.124	0.461	0.545	0.499	0.865	Hipertensão
0.347	0.037	0.429	0.347	0.384	0.911	Diabetes com Hipertensão
0	0	0	0	0	0.12	Diabetes Gestacional
Weighted Avg.	0.764	0.211	0.727	0.764	0.74	0.922

=== Confusion Matrix ===

a	b	c	d	e	<-- classified as
9	80	112	42	0	a = Diabetes
9	2041	102	12	0	b = Normal
16	163	280	55	0	c = Hipertensão
12	28	114	82	0	d = Diabetes com Hipertensão
0	1	0	0	0	e = Diabetes Gestacional

8.1.4. Algoritmo LMT+Discretize-Rfirst-last para o CS Tábua.

=== Run information ===

Scheme:weka.classifiers.trees.LMT -I -1 -M 15 -W 0.0

Relation: Class_Tábua-weka.filters.supervised.attribute.Discretize-Rfirst-last

Instances: 5928

Attributes: 10

Utente

Consultas

Género Sexual

Óbito

Ocupação

Idade

IMC

PAS

PAD

Diagnóstico

Test mode:5-fold cross-validation

=== Classifier model (full training set) ===

Logistic model tree

: LM_1:94/94 (5928)

Number of Leaves : 1

Size of the Tree : 1

LM_1:

Class 0 :

-0.12 +

[Utente='(5252.5-inf)'] * -0.13 +

[Consultas='(-inf-4.5]'] * -0.35 +

[Consultas='(4.5-5.5]'] * -0.38 +

[Consultas='(17.5-inf)'] * 0.01 +

[Ocupação=Reformado] * 0.55 +

[Ocupação=Não Activo] * 0.77 +

[Ocupação=Estudante] * 1.04 +

[Ocupação=Activo] * 0.65 +

[Ocupação=Não Aplicável] * -3 +

[Idade='(-inf-26.5]'] * -2.59 +

[Idade='(26.5-33.5]'] * -0.9 +

[Idade='(33.5-42.5]'] * -0.77 +

[Idade='(42.5-47.5]'] * -0.6 +

[Idade='(47.5-55.5]'] * -0.4 +

[Idade='(55.5-64.5]'] * 0.12 +

[Idade='(64.5-71.5]'] * 0.07 +

[Idade='(71.5-inf)'] * 0.21 +

[IMC='(-inf-17.5]'] * -1.29 +

[IMC='(17.5-22.5]'] * -0.06 +

[PAS='(-inf-102.5]'] * -1.44 +

[PAS='(102.5-110.5]'] * -0.83 +

[PAS='(110.5-120.5]'] * -0.05 +

[PAS='(120.5-126.5]'] * -0.02 +

[PAS='(126.5-131.5]'] * 0.17 +

[PAS='(131.5-140.5]'] * 0.08 +

[PAS='(140.5-147.5]'] * 0.19 +

[PAS='(147.5-inf)'] * 0.02 +

[PAD='(-inf-50.5]'] * 0.04 +

[PAD='(50.5-55.5]'] * -0.09 +

[PAD='(55.5-59.5]'] * -0.17 +

[PAD='(59.5-60.5]'] * 0.13 +

[PAD='(70.5-79.5]'] * 0.23 +

[PAD='(79.5-80.5]'] * 0.4 +

[PAD='(80.5-inf)'] * 0.84

Class 1 :

1.38 +

[Utente='(-inf-2002.5]'] * -0.09 +

[Utente='(2002.5-2765.5]'] * -0.05 +

[Utente='(5252.5-inf)'] * 0.48 +

[Consultas='(-inf-4.5]'] * 1.01 +

[Consultas='(8.5-17.5]'] * -0.31 +

[Consultas='(17.5-inf)'] * -0.85 +

[Género Sexual] * 0.35 +

[Ocupação=Reformado] * -0.27 +

[Ocupação=Activo] * -0.22 +

[Idade='(-inf-26.5]'] * 1.26 +

[Idade='(26.5-33.5]'] * 1.62 +

[Idade='(33.5-42.5]'] * 0.95 +

[Idade='(42.5-47.5]'] * 0.64 +

[Idade='(47.5-55.5]'] * 0.2 +

[Idade='(64.5-71.5]'] * -0.67 +

[Idade='(71.5-inf)'] * -1.39 +

[IMC='(-inf-17.5]'] * 0.56 +

[IMC='(17.5-22.5]'] * 0.79 +

[IMC='(22.5-26.5]'] * 0.4 +

[IMC='(30.5-inf)'] * -0.4 +

[PAS='(102.5-110.5]'] * 0.06 +

[PAS='(110.5-120.5]'] * 0.25 +

[PAS='(131.5-140.5]'] * -0.46 +

[PAS='(140.5-147.5]'] * -0.72 +

[PAS='(147.5-inf)'] * -1.51 +

[PAD='(50.5-55.5]'] * -0.52 +

[PAD='(55.5-59.5]'] * -0.07 +

[PAD='(59.5-60.5]'] * 0.33 +

[PAD='(60.5-69.5]'] * -0.2 +

[PAD='(69.5-70.5]'] * 0.11 +

[PAD='(70.5-79.5]'] * -0.12 +

[PAD='(80.5-inf)'] * 0.03

Class 2 :

-0.02 +

[Utente='(-inf-2002.5]'] * 0.88 +

[Utente='(2002.5-2765.5]'] * 0.1 +

[Utente='(2765.5-5252.5]'] * -1.22 +

[Utente='(5252.5-inf)'] * -0.67 +

[Consultas='(-inf-4.5]'] * -0.89 +

[Consultas='(4.5-5.5]'] * -0.18 +

[Consultas='(5.5-8.5]'] * -0.12 +

[Consultas='(8.5-17.5]'] * 0.12 +

[Consultas='(17.5-inf)'] * 0.34 +

[Género Sexual] * -0.37 +

[Ocupação=Reformado] * -0.16 +

[Ocupação=Activo] * 0.09 +

[Idade='(-inf-26.5]'] * -6.4 +

[Idade='(26.5-33.5]'] * -5.36 +

[Idade='(33.5-42.5]'] * -1.6 +

[Idade='(42.5-47.5]'] * -0.8 +

[Idade='(47.5-55.5]'] * -0.52 +

[Idade='(71.5-inf)'] * 0.07 +

[IMC='(-inf-17.5]'] * -0.43 +

[IMC='(17.5-22.5]'] * -0.52 +

[IMC='(22.5-26.5]'] * -0.11 +

[IMC='(26.5-30.5]'] * 0.29 +

[IMC='(30.5-inf)'] * 0.75 +

[PAS='(-inf-102.5]'] * -1.01 +

[PAS='(102.5-110.5]'] * -1.56 +

[PAS='(110.5-120.5]'] * -0.69 +

[PAS='(120.5-126.5]'] * -0.08 +

[PAS='(140.5-147.5]'] * 0.12 +

[PAS='(147.5-inf)'] * 0.09 +

[PAD='(-inf-50.5]'] * -0.2 +

[PAD='(50.5-55.5]'] * 0.81 +

[PAD='(55.5-59.5]'] * 1.03 +

[PAD='(59.5-60.5]'] * -0.13 +

[PAD='(60.5-69.5]'] * 0.17 +

[PAD='(69.5-70.5]'] * -0.36 +

[PAD='(79.5-80.5]'] * -0.98

Class 3 :

-1.69 +

[Utente='(-inf-2002.5]'] * 1.27 +

[Utente='(2002.5-2765.5]'] * -0.05 +

[Utente='(2765.5-5252.5]'] * -0.28 +

[Utente='(5252.5-inf)'] * 0.56 +

[Consultas='(-inf-4.5]'] * -0.39 +

[Consultas='(4.5-5.5]'] * 0.05 +

[Consultas='(8.5-17.5]'] * 0.46 +

[Consultas='(17.5-inf)'] * -0.02 +

[Género Sexual] * -0.58 +

[Ocupação=Reformado] * -0.36 +

[Ocupação=Não Activo] * 0.34 +

[Ocupação=Não Aplicável] * -0.5 +

[Idade='(-inf-26.5]'] * -1.06 +

[Idade='(26.5-33.5]'] * -0.76 +

[Idade='(33.5-42.5]'] * -0.28 +

[Idade='(42.5-47.5]'] * 0.81 +

[Idade='(47.5-55.5]'] * 0.03 +

[Idade='(55.5-64.5]'] * 0.34 +

[Idade='(64.5-71.5]'] * 0.07 +

[Idade='(71.5-inf)'] * -0.73 +

[IMC='(-inf-17.5]'] * -0.37 +

[IMC='(22.5-26.5]'] * -0.1 +

[IMC='(30.5-inf)'] * 0.06 +

[PAS='(110.5-120.5]'] * 0.15 +

[PAS='(120.5-126.5]'] * 1.18 +

[PAS='(126.5-131.5]'] * 0.66 +

[PAS='(140.5-147.5]'] * -0.75 +

[PAS='(147.5-inf)'] * -1.03 +

[PAD='(-inf-50.5]'] * 1.34 +

[PAD='(50.5-55.5]'] * 0.37 +

[PAD='(55.5-59.5]'] * -0.52 +

[PAD='(59.5-60.5]'] * -0.5 +

[PAD='(60.5-69.5]'] * 0.04 +

[PAD='(69.5-70.5]'] * -0.75 +

[PAD='(79.5-80.5]'] * -0.04 +

[PAD='(80.5-inf)'] * -0.35

Time taken to build model: 395.68 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	4425	74.6457 %
Incorrectly Classified Instances	1503	25.3543 %
Kappa statistic	0.5545	
Mean absolute error	0.1643	
Root mean squared error	0.287	
Relative absolute error	56.3439 %	
Root relative squared error	75.1625 %	
Total Number of Instances	5928	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.662	0.169	0.601	0.662	0.63	0.863	Hipertensão
	0.892	0.181	0.866	0.892	0.879	0.94	Normal
	0.427	0.059	0.516	0.427	0.467	0.9	Diabetes com Hipertensão
	0.04	0.001	0.462	0.04	0.074	0.793	Diabetes
Weighted Avg.	0.746	0.158	0.737	0.746	0.736	0.91	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
1090	332	225	0	a = Hipertensão

317 3002 46 1 | b = Normal

369 63 327 6 | c = Diabetes com Hipertensão

39 69 36 6 | d = Diabetes

8.1.5. Algoritmo LMT+Discretize-Rfirst-last para classificação binária

=== Run information ===

Scheme:weka.classifiers.trees.LMT -I -1 -M 15 -W 0.0

Relation: Classification-weka.filters.unsupervised.attribute.Remove-R1,4-weka.filters.supervised.attribute.Discretize-Rfirst-last

Instances: 28416

Attributes: 8

Consultas

Género Sexual

Ocupação

Idade

IMC

PAS

PAD

Diagnóstico

Test mode:5-fold cross-validation

=== Classifier model (full training set) ===

Logistic model tree

: LM_1:60/60 (28416)

Number of Leaves : 1

Size of the Tree : 1

LM_1:

Class 0 :

0.47 +

[Consultas='(-inf-5.5]'] * 0.34 +

[Consultas='(5.5-8.5]'] * 0.19 +

[Consultas='(8.5-11.5]'] * 0.06 +

[Consultas='(15.5-22.5]'] * -0.17 +

[Consultas='(22.5-36.5]'] * -0.25 +

[Consultas='(36.5-inf)'] * -0.46 +

[Género Sexual] * 0.28 +

[Ocupação=Reformado] * -0.15 +

[Idade='(-inf-22.5]'] * 1.62 +

[Idade='(22.5-27.5]'] * 1.3 +

[Idade='(27.5-35.5]'] * 1.01 +

[Idade='(35.5-42.5]'] * 0.61 +

[Idade='(42.5-48.5]'] * 0.2 +

[Idade='(52.5-55.5]'] * -0.12 +

[Idade='(55.5-57.5]'] * -0.21 +

[Idade='(57.5-63.5]'] * -0.41 +

[Idade='(63.5-70.5]'] * -0.63 +

[Idade='(70.5-inf)'] * -1.05 +

[IMC='(-inf-17.5]'] * 0.37 +

[IMC='(18.5-20.5]'] * 0.21 +

[IMC='(20.5-22.5]'] * 0.13 +

[IMC='(25.5-27.5]'] * -0.17 +

[IMC='(27.5-29.5]'] * -0.25 +

[IMC='(29.5-inf)'] * -0.46 +

[PAS='(-inf-100.5]'] * 0.42 +

[PAS='(100.5-112.5]'] * 0.34 +

[PAS='(120.5-126.5]'] * -0.19 +

[PAS='(126.5-132.5]'] * -0.32 +

[PAS='(132.5-140.5]'] * -0.51 +

[PAS='(140.5-160.5]'] * -0.8 +

[PAS='(160.5-189.5]'] * -0.94 +

[PAS='(189.5-inf)'] * -1.53 +

[PAD='(53.5-60.5]'] * 0.14 +

[PAD='(60.5-70.5]'] * 0.16 +

[PAD='(70.5-75.5]'] * 0.05 +

[PAD='(80.5-90.5]'] * -0.06 +

[PAD='(90.5-inf)'] * -0.09

Class 1 :

-0.47 +

[Consultas='(-inf-5.5]'] * -0.34 +

[Consultas='(5.5-8.5]'] * -0.19 +

[Consultas='(8.5-11.5]'] * -0.06 +

[Consultas='(15.5-22.5]'] * 0.17 +

[Consultas='(22.5-36.5]'] * 0.25 +

[Consultas='(36.5-inf)'] * 0.46 +

[Género Sexual] * -0.28 +

[Ocupação=Reformado] * 0.15 +

[Idade='(-inf-22.5]'] * -1.62 +

[Idade='(22.5-27.5]'] * -1.3 +

[Idade='(27.5-35.5]'] * -1.01 +

[Idade='(35.5-42.5]'] * -0.61 +

[Idade='(42.5-48.5]'] * -0.2 +

[Idade='(52.5-55.5]'] * 0.12 +

[Idade='(55.5-57.5]'] * 0.21 +

[Idade='(57.5-63.5]'] * 0.41 +

[Idade='(63.5-70.5]'] * 0.63 +

[Idade='(70.5-inf)'] * 1.05 +

[IMC='(-inf-17.5]'] * -0.37 +

[IMC='(18.5-20.5]'] * -0.21 +

[IMC='(20.5-22.5]'] * -0.13 +

[IMC='(25.5-27.5]'] * 0.17 +

[IMC='(27.5-29.5]'] * 0.25 +

[IMC='(29.5-inf)'] * 0.46 +

[PAS='(-inf-100.5]'] * -0.42 +

[PAS='(100.5-112.5]'] * -0.34 +

[PAS='(120.5-126.5]'] * 0.19 +

[PAS='(126.5-132.5]'] * 0.32 +

[PAS='(132.5-140.5]'] * 0.51 +

[PAS='(140.5-160.5]'] * 0.8 +

[PAS='(160.5-189.5]'] * 0.94 +

[PAS='(189.5-inf)'] * 1.53 +

[PAD='(53.5-60.5]'] * -0.14 +

[PAD='(60.5-70.5]'] * -0.16 +

[PAD='(70.5-75.5]'] * -0.05 +

[PAD='(80.5-90.5]'] * 0.06 +

[PAD='(90.5-inf)'] * 0.09

Time taken to build model: 1143.46 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	24743	87.0742 %
Incorrectly Classified Instances	3673	12.9258 %
Kappa statistic	0.726	
Mean absolute error	0.187	
Root mean squared error	0.3041	
Relative absolute error	39.8749 %	
Root relative squared error	62.8064 %	
Total Number of Instances	28416	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.887	0.157	0.904	0.887	0.896	0.942	Normal
	0.843	0.113	0.818	0.843	0.83	0.942	Doença
Weighted Avg.	0.871	0.14	0.872	0.871	0.871	0.942	

=== Confusion Matrix ===

a	b	<-- classified as
15751	1999	a = Normal
1674	8992	b = Doença

8.2. Modelos de Associação

8.2.1. Algoritmo PredictiveAPriori para o CS Arnaldo Sampaio

Prescrições - Diabetes Mellitus

1. Glibenclamida=yes Glibenclamida + Metformina=yes 15 ==> Metformina=yes 15 acc:(0.94105)

2. G-lancets No-dol=yes Metformina=yes 14 ==> Glucocard G sensor (50)=yes 14 acc:(0.93738)

3. Acetilcisteína=yes One Touch Ultra=yes 8 ==> Metformina=yes 8 acc:(0.89993)

4. Acarbose=yes Glibenclamida=yes 7 ==> Metformina=yes 7 acc:(0.88883)

5. Acarbose=yes Omeprazol=yes 7 ==> Metformina=yes 7 acc:(0.88883)

6. G-lancets No-dol=yes Metformina + Sitagliptina=yes 6 ==> Glucocard G sensor (50)=yes 6 acc:(0.87495)

7. Glibenclamida=yes Nimesulida=yes 6 ==> Metformina=yes 6 acc:(0.87495)

8. Ácido acetilsalicílico=yes Furosemida=yes 21 ==> Metformina=yes 19 acc:(0.86957)

9. Ácido acetilsalicílico=yes Amoxicilina + Ácido clavulânico=yes 20 ==> Metformina=yes 18 acc:(0.86364)

10. Acetilcisteína=yes Claritromicina=yes 12 ==> Metformina=yes 11 acc:(0.85724)

Prescrições - Diabetes Mellitus com Hipertensão

1. Paracetamol=yes Tiocolquicosido=yes 26 ==> Metformina=yes 26 acc:(0.96407)

2. Amoxicilina + Ácido clavulânico=yes Rosuvastatina=yes 16 ==> Metformina=yes 16 acc:(0.94431)

3. Diclofenac=yes Perindopril=yes 15 ==> Metformina=yes 15 acc:(0.94105)

4. Alopurinol=yes Lercanidipina=yes 14 ==> Metformina=yes 14 acc:(0.93738)

5. Azitromicina=yes Pantoprazol=yes 12 ==> Metformina=yes 12 acc:(0.92847)

6. Fluoxetina=yes Rosuvastatina=yes 11 ==> Metformina=yes 11 acc:(0.92298)

7. Acarbose=yes Carvedilol=yes 10 ==> Metformina=yes 10 acc:(0.91658)

8. Ácido acetilsalicílico=yes Valsartan + Hidroclorotiazida=yes 42 ==> Metformina=yes 39 acc:(0.90909)

9. Alprazolam=yes Lorazepam=yes 9 ==> Metformina=yes 9 acc:(0.90902)

10. Perindopril=yes Pitavastatina=yes 8 ==> Metformina=yes 8 acc:(0.89993)

Prescrições - Diabetes Mellitus Gestacional

Resulted on a Memory Error

Prescrições - Hipertensão

1. Alprazolam=yes Vacina contra a gripe=yes Varfarina=yes 9 ==> Paracetamol=yes 9 acc:(0.90902)

2. Claritromicina=yes Tansulosina=yes 8 ==> Vacina contra a gripe=yes 8 acc:(0.89993)

3. Amiodarona=yes Estriol=yes 6 ==> Paracetamol=yes 6 acc:(0.87495)

4. Azitromicina=yes Metamizol magnésico=yes 6 ==> Diclofenac=yes 6 acc:(0.87495)

5. Ácido fusídico=yes Mexazolam=yes 5 ==> Paracetamol=yes 5 acc:(0.8571)

6. Ambroxol=yes Amlodipina=yes 5 ==> Paracetamol=yes 5 acc:(0.8571)

7. Beta-histina=yes Irbesartan=yes 5 ==> Paracetamol=yes 5 acc:(0.8571)

8. Etofenamato=yes Trimetazidina=yes 11 ==> Vacina contra a gripe=yes 10 acc:(0.84624)

9. Aceclofenac=yes Mexazolam=yes 4 ==> Omeprazol=yes 4 acc:(0.8333)

10. Ácido fusídico=yes Varfarina=yes 4 ==> Paracetamol=yes 4 acc:(0.8333)

8.2.2. Algoritmo PredictiveAPriori para o CS Eiras

MCDT - Diabetes Mellitus

1. COLESTEROL TOTAL, S/L =yes 139 ==> TRIGLICÉRIDOS, S/U/L =yes 139 acc:(0.99178)

2. TRIGLICÉRIDOS, S/U/L =yes 139 ==> COLESTEROL TOTAL, S/L =yes 139 acc:(0.99178)

3. URINA, ANÁLISE SUMÁRIA (INCLUI ANÁLISE DO SEDIMENTO)=yes 117 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 117 acc:(0.99064)

4. COLESTEROL TOTAL, S/L =yes URINA, ANÁLISE SUMÁRIA (INCLUI ANÁLISE DO SEDIMENTO)=yes 112 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes TRIGLICÉRIDOS, S/U/L =yes 112 acc:(0.99031)

5. TRIGLICÉRIDOS, S/U/L =yes URINA, ANÁLISE SUMÁRIA (INCLUI ANÁLISE DO SEDIMENTO)=yes 112 ==> COLESTEROL TOTAL, S/L =yes GLUCOSE, DOSEAMENTO, S/U/L =yes 112 acc:(0.99031)

6. HEMOGRAMA COM FÓRMULA LEUCOCITÁRIA ...=yes HGB A1C=yes URINA, ANÁLISE SUMÁRIA (INCLUI ANÁLISE DO SEDIMENTO)=yes 96 ==> COLESTEROL TOTAL, S/L =yes 96 acc:(0.98901)

7. CREATININA, S/U =yes HEMOGRAMA COM FÓRMULA LEUCOCITÁRIA ...=yes HGB A1C=yes URINA, ANÁLISE SUMÁRIA (INCLUI ANÁLISE DO SEDIMENTO)=yes 94 ==> TRIGLICÉRIDOS, S/U/L =yes 94 acc:(0.98881)

8. HGB A1C=yes 142 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 141 acc:(0.98687)

9. COLESTEROL TOTAL, S/L =yes HGB A1C=yes 130 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes TRIGLICÉRIDOS, S/U/L =yes 129 acc:(0.98559)

10. ÁCIDO ÚRICO, S/U/L =yes 62 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 62 acc:(0.98386)

MCDT - Diabetes Mellitus com Hipertensão

1. URINA - EXAME DIRECTO, CULTURAL, IDENTIFICAÇÃO E TSA (UROCULTURA)=yes URINA, ANÁLISE SUMÁRIA (INCLUI ANÁLISE DO SEDIMENTO)=yes 299 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes HGB A1C=yes 299 acc:(0.99449)

2. COLESTEROL TOTAL, S/L =yes URINA - EXAME DIRECTO, CULTURAL, IDENTIFICAÇÃO E TSA (UROCULTURA)=yes URINA, ANÁLISE SUMÁRIA (INCLUI ANÁLISE DO SEDIMENTO)=yes 297 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes TRIGLICÉRIDOS, S/U/L =yes 297 acc:(0.99448)

3. HGB A1C=yes PSA TOTAL=yes 251 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 251 acc:(0.99414)

4. PSA TOTAL=yes TRIGLICÉRIDOS, S/U/L =yes 250 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 250 acc:(0.99413)

5. ÁCIDO ÚRICO, S/U/L =yes IONOGRAMA (NA, K, CL), S/U=yes TRIGLICÉRIDOS, S/U/L =yes 250 ==> COLESTEROL TOTAL, S/L =yes 250 acc:(0.99413)

6. COLESTEROL TOTAL, S/L =yes PSA TOTAL=yes 249 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes TRIGLICÉRIDOS, S/U/L =yes 249 acc:(0.99412)

7. CREATININA, S/U =yes PSA TOTAL=yes 248 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 248 acc:(0.99411)

8. FOSFATASE ALCALINA, S =yes HEMOGRAMA COM FÓRMULA LEUCOCITÁRIA=yes 233 ==> CREATININA, S/U =yes 233 acc:(0.99396)

9. CREATININA, S/U =yes FOSFATASE ALCALINA, S =yes TRIGLICÉRIDOS, S/U/L =yes 233 ==> COLESTEROL TOTAL, S/L =yes 233 acc:(0.99396)

10. ÁCIDO ÚRICO, S/U/L =yes FOSFATASE ALCALINA, S =yes 208 ==> CREATININA, S/U =yes 208 acc:(0.99362)

MCDT - Diabetes Mellitus Gestacional

1. GLUCOSE, DOSEAMENTO, S/U/L =yes 3 ==> HEMOGRAMA COM FÓRMULA LEUCOCITÁRIA=yes 3 acc:(0.79997)

2. HEMOGRAMA COM FÓRMULA LEUCOCITÁRIA=yes 3 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 3 acc:(0.79997)

3. ANTIGÉNIO HBS=yes 2 ==> EXAME CITOLÓGICO CERVICO-VAGINAL=yes GLUCOSE, DOSEAMENTO, S/U/L =yes 2 acc:(0.74998)

4. ANTIGÉNIO HBS=yes 2 ==> EXAME CITOLÓGICO CERVICO-VAGINAL=yes HEMOGRAMA COM FÓRMULA LEUCOCITÁRIA=yes 2 acc:(0.74998)

5. CREATININA, S/U =yes 2 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes HEMOGRAMA COM FÓRMULA LEUCOCITÁRIA=yes 2 acc:(0.74998)

6. EXAME CITOLÓGICO CERVICO-VAGINAL=yes 2 ==> ANTIGÉNIO HBS=yes GLUCOSE, DOSEAMENTO, S/U/L =yes 2 acc:(0.74998)

7. EXAME CITOLÓGICO CERVICO-VAGINAL=yes 2 ==> ANTIGÉNIO HBS=yes HEMOGRAMA COM FÓRMULA LEUCOCITÁRIA=yes 2 acc:(0.74998)

8. URINA - EXAME DIRECTO, CULTURAL, IDENTIFICAÇÃO E TSA (UROCULTURA)=yes 2 ==> ANTIGÉNIO HBS=yes 2 acc:(0.74998)

9. VDRL=yes 2 ==> ANTIGÉNIO HBS=yes 2 acc:(0.74998)

10. ANTIGÉNIO HBS=yes EXAME CITOLÓGICO CERVICO-VAGINAL=yes 2 ==> VDRL=yes 2 acc:(0.74998)

MCDT - Hipertensão

1. ÁCIDO ÚRICO, S/U/L =yes TIROXINA LIVRE (FT4), S =yes TRIGLICÉRIDOS, S/U/L =yes 293 ==> COLESTEROL TOTAL, S/L =yes 293 acc:(0.99446)

2. PSA TOTAL=yes URINA, ANÁLISE SUMÁRIA (INCLUI ANÁLISE DO SEDIMENTO)=yes 286 ==> COLESTEROL TOTAL, S/L =yes 286 acc:(0.99441)

3. ECOCARDIOGRAMA TRANSTORÁCICO BIDIMENSIONAL=yes TRIGLICÉRIDOS, S/U/L =yes 281 ==> COLESTEROL TOTAL, S/L =yes 281 acc:(0.99438)

4. ÁCIDO ÚRICO, S/U/L =yes HEMOGRAMA COM FÓRMULA LEUCOCITÁRIA=yes TIROXINA LIVRE (FT4), S =yes 273 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 273 acc:(0.99433)

5. ÁCIDO ÚRICO, S/U/L =yes ECOCARDIOGRAMA TRANSTORÁCICO BIDIMENSIONAL=yes 260 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 260 acc:(0.99422)

6. ECOCARDIOGRAMA TRANSTORÁCICO BIDIMENSIONAL=yes HEMOGRAMA COM FÓRMULA LEUCOCITÁRIA=yes TRIGLICÉRIDOS, S/U/L =yes 254 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 254 acc:(0.99417)

7. ÁCIDO ÚRICO, S/U/L =yes TIROXINA LIVRE (FT4), S =yes URINA, ANÁLISE SUMÁRIA (INCLUI ANÁLISE DO SEDIMENTO)=yes 240 ==> COLESTEROL TOTAL, S/L =yes 240 acc:(0.99403)

8. ÁCIDO ÚRICO, S/U/L =yes ECOCARDIOGRAMA TRANSTORÁCICO BIDIMENSIONAL=yes URINA, ANÁLISE SUMÁRIA (INCLUI ANÁLISE DO SEDIMENTO)=yes 225 ==> COLESTEROL TOTAL, S/L =yes 225 acc:(0.99386)

9. ÁCIDO ÚRICO, S/U/L =yes ECOCARDIOGRAMA TRANSTORÁCICO BIDIMENSIONAL=yes IONOGRAMA (NA, K, CL), S/U=yes 214 ==> CREATININA, S/U =yes 214 acc:(0.99371)

10. ECOCARDIOGRAMA TRANSTORÁCICO BIDIMENSIONAL=yes HEMOGRAMA COM FÓRMULA LEUCOCITÁRIA=yes IONOGRAMA (NA, K, CL), S/U=yes 212 ==> CREATININA, S/U =yes 212 acc:(0.99368)

Prescrições – Diabetes Mellitus

1. Tiocolquicosido=yes 10 ==> Diclofenac=yes 10 acc:(0.91658)

2. Fenofibrato=yes 8 ==> Metformina=yes 8 acc:(0.89993)

3. Ácido acetilsalicílico=yes Metformina + Sitagliptina=yes 7 ==> Metformina=yes 7 acc:(0.88883)

4. Ácido acetilsalicílico=yes Claritromicina=yes 6 ==> Ibuprofeno=yes 6 acc:(0.87495)

5. Ácido acetilsalicílico=yes One Touch Ultra=yes 12 ==> Metformina=yes 11 acc:(0.85724)

6. Espironolactona=yes 9 ==> Furosemida=yes 8 acc:(0.81825)

7. Aminofilina=yes 8 ==> Acetilcisteína=yes 7 acc:(0.80006)

8. Aminofilina=yes 8 ==> Furosemida=yes 7 acc:(0.80006)

9. Sertaconazol=yes 8 ==> Metformina=yes 7 acc:(0.80006)

10. Trazodona=yes 8 ==> Metformina=yes 7 acc:(0.80006)

Prescrições – Diabetes Mellitus com Hipertensão

1. Agulha B-D Micro-Fine+8mm=yes Sinvastatina=yes Vacina contra a gripe=yes 19 ==> Metformina=yes 19 acc:(0.95222)

2. Azitromicina=yes Varfarina=yes 18 ==> Paracetamol=yes 18 acc:(0.94985)

3. Sinvastatina=yes Vildagliptina=yes 17 ==> Metformina=yes 17 acc:(0.94723)

4. Acetilcisteína=yes Bisoprolol=yes 16 ==> Vacina contra a gripe=yes 16 acc:(0.94431)

5. Azitromicina=yes Claritromicina=yes 14 ==> Paracetamol=yes 14 acc:(0.93738)

6. Amlodipina=yes Amoxicilina=yes 13 ==> Paracetamol=yes 13 acc:(0.93322)

7. Ácido fusídico=yes Azitromicina=yes 12 ==> Paracetamol=yes 12 acc:(0.92847)

8. Acetilcisteína=yes Clopidogrel=yes Pantoprazol=yes 10 ==> Vacina contra a gripe=yes 10 acc:(0.91658)

9. Ácido fusídico=yes Estriol=yes 9 ==> Paracetamol=yes 9 acc:(0.90902)

10. Acetilcisteína=yes Ácido fusídico=yes Varfarina=yes 9 ==> Paracetamol=yes 9 acc:(0.90902)

Prescrições – Diabetes Mellitus Gestacional

No large itemsets and rules found!

Prescrições - Hipertensão

1. Amoxicilina + Ácido clavulânico=yes Etoricoxib=yes Glucosamina=yes 16 ==> Paracetamol=yes 16 acc:(0.94431)

2. Amoxicilina + Ácido clavulânico=yes Desloratadina=yes Glucosamina=yes 15 ==> Paracetamol=yes 15 acc:(0.94105)

3. Acetilcisteína=yes Atorvastatina=yes Azitromicina=yes 14 ==> Paracetamol=yes 14 acc:(0.93738)

4. Acetilcisteína=yes Alprazolam=yes Bioflavonóides=yes 13 ==> Paracetamol=yes 13 acc:(0.93322)

5. Acetilcisteína=yes Amoxicilina + Ácido clavulânico=yes Metamizol magnésico=yes 13 ==> Paracetamol=yes 13 acc:(0.93322)

6. Amoxicilina + Ácido clavulânico=yes Flupirtina=yes 12 ==> Paracetamol=yes 12 acc:(0.92847)

7. Amoxicilina + Ácido clavulânico=yes Etofenamato=yes 11 ==> Paracetamol=yes 11 acc:(0.92298)

8. Amoxicilina + Ácido clavulânico=yes Etoricoxib=yes Ibuprofeno=yes 10 ==> Paracetamol=yes 10 acc:(0.91658)

9. Acetilcisteína=yes Sertralina=yes 18 ==> Paracetamol=yes 17 acc:(0.90014)

10. Acetilcisteína=yes Amoxicilina + Ácido clavulânico=yes Mometasona=yes 8 ==> Desloratadina=yes 8 acc:(0.89993)

8.2.3. Algoritmo PredictiveAPriori para o CS Fundão

Prescrições – Diabetes Mellitus

1. Digoxina=yes 8 ==> Furosemida=yes 8 acc:(0.89993)

2. Fluconazol=yes Paracetamol=yes 6 ==> Metformina=yes 6 acc:(0.87495)

3. Ácido acetilsalicílico=yes Lercanidipina=yes 5 ==> Metformina=yes 5 acc:(0.8571)

4. Amoxicilina + Ácido clavulânico=yes Paracetamol + Codeína=yes 5 ==> Ibuprofeno=yes 5 acc:(0.8571)

5. Tramadol=yes 10 ==> Metformina=yes 9 acc:(0.83341)

6. Fluconazol=yes Picetoprofeno=yes 4 ==> Metformina=yes 4 acc:(0.8333)

7. Aceclofenac=yes 8 ==> Metformina=yes 7 acc:(0.80006)

8. Cefixima=yes 8 ==> Ibuprofeno=yes 7 acc:(0.80006)

9. Escina=yes 8 ==> Metformina=yes 7 acc:(0.80006)

10. Fluconazol=yes 18 ==> Metformina=yes 15 acc:(0.8)

Prescrições – Diabetes Mellitus com Hipertensão

1. Paracetamol + Tiocolquicosido=yes 12 ==> Metformina=yes 12 acc:(0.92847)

2. Atorvastatina=yes Clopidogrel=yes 8 ==> Furosemida=yes 8 acc:(0.89993)

3. Clopidogrel=yes Metformina + Sitagliptina=yes 8 ==> Furosemida=yes 8 acc:(0.89993)

4. Diclofenac=yes Sinvastatina=yes 15 ==> Metformina=yes 14 acc:(0.88247)

5. Accu-Chek Aviva 50 Tiras Teste=yes Amoxicilina + Ácido clavulânico=yes 6 ==> Metformina=yes 6 acc:(0.87495)

6. Accu-Chek Aviva 50 Tiras Teste=yes Lansoprazol=yes 6 ==> Metformina=yes 6 acc:(0.87495)

7. Bioflavonóides=yes Sinvastatina=yes 6 ==> Metformina=yes 6 acc:(0.87495)

8. Candesartan=yes 10 ==> Metformina=yes 9 acc:(0.83341)

9. Vildagliptina=yes 10 ==> Metformina=yes 9 acc:(0.83341)

10. Accu-Chek Aviva 50 Tiras Teste=yes Beta-histina=yes 4 ==> Metformina=yes 4 acc:(0.8333)

Prescrições – Diabetes Mellitus Gestacional

No large itemsets and rules found!

Prescrições - Hipertensão

1. Cetoprofeno=yes Metamizol magnésico=yes 8 ==> Tiocolquicosido=yes 8 acc:(0.89993)

2. Ácido acetilsalicílico=yes Cetoprofeno=yes 7 ==> Tiocolquicosido=yes 7 acc:(0.88883)

3. Amoxicilina + Ácido clavulânico=yes Cefixima=yes 6 ==> Desloratadina=yes 6 acc:(0.87495)

4. Claritromicina=yes Furoato de fluticasona=yes 6 ==> Ibuprofeno=yes 6 acc:(0.87495)

5. Ácido acetilsalicílico=yes Ácido alendrónico + Colecalciferol=yes 5 ==> Sinvastatina=yes 5 acc:(0.8571)

6. Ácido alendrónico + Colecalciferol=yes Glucosamina=yes 5 ==> Sinvastatina=yes 5 acc:(0.8571)

7. Amoxicilina + Ácido clavulânico=yes Diclofenac=yes 5 ==> Ibuprofeno=yes 5 acc:(0.8571)

8. Acetilcisteína=yes Amlodipina + Valsartan=yes 4 ==> Desloratadina=yes 4 acc:(0.8333)

9. Acetilcisteína=yes Ebastina=yes 4 ==> Ibuprofeno=yes 4 acc:(0.8333)

10. Cetoprofeno=yes 23 ==> Tiocolquicosido=yes 19 acc:(0.8)

8.2.4. Algoritmo PredictiveAPriori para o CS Tábua

MCDT – Diabetes Mellitus

1. COLESTEROL HDL=yes 119 ==> COLESTEROL TOTAL, S/L =yes 119 acc:(0.99077)

2. COLESTEROL TOTAL, S/L =yes 119 ==> COLESTEROL HDL=yes 119 acc:(0.99077)

3. TRIGLICÉRIDOS, S/U/L =yes 118 ==> COLESTEROL HDL=yes 118 acc:(0.9907)

4. MICROALBUMINURIA EM URINA 24H=yes 100 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes HGB A1C=yes 100 acc:(0.98938)

5. COLESTEROL HDL=yes ECG=yes 86 ==> CREATININA, S/U =yes 86 acc:(0.98793)

6. HGB A1C=yes 129 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 128 acc:(0.98548)

7. COLESTEROL HDL=yes GAMA GT=yes 69 ==> CREATININA, S/U =yes 69 acc:(0.98534)

8. ÁCIDO ÚRICO, S/U/L =yes 65 ==> CREATININA, S/U =yes 65 acc:(0.98454)

9. COLESTEROL HDL=yes 119 ==> COLESTEROL TOTAL, S/L =yes TRIGLICÉRIDOS, S/U/L =yes 118 acc:(0.9842)

10. COLESTEROL HDL=yes 119 ==> COLESTEROL TOTAL, S/L =yes GLUCOSE, DOSEAMENTO, S/U/L =yes 118 acc:(0.9842)

MCDT – Diabetes Mellitus com Hipertensão

1. ÁCIDO ÚRICO, S/U/L =yes HEMOGRAMA COM FÓRMULA LEUCOCITÁRIA ...=yes HGB A1C=yes 296 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 296 acc:(0.99447)

2. ÁCIDO ÚRICO, S/U/L =yes GAMA GT=yes 289 ==> COLESTEROL TOTAL, S/L =yes GLUCOSE, DOSEAMENTO, S/U/L =yes 289 acc:(0.99443)

3. ÁCIDO ÚRICO, S/U/L =yes TGO=yes 278 ==> COLESTEROL TOTAL, S/L =yes 278 acc:(0.99436)

4. UREIA, S/U =yes 273 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 273 acc:(0.99433)

5. ÁCIDO ÚRICO, S/U/L =yes ECG=yes 273 ==> COLESTEROL TOTAL, S/L =yes GLUCOSE, DOSEAMENTO, S/U/L =yes 273 acc:(0.99433)

6. ÁCIDO ÚRICO, S/U/L =yes IONOGRAMA (NA, K, CL), S/U=yes 241 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 241 acc:(0.99404)

7. BILIRRUBINA TOTAL E DIRECTA, S/L =yes 228 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 228 acc:(0.9939)

8. ÁCIDO ÚRICO, S/U/L =yes TGP=yes 219 ==> COLESTEROL TOTAL, S/L =yes 219 acc:(0.99378)

9. ÁCIDO ÚRICO, S/U/L =yes URINA, ANÁLISE SUMÁRIA (INCLUI ANÁLISE DO SEDIMENTO)=yes 204 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 204 acc:(0.99355)

10. ÁCIDO ÚRICO, S/U/L =yes BILIRRUBINA TOTAL E DIRECTA, S/L =yes 203 ==> COLESTEROL TOTAL, S/L =yes 203 acc:(0.99353)

MCDT - Hipertensão

1. COLESTEROL HDL=yes FOSFATASE ALCALINA, S =yes TGO=yes 299 ==> COLESTEROL TOTAL, S/L =yes 299 acc:(0.99449)

2. FOSFATASE ALCALINA, S =yes TRIGLICÉRIDOS, S/U/L =yes 298 ==> COLESTEROL TOTAL, S/L =yes 298 acc:(0.99448)

3. BILIRRUBINA TOTAL E DIRECTA, S/L =yes UREIA, S/U =yes 297 ==> CREATININA, S/U =yes TGO=yes 297 acc:(0.99448)

4. ÁCIDO ÚRICO, S/U/L =yes COLESTEROL HDL=yes PSA TOTAL=yes 283 ==> GLUCOSE, DOSEAMENTO, S/U/L =yes 283 acc:(0.99439)

5. COLESTEROL HDL=yes FOSFATASE ALCALINA, S =yes TGP=yes 282 ==> COLESTEROL TOTAL, S/L =yes 282 acc:(0.99439)

6. MICROALBUMINURIA EM URINA 24H=yes TGP=yes 260 ==> TGO=yes 260 acc:(0.99422)

7. ÁCIDO ÚRICO, S/U/L =yes COLESTEROL HDL=yes TSH=yes 228 ==> COLESTEROL TOTAL, S/L =yes 228 acc:(0.9939)

8. BILIRRUBINA TOTAL E DIRECTA, S/L =yes ECG=yes 227 ==> GAMA GT=yes 227 acc:(0.99388)

9. FOSFATASE ALCALINA, S =yes VS=yes 218 ==> GAMA GT=yes 218 acc:(0.99376)

10. BILIRRUBINA TOTAL E DIRECTA, S/L =yes VS=yes 210 ==> GAMA GT=yes 210 acc:(0.99365)

Prescrições – Diabetes Mellitus

1. Sertaconazol=yes 11 ==> Metformina=yes 11 acc:(0.92298)

2. Amoxicilina + Ácido clavulânico=yes Desloratadina=yes 10 ==> Metformina=yes 10 acc:(0.91658)

3. Amoxicilina + Ácido clavulânico=yes Ibuprofeno=yes 10 ==> Metformina=yes 10 acc:(0.91658)

4. Azitromicina=yes 7 ==> Paracetamol=yes 7 acc:(0.88883)

5. Claritromicina=yes 7 ==> Metformina=yes 7 acc:(0.88883)

6. Amoxicilina + Ácido clavulânico=yes Domperidona=yes 6 ==> Paracetamol=yes 6 acc:(0.87495)

7. Acetilcisteína=yes Paracetamol=yes 12 ==> Metformina=yes 11 acc:(0.85724)

8. Bioflavonóides=yes 5 ==> Metformina=yes 5 acc:(0.8571)

9. Amoxicilina + Ácido clavulânico=yes Naproxeno=yes 5 ==> Metformina=yes 5 acc:(0.8571)

10. Cetirizina=yes 10 ==> Metformina=yes 9 acc:(0.83341)

Prescrições – Diabetes Mellitus com Hipertensão

1. Cetirizina=yes Metamizol magnésico=yes 19 ==> Paracetamol=yes 19 acc:(0.95222)

2. Desloratadina=yes Tramadol=yes 13 ==> Ácido acetilsalicílico=yes 13 acc:(0.93322)

3. Desloratadina=yes Insulina aspártico (solúvel + protamina)=yes 12 ==> Vacina contra a gripe=yes 12 acc:(0.92847)

4. Aminofilina=yes Insulina humana=yes 11 ==> Furosemida=yes 11 acc:(0.92298)

5. Acarbose=yes Omeprazol=yes 9 ==> Ácido acetilsalicílico=yes 9 acc:(0.90902)

6. Acetilcisteína=yes Aminofilina=yes 19 ==> Furosemida=yes 18 acc:(0.90491)

7. Acetilcisteína=yes Precision Xtra Plus Test Strips=yes 8 ==> Paracetamol=yes 8 acc:(0.89993)

8. Clopidogrel=yes Digoxina=yes 7 ==> Furosemida=yes 7 acc:(0.88883)

9. Acetilcisteína=yes Amoxicilina=yes 23 ==> Paracetamol=yes 21 acc:(0.88)

10. Aminofilina=yes Metformina=yes 22 ==> Furosemida=yes 20 acc:(0.875)

Prescrições - Hipertensão

1. Lactulose=yes Picetoprofeno=yes 19 ==> Paracetamol=yes 19 acc:(0.95222)

2. Acetilcisteína=yes Sucralfato=yes 16 ==> Paracetamol=yes 16 acc:(0.94431)

3. Beta-histina=yes Diazepam=yes 16 ==> Paracetamol=yes 16 acc:(0.94431)

4. Azitromicina=yes Oxazepam=yes 14 ==> Paracetamol=yes 14 acc:(0.93738)

5. Lactulose=yes Varfarina=yes 11 ==> Paracetamol=yes 11 acc:(0.92298)

6. Acetilcisteína=yes Ácido alendrónico + Colecalciferol=yes 10 ==> Paracetamol=yes 10 acc:(0.91658)

7. Acetilcisteína=yes Lactulose=yes 21 ==> Paracetamol=yes 20 acc:(0.91321)

8. Beta-histina=yes Lactulose=yes 9 ==> Paracetamol=yes 9 acc:(0.90902)

9. Acetilcisteína=yes Tramadol=yes 18 ==> Paracetamol=yes 17 acc:(0.90014)

10. Acetilcisteína=yes Pregabalina=yes 8 ==> Paracetamol=yes 8 acc:(0.89993)