U.PORTO

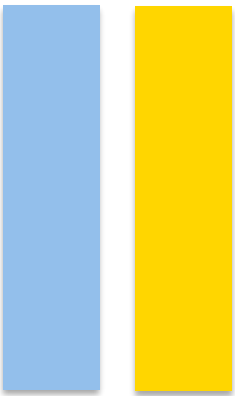
CICLO DE ESTUDOS
MESTRADO EM INFORMÁTICA MÉDICA

Previsão de número de dias de internamento em doentes diabéticos – Uma abordagem de Machine Learning

Florbela Santos Nunes

M

2020



SEDE ADMINISTRATIVA

FACULDADE DE **MEDICINA**

FACULDADE DE **CIÊNCIAS**

13ª ed

MIM

Previsão de número de dias de internamento em doentes diabéticos – Uma abordagem de Machine Learning

Florbela Nunes

MESTRADO EM
INFORMÁTICA MÉDICA
2º CICLO DE ESTUDOS

OUT|2020 (30 de outubro de 2020)

13ª ed

MIM

Previsão de número de dias de internamento em doentes diabéticos – Uma abordagem de Machine Learning

Florbela Nunes

MESTRADO EM
INFORMÁTICA MÉDICA
2º CICLO DE ESTUDOS

ORIENTADORES:

Inês Dutra, Professora Doutora Auxiliar Faculdade de Ciências da Univ. do Porto

Pedro Brandão, Professor Doutor Auxiliar Faculdade de Ciências da Univ. do Porto; Instituto de Telecomunicações

OUT|2020 (30 de outubro de 2020)

Agradecimentos

"Tudo o que um sonho precisa para ser realizado
é alguém que acredite que ele possa ser realizado"
Roberto Shinyashiki.

Aos orientadores deste trabalho, Professora Doutora Inês de Castro Dutra e ao Professor Doutor Pedro Brandão, pelos esclarecimentos, críticas construtivas, correções e sugestões. Agradeço a confiança que depositaram em mim, a constante disponibilidade demonstrada, os incentivos e o fundamental apoio.

Quero ainda deixar o meu agradecimento à minha família que acreditaram em mim desde o primeiro momento. Obrigada pelas palavras de incentivo, por todo o carinho e por serem os alicerces da minha vida.

A ti, Marcelo, por fazeres com que cada desafio seja um pouco mais fácil de ultrapassar. Por compreenderes cada ausência, por acreditares em mim e por todas as palavras de motivação.

Sumário

A diabetes é uma enfermidade crónica e progressiva, sendo considerada um problema crescente de saúde pública que afeta milhões de pessoas a nível mundial. Sabe-se, por estudos realizados, que no geral o risco de morte em pessoas com diabetes é quase o dobro em relação a pessoas da mesma idade que não tenham diabetes e que pessoas com a doença têm maior risco de serem hospitalizadas em relação a pessoas que não sejam diabéticas. Sendo que nos pacientes com a doença as estadias são mais prolongadas e acarretam mais despesas. A duração da estadia explica 85 a 90% da variação nos custos hospitalares entre pacientes.

Dado que os hospitais têm leitos e capacidade financeira limitados para receber pacientes internados, é evidente ser necessário tomar medidas de prevenção e controlo do prolongamento hospitalar destes pacientes. O principal objetivo deste trabalho é a utilização de técnicas de *data mining* para extrair informações úteis relativamente aos internamentos hospitalares de pacientes diabéticos e obter *insights* sobre a previsão do número de dias de internamento e sobre os principais fatores e elementos que afetam a duração da hospitalização do paciente diabético.

Utilizou-se as técnicas de KDD- *Knowledge Discovery from Data* ao longo do trabalho com o propósito de extrair conhecimento útil à cerca da amostra em estudo- 101766 episódios diferentes em que a diabetes foi inserida como diagnóstico principal ou secundário. A aplicação do processo de *data mining* compreendeu diferentes técnicas da área da estatística e da aprendizagem automática para identificar fatores de risco e prever o prolongamento do internamento hospitalar. Num primeiro momento foi realizado um estudo estatístico básico da amostra e selecionadas as variáveis a incluir no estudo para realizar a previsão. De um conjunto com um total de 50 variáveis, chegou-se a 15 variáveis selecionadas. Posteriormente, foram exploradas regras de associação para perceber quais os antecedentes com maior influência, cujo consequente seja o internamento curto/longo. Foi ainda explorado o algoritmo Random Forest para estabelecer a previsão do internamento hospitalar como uma variável binomial, tendo-se conseguido obter uma *accuracy* de 81%. Por fim, implementou-se o algoritmo obtido numa interface com o utilizador, de forma a ser potencialmente utilizado pelos auxiliares de saúde como uma ferramenta de

auxílio no momento do primeiro contacto com o paciente no internamento hospitalar.

Palavras-Chave: *Data mining,* KDD- *Knowledge Discovery from Data,* Regras de Associação*, Random Forest*

Abstract

Diabetes is a chronic and progressive disease, being considered a growing public health problem that affects millions of people worldwide. It is known from studies that in general the risk of death in people with diabetes is almost double than of people of the same age who do not have diabetes and that people with the disease have a higher risk of being hospitalized compared to people who are not diabetic. Since in patients with the disease, stays are longer and cost more. The length of stay explains 85 to 90% of the variation in hospital costs between patients.

Given that hospitals have limited beds and financial capacity to receive inpatients, it is evident that it is necessary to take measures to prevent and control the hospital extension of these patients. The main objective of this work is to use data mining techniques to extract useful information regarding the hospitalizations of diabetic patients and to gain insights on the forecast of the number of days of hospitalization and on the main factors and elements that affect the duration of hospitalization diabetic patient.

The KDD-Knowledge Discovery from Data techniques were used throughout the work with the purpose of extracting useful knowledge about the sample under study- 101766 different episodes in which diabetes was inserted as the main or secondary diagnosis. The application of the data mining process included different techniques in the field of statistics and machine learning to identify risk factors and predict the length of hospital stay. At first, a basic statistical study of the sample was carried out and the variables to be included in the study were selected to make the forecast. From a set with a total of 50 variables, 15 variables were selected. Subsequently, association rules were explored to understand which antecedents had the greatest influence, the consequence of which was short / long hospitalization. The Random Forest algorithm was also explored to establish the forecast of hospital stay as a binomial variable, with an accuracy of 81%. Finally, the algorithm obtained in an interface with the user was implemented, in order to be potentially used by health assistants as an aid tool when first contacting the patient during hospitalization.

Keywords: Data mining, KDD-Knowledge Discovery from Data, Association Rules, Random Forest

Índice

Índice de Figuras

Índice de Tabelas

Acrónimos

A1c	Hemoglobina glicada
AUC	*Area Under Curve*
AVC	Acidente Vascular Cerebral
DAC	Doença Arterial Coronária
EAM	Enfarte Agudo do Miocárdio
IMC	Índice de massa corporal
KDD	*Knowledge-Discovery in Databases*
MBA	*Market Basket Analysis*
MTL	Aprendizado de Múltiplas Tarefas
OCDE	Organização para a Cooperação e Desenvolvimento Econômico
OMS	Organização Mundial de Saúde
RF	*Random Forest*
ANN	*Artificial neural network*
ROC	*Receiver Operating Characteristic*
SVM	*Support Vector Machine*

1. Introdução / Motivação

A doença da diabetes, também designada por Diabetes Mellitus, é uma enfermidade cronica e progressiva que pode ser encontrada em quase todas as populações do mundo, evidências epidemiológicas estimam que sem programas eficazes de prevenção e controle, a diabetes continuará a aumentar globalmente (International Diabetes Federation, 2003).

A OMS – Organização Mundial da Saúde, reporta que a doença da diabetes pode vir a ser a sétima causa de morte até 2030 (World Health Organization, 2016). Em Portugal, a OMS em 2016, declarou que a doença mata 12 pessoas por dia. Já o Relatório Anual do Observatório Nacional da Diabetes de 2014 assegurou que a prevalência estimada da diabetes na população com idades compreendidas entre os 20 e os 79 anos (7,7 milhões de indivíduos) foi de 13,1%, isto é, mais de um milhão de portugueses neste grupo etário tem diabetes. O rápido aumento na prevalência da diabetes, impulsionado pelo aumento da prevalência da obesidade e pelo envelhecimento da população, faz com que a doença da diabetes seja um problema sério de saúde pública que tem implicações para os indivíduos, comunidade e serviços de saúde e humanos (Comino et al., 2015). O presidente da Secção Regional Norte da Ordem dos Médicos, Miguel Guimarães numa entrevista dada em 2016, destacou os elevados custos que a diabetes acarreta para a sociedade, em termos de internamentos e consumo de medicamentos:

"A diabetes tem um peso muito grande naquilo que é o orçamento de estado para a Saúde e, por outro lado, tem consequências nocivas para os doentes"
(Aragão, 2016)

Estudo feito nos Estados Unidos relata que quase um terço dos pacientes com diabetes podem necessitar de duas ou mais hospitalizações anualmente. Sendo que nos pacientes com a doença as estadias são mais prolongadas e

acarretam mais despesas (Knecht et al., 2006). A duração da estadia explica 85 a 90% da variação nos custos hospitalares entre pacientes (Gentimis, Alnaser, Durante, Cook, & Steele, 2018). Dado que os hospitais têm leitos e capacidade financeira limitados para receber pacientes internados, é extremamente importante encontrar maneiras de reduzir e prever os custos com a saúde. Nas últimas décadas, os hospitais conseguiram acumular um grande volume de dados que permitem avaliar e comparar o desempenho clínico. As técnicas de *data mining* oferecem um primeiro passo e uma ajuda para extrair informações úteis desses dados e obter *insights* sobre a previsão do número de dias de internamento e sobre os principais fatores e elementos que afetam a duração de uma hospitalização do paciente (Livieris, Kotsilieris, Dimopoulos, & Pintelas, 2018).

Na Figura 1 estão apresentadas algumas das vantagens da precisão do prolongamento hospitalar. Existem benefícios não só para os pacientes, na sua generalidade, por se traduzir em cuidados de saúde personalizados, mas também ao hospital por ser num novo indicar auxiliar nas tomadas de decisão hospitalares.

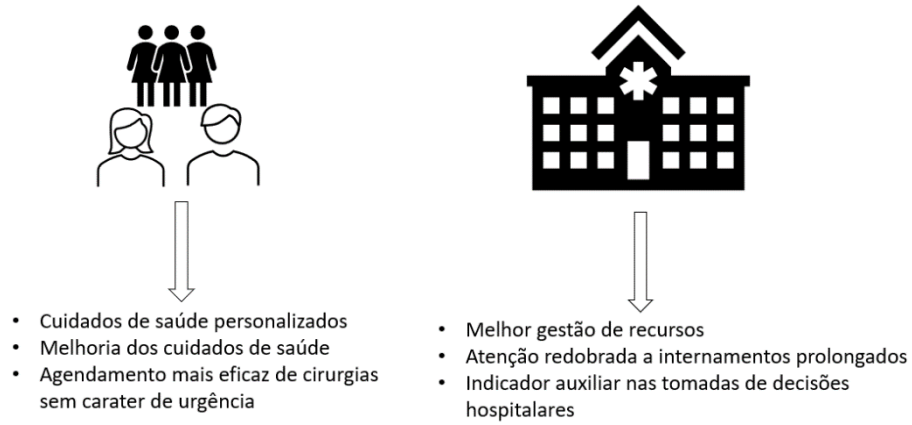


Figura 1 - Vantagens para o paciente e para as unidades hospitalares da previsão do número de dias de internamento.

1.1 Objetivos

Este projeto tem por objetivo, num primeiro momento, a discussão e análise do *data set* em estudo. Estudar os dados quanto às informações estatísticas dos indivíduos: sexo, idade, etnia. Como quanto às características dos internamentos: tipo e fonte de admissão, diagnósticos, especialidade médica, resultado de teste de A1c e de glicose e a readmissão hospitalar.

É também objetivo deste trabalho estudar o número de dias de internamento como um problema de classificação. Tentar perceber quais as variáveis mais influentes no momento da previsão do prolongamento hospitalar e que fatores, quando presentes, podem influenciar com maior ou menor probabilidade um internamento hospitalar mais prolongado ou reduzido. Posteriormente, pretende-se chegar a um modelo de previsão do número de dias de internamento hospitalar a partir de variáveis passiveis de estarem disponíveis logo no primeiro momento do processo de internamento hospitalar. Após a obtenção do modelo de previsão, pretende-se desenvolver uma aplicação de interface com o utilizador que possibilite a previsão do prolongamento hospitalar a partir da introdução dos *inputs* necessários para essa previsão.

1.2 Pergunta de Investigação

A pergunta de investigação foi elaborada com base na análise da metodologia PICO, que se descreve em seguida:

- **P**opulação: Doentes diabéticos.
- **I**ntervenção: Aplicação de técnicas de *data mining* em dados de internamento
- **C**ontrolo: -
- ***O****utcome*: Previsão do número de dias de internamento

Dessa forma, os princípios sobre as quais se apoia o resultado, abarcam a seguinte pergunta de investigação:

"É possível prever o número de dias de internamento de doentes diabéticos a partir de técnicas de *data mining*/ machine learning?"

1.3 Estrutura da Dissertação

De forma esquemática a dissertação está organizada da forma como se representa na Figura 2.

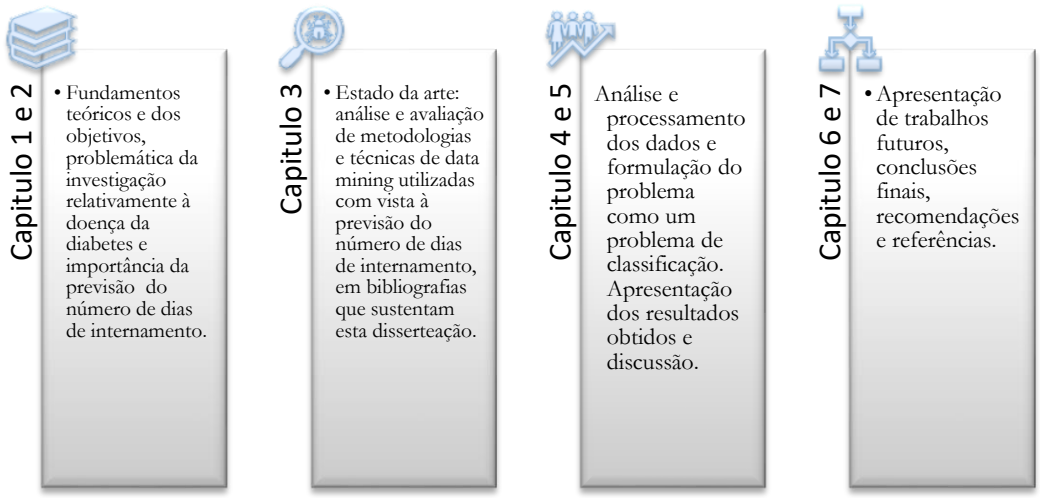


Figura 2- Esquematização da estrutura da dissertação

2. Fundamentos e Terminologia

Neste capítulo será apresentado o background para o estudo da permanência, readmissões e mortalidade hospitalar, em doentes diabéticos, com a intenção de detalhar conceitos importantes e auxiliar a compreensão da presente dissertação. Será feita uma primeira abordagem à doença da diabetes, desde a definição da doença a uma quantificação geral da doença em Portugal e no Mundo. Posteriormente, será apresentada uma abordagem aos conceitos relacionados ao tratamento e manipulação de dados, tendo como objetivo a clarificação dos conceitos de *data mining*, bem como a usabilidade e performance de diferentes técnicas de machine learning.

2.1 Diabetes Mellitus

A doença da diabetes, também designada por Diabetes Mellitus, é uma enfermidade cronica e progressiva que pode ser encontrada em quase todas as populações do mundo. Evidências epidemiológicas estimam que sem programas eficazes de prevenção e controle, a diabetes continuará a aumentar globalmente (International Diabetes Federation, 2003).

A diabetes é caracterizada por anómalas concentrações de glucose no sangue – principal fonte de energia do ser humano, podendo ser resultante devido a deficiência na secreção de insulina, da ação da insulina, ou mesmo da combinação de ambos os fatores, sendo esta última a classe mais frequente

(American Diabetes Association, 2014). Este grupo de condições pode ser subdividido em 4 tipos clinicamente distintos:

Tipo 1: É caracterizada por uma falha na sua totalidade da produção de insulina, devido à destruição autoimune de células beta do pâncreas, que impede a segregação desta hormona. Este tipo, é por isso, também conhecido como diabetes insulinodependentes, pois exige a administração de insulina (Observatório Nacional da Diabetes, 2010). Diz respeito a cerca de 5% a 10% de todos os casos de diabetes e atinge maioritariamente crianças ou jovens. Até ao momento são desconhecidos métodos de prevenção da diabetes tipo 1, uma vez que os seus fatores de risco passam por causas autoimunes, genéticas e ambientais (Deshpande, Harris-Hayes, & Schootman, 2008).

Tipo 2: Esta é a categoria da diabetes mais frequente, prevendo-se que englobe cerca de 90% a 95% de todos os casos da doença (Deshpande et al., 2008). É causada pela combinação da resistência à ação da insulina e de uma inadequada resposta secretora de insulina compensatória. Indivíduos com esta patologia podem estar longos períodos de forma assintomática, sendo que durante este intervalo podem apresentar irregularidade no metabolismo de carboidratos pela medição de glicose plasmática em jejum (American Diabetes Association, 2014).

Diabetes Gestacionais: Ocorre durante a gravidez e afeta 1 em cada 20 mulheres (Ali & Dornhorst, 2011). As mulheres com esta patologia devem fazer um controlo rigoroso de forma a prevenir complicações no nascimento e desenvolvimento infantil. Estima-se que mulheres com diabetes gestacionais têm cerca de 20% a 50% de risco de desenvolver diabetes tipo 2 ao longo da vida (Deshpande et al., 2008).

Outros Tipos da Diabetes: Existem outros tipos de diabetes, menos comuns, que estão relacionadas com defeitos monogenéticos das funções nas células beta do pâncreas. São conhecidos como diabetes de maturidade dos jovens (MODY). É caracterizado por uma reduzida produção de insulina com defeitos mínimos ou inexistentes na ação da insulina (American Diabetes Association, 2014).

A diabetes pode aumentar o risco de morte prematura e levar a complicações em diferentes partes do corpo, nomeadamente pé, rins e olhos

podendo as complicações serem designadas por vasculares e não vasculares (World Health Organization, 2016). As complicações vasculares incluem problemas como retinopatia, nefropatia e neuropatia – microvasculares e doenças arteriais coronárias, doenças periféricas, vasculares e cerebrovasculares – complicações macrovasculares, que podem levar ao Enfarte Agudo do Miocárdio (EAM) e ao Acidente Vascular Cerebral (AVC) (IJsselmuiden & Faden, 1992). Complicações não vasculares incluem problemas como gastroparesia, disfunção sexual e alterações na pele (Tripathi & Srivastava, 2006).

2.1.1 A Diabetes em Portugal e no Mundo

Diabetes Mellitus é um problema crescente de saúde pública que afeta adversamente a vida de milhões de pessoas em todos o mundo. Esta doença requer assistência médica continuada e autogestão do paciente, a fim de evitar complicações a longo prazo. Estas complicações podem resultar num aumento significativo da carga económica total da doença (Menzin et al., 2010). Em 2002, a diabetes foi classificada como a sexta principal causa de morte. No geral o risco de morte em pessoas com diabetes é quase o dobro em relação a pessoas da mesma idade que não tenham diabetes (Deshpande et al., 2008).

Em 2019, a federação de diabetes internacional estimou que nesse ano existissem cerca de 350 milhões de pessoas em todo o mundo com diabetes e previu que até 2030 este número aumente para 417 milhões e para 486 milhões até 2045. Os gastos devido a diabetes têm um impacto significativo nos orçamentos de saúde em todo o mundo. Se a previsão para o ano de 2045 for cumprida, o gasto total com a doença nesse ano atingirá os 845 biliões de dólares internacionais (Atlas, 2019).

Na Figura 3 é possível verificar a taxa de prevalência da diabetes de adultos com idade compreendida entre 20 e 79 anos, cujo tipo de diabetes é 1 ou 2, em 2015 (*Health at a Glance 2017*, 2017). Constata-se que Portugal tinha em 2015 uma prevalência de 9.9% em adultos, estando acima da média da OCDE35.

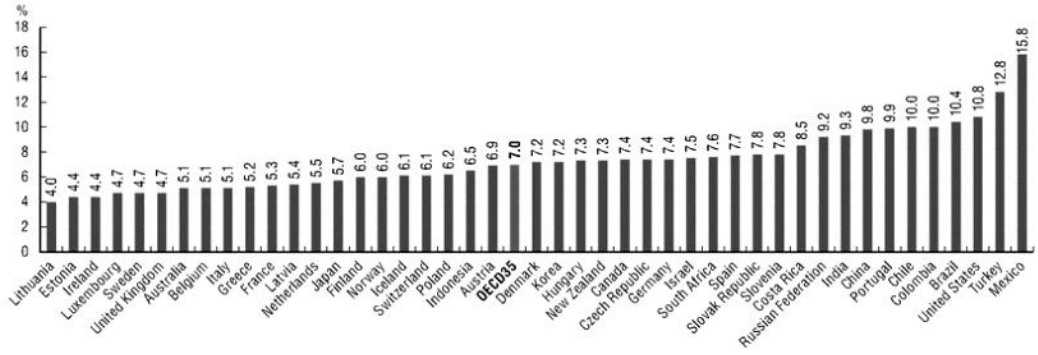
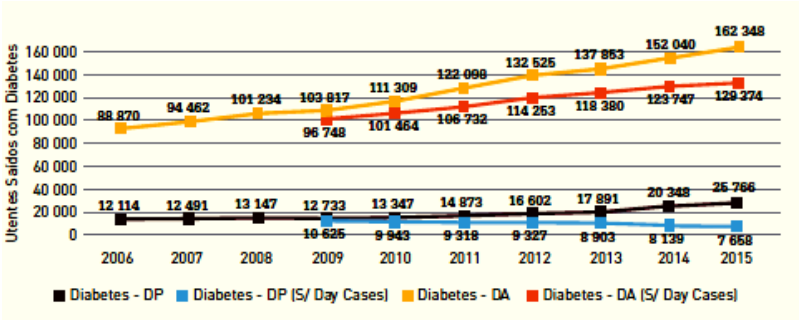


Figura 3 - Taxa de prevalência de diabetes de adultos por país – 2015 (*Health at a Glance 2017*, 2017).

2.1.2 Hospitalização - Diabetes

Pessoas com a doença da diabetes têm maior risco de serem hospitalizadas em relação a pessoas que não sejam diabéticas. Um estudo feito nos Estados Unidos relata que quase um terço dos pacientes com diabetes podem necessitar de duas ou mais hospitalizações anualmente. Sendo que nos pacientes com a doença as estadias são mais prolongadas e acarretam mais despesas. (Knecht et al., 2006).

Em Portugal, o número de episódios nos hospitais do Serviço Nacional de Saúde em que a doença da diabetes é o diagnóstico principal têm vindo a diminuir até ao ano de 2015, excluindo os *Day Case* – internamento com uma duração inferior a 24 horas. Por outro lado, o número total de internados em que a doença da diabetes surge como diagnóstico secundário tem aumentado de forma acentuada (aumentou 82.7% entre 2006 e 2015). (Figura 4) (Observatório da diabetes, 2016).



Legenda:
DA- Diagnóstico Associados

DP– Diagnóstico Principal

Day Case- Internamento com uma duração inferior a 24 horas

Figura 4- Utentes saídos dos Internamentos com Diabetes dos Hospitais do SNS – 2015 (Observatório da diabetes, 2016).

A representatividade da diabetes no universo dos utentes hospitalizados no SNS tem vindo a aumentar nos últimos anos, nomeadamente nos internamentos com uma duração superior a 1 dia. Em 2015 verificou-se ainda que excluindo os *Day Case*, a durabilidade dos internamentos é superior relativamente ao registado globalmente no SNS (Observatório da diabetes, 2016).

2.1.3 Tempo de Permanência Hospitalar

Com o crescente aumento da prevalência da diabetes e com o consequente impacto nos serviços de saúde, inclusive os hospitais, é cada vez mais importante o planeamento adequado da prestação de cuidados. Sabe-se, por estudos anteriores que pessoas com a diabetes têm longos períodos de internamento, em comparação com pessoas sem a doença (Carral et al., 2001). O tempo de permanência de um paciente define-se pelo intervalo desde a admissão à atribuição do estado de alta. A previsão do número de dias de internamento pode ser utilizada para diferentes finalidades. Exemplos destas finalidades são: uma melhor gestão de recursos hospitalares, emissão de alertas em situações em que medidas preventivas redobradas devam ser tomadas - nos casos em que previsão do número de dias de internamento é superior ao esperado. De referir ainda que a estimativa da duração dos internamentos, permitirá um agendamento mais

eficaz das cirurgias sem caráter de urgência – cirurgias eletivas, de acordo com a disponibilidade dos leitos esperada (Andersson, 2019).

A previsão do período total de internamento é benéfica para todos os elementos intervenientes: paciente, prestadores de cuidados de saúde, contribuintes do hospital e hospital. O paciente beneficia, na medida em que a previsão permite um planeamento personalizado para o seu internamento, podendo contribuir para a qualidade dos cuidados de saúde prestados. Para os prestadores de cuidados de saúde, pode ser um auxiliar nas tomadas de decisões. Os contribuintes do hospital, são responsáveis pelo pagamento dos cuidados de saúde e, sabe-se que as durações das estadias explicam 85 a 90% da variação dos custos hospitalares entre pacientes, pelo que, uma previsão eficaz pode traduzir-se numa boa previsão de custos. O hospital deseja otimizar a ocupação dos leitos, para possibilitar uma melhor assistência (Gentimis et al., 2018).

2.2 *Data Mining*

Ao longo do tempo o armazenamento e organização de grande volume de dados tem sido efetuado de forma cada vez mais eficiente, contudo, tem-se sentido a necessidade de utilizar devidamente essa grande quantidade de informação, de forma a transformá-la em conhecimento útil. *Data mining* permite extrair conhecimento através do reconhecimento de padrões e relacionamento entre variáveis, a partir de uma base de dados. O conhecimento é atingido a partir de técnicas confiáveis e validadas por comprovação estatística (Côrtes, Porcaro, & Lifschitz, 2002). *Data mining* engloba muitas técnicas de vários domínios, como estatística, machine learning, reconhecimento de padrões, sistemas de armazenamento de dados, visualização de dados e aplicação de algoritmos.

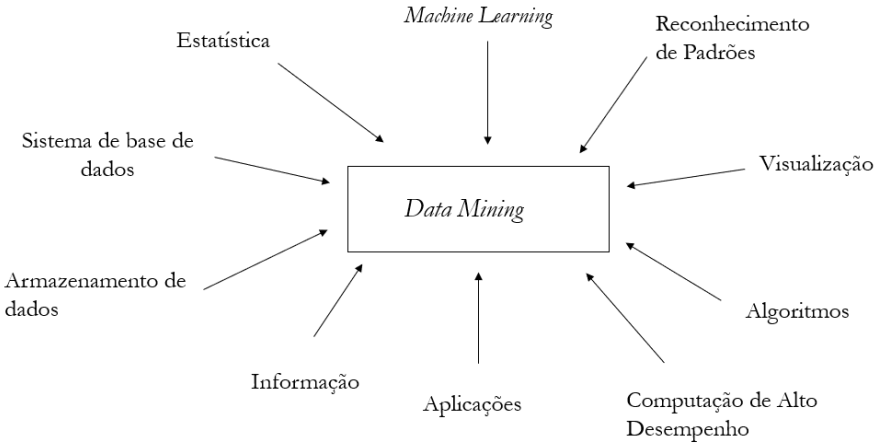


Figura 5 - Várias técnicas de diferentes domínios utilizados em *data mining*.

Data mining não tem uma definição única e é também conhecido como *Knowledge Discovery from Data* – descoberta de conhecimento a partir de dados, ou, KDD. Atualmente são vários os processos que padronizam e definem as fases e atividades de *data mining* (Camilo & Silva, 2009). O processo KDD consiste numa sequência iterativa, que normalmente envolve limpeza, integração, seleção, transformação, avaliação de padrões de dados, finalizando com a apresentação do conhecimento obtido. A esquematização de todo o processo encontra-se apresentada na Figura 6.

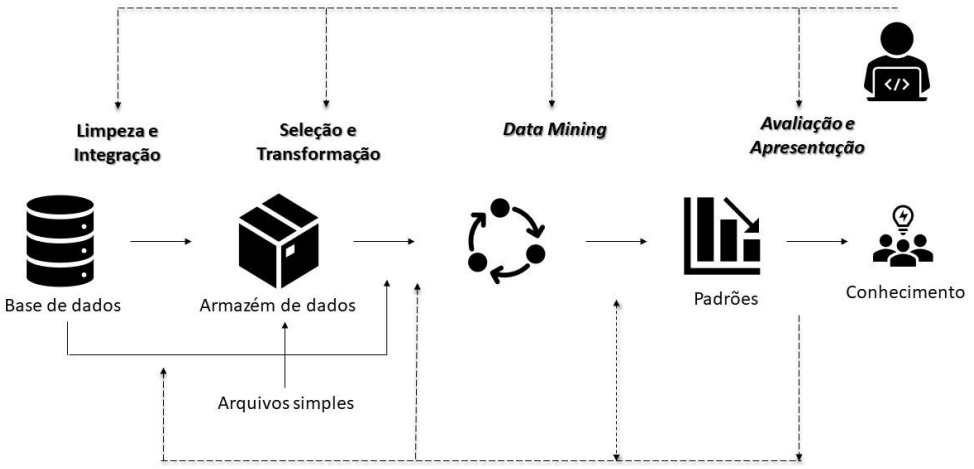


Figura 6- Diferentes fases do processo iterativo de *data mining*.

2.2.1 Limpeza de dados

Os dados do mundo real tendem a ser incompletos, a conter valores errados e a ser pouco consistentes, pelo que as rotinas de limpeza dos dados são muito importantes, pois permitem preencher valores em falta – *missing values*, suavizar dados ruidosos – *noise data*, a partir da remoção de *outliers* e a resolver inconsistências dos dados. As técnicas usadas vão desde a remoção de um registo com problemas, podendo existir a atribuição de valores considerados padrão, ou mesmo a aplicação de técnicas de agrupamento, de modo a atingir a descoberta de melhores valores.

Valores em falta: referem-se a registos, que podem ser relevantes, e que não possuem valores. Existem alguns métodos para preencher esses atributos:

• Exclusão de casos: excluiu do conjunto de dados as linhas que possuam pelo menos um atributo não preenchido. Este é o método mais simples, mas, em geral, não é o método mais adequado, à exceção se a linha a ser eliminada contiver vários atributos com valores em falta.

• Preenchimento manual de valores: na maioria das vezes é uma opção que consome muito tempo e nem sempre é exequível, principalmente quando existe um grande volume de dados com informação desconhecida.

• Substituição por valores globais constantes: os valores ausentes são substituídos por um único valor tomado como padrão, de que é exemplo "desconhecido" ou "*null*".

• Preenchimento com medidas estatísticas: utilização de medidas estatísticas para preenchimento dos valores ausentes. A média e a moda são dois dos exemplos de medidas estatísticas utilizadas. A média é utilizada quando se trata de atributos numéricos e a moda, por sua vez, em atributos categóricos. Uma variação desta metodologia é utilizada em problemas de classificação, em que ao invés de ser considerado todo o volume de dados, o cálculo das medidas estatísticas é feito para cada classe (Kelly, 2014).

• Preenchimento com métodos de *data mining*: mesmo aquando da etapa de pré-processamento. Os algoritmos de *data mining* podem ser utilizados para preencher valores em falta. Por exemplo, utilização de valores de maior probabilidade para preencher os valores ausentes. Este valor pode ser estimado a partir de técnicas de formalismo *bayseano*, indução por árvore de decisão ou técnicas de regressão. Este método de preenchimento de atributos é o mais popular por considerar mais informação para prever os valores ausentes (Côrtes et al., 2002).

Noise data: o ruído – *noise data*, é uma componente aleatória de um erro de medição de uma variável. Existem algumas técnicas para suavizar (*smooth*) essas variáveis do tipo numéricas:

• Binning: Este método visa suavizar um valor de dados classificados, tendo em conta a sua vizinhança. Os valores são ordenados e posteriormente são repartidos por grupos, em que cada grupo contém o mesmo número de valores. Diferentes métodos podem ser utilizados com a finalidade de ajustar os valores

de grupos. Em cada grupo é utilizado um critério na preferência de uma medida de ajuste, que pode passar por um valor aritmético, mediana, ou um valor limite. Depois, em cada grupo, os valores são substituídos pelas medidas calculadas (Han, Kamber, & Pei, 2016).

• Análise de *Outliers*: *outlier*, ou valor atípico são valores que apresentam discrepância relativamente aos restantes valores da mesma série. Os *outliers* podem ser detetados a partir de agrupamento – *clustering*. O método de *clustering* envolve o agrupamento de objetos para que os objetos dentro de um cluster tenham elevada similaridade, mas sejam diferentes dos objetos noutros clusters. Na Figura 7 é apresentada a esquematização de um agrupamento com três clusters, os *outliers* podem ser identificados como os valores que se situam fora do conjunto de clusters (Côrtes et al., 2002).

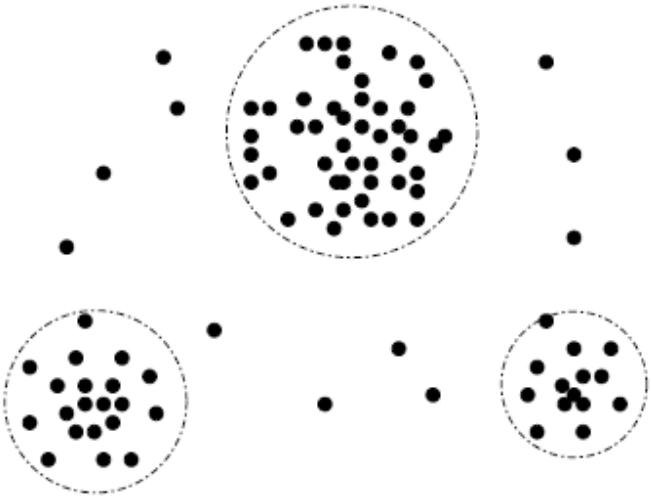


Figura 7 - Esquematização de três clusters.

• Regressão: a suavização a partir de regressão é uma técnica utilizada que torna valores de dados numa função. A regressão pode ser linear ou linear múltipla. A primeira permite encontrar a melhor linha para ajustar dois atributos/variáveis, de modo que um atributo possa ser utilizado para prever outro. A regressão linear múltipla permite que mais de dois atributos estejam

envolvidos e sejam ajustados a uma superfície multidimensional (Han et al., 2016).

2.2.2 Integração, seleção e tratamento dos dados

A integração dos dados permite de forma consistente e coerente, a integração de diversas fontes de dados. Esta necessidade de obter um repositório único e consistente é sentida essencialmente quando os dados surgem de diversas fontes, por exemplo, arquivos de texto, base de dados, imagens, vídeos, entre outras. É realizada uma análise detalhada dos dados verificando a existência de possíveis redundâncias., dependências entre variáveis e valores conflitantes (Côrtes et al., 2002).

Posteriormente é necessário proceder à seleção de dados que sejam relevantes para aplicação das técnicas de *data mining*. Uma vez que teoricamente esta etapa ocorre depois da integração dos dados, permite a seleção apenas dos dados que irão, efetivamente, ser usados. Uma vez que na fase anterior já foi garantida a coerência entre as diversas fontes dos dados.

Quanto ao tratamento dos dados, permite a transformação ou consolidação dos dados no formato mais adequado para o processo de *data mining*. A transformação pode envolver processos de remoção de dados ruidosos (como já identificado, técnicas de *binning*, agrupamento e regressão), agregação, por exemplo hospitalizações diárias podem ser agregadas em hospitalizações semanais, ou mensais. Nesta fase podem ainda ser aplicados processos de generalização, em que um atributo mais pormenorizado pode ser generalizado, por exemplo, categorização em criança, jovem, adulto, mediante a variável idade. Outra característica da transformação de dados é a possibilidade de normalização dos dados e de construção de atributos. Na normalização, é possível atribuir uma nova escala a um dado atributo, de modo a que os valores compreendidos desse atributo possam estar dentro de um determinado intervalo, por exemplo, entre -1 a 0. Por fim, na construção de atributos, novos atributos podem ser gerados a partir de alguns já existentes, de que é exemplo o cálculo do IMC a partir das variáveis peso e altura pré-existentes (Han et al., 2016).

2.3 Data Mining, avaliação e apresentação

O *Data Mining* é a fase fundamental, em que as técnicas de análise e extração de dados são aplicadas. Envolve os métodos de identificação da finalidade do processo de *data mining*, estudo das técnicas mais adequadas a serem aplicadas e por fim, abordagem da aplicação dos seus processos.

A avaliação é a deteção de padrões relevantes entre os vários apresentados pelo processo de *data mining*, tendo em consideração as medidas de interesse. Nem todos os padrões obtidos podem ser considerados interessantes para o estudo em questão. Pelo que, nesta fase, tendo sempre por base a significância estatística, deve ser feito um estudo e avaliação dos resultados, identificando os padrões que devem ser utilizados.

Por fim, a apresentação do conhecimento obtido, refere-se à utilização de meios de representação e visualização da informação por forma a apresentar o conhecimento adquirido. Nesta fase devem ser apresentadas as descobertas atingidas, bem como a determinação da forma que esse conhecimento pode ser usado para melhorar o processo de tomada de decisão. É nesta fase final também que se espera a realização de um balanço dos aspetos positivos e negativos do projeto desenvolvido e de planeamento de projetos futuros.

2.4 Métodos ou Técnicas

Uma área de influência do *data mining* é o *machine learning*. Esta área permite, a partir da análise de dados, aprender de forma automática, para posteriormente fazer previsões em dados que sejam desconhecidos (Paiva, 2016). Uma vertente importante do machine learning é a aprendizagem e existem dois tipos diferentes:

• **Aprendizagem supervisionada:** são apresentados alguns conjuntos padrões de entrada e os respetivos padrões de saída. Requer, portanto, um conhecimento prévio do comportamento esperado. O algoritmo percorre primeiramente um conjunto de dados de treino em que é conhecido o valor da variável classe, de forma a existir aprendizagem sobre os dados. Posteriormente

o algoritmo é colocado num conjunto de dados, designado por modelo classificador, para que já seja capaz de fazer previsão da variável de classe ao analisar outros dados no mesmo formato. Este tipo de aprendizagem pode ser considerado como um método de classificação, ou de modelo de regressão (Paiva, 2016).

• **Aprendizagem não supervisionada:** não existe um agente externo. Somente os padrões de entrada estão disponíveis. Este tipo de aprendizagem só é possível quando existe redundância nos dados de entrada, para ser possível encontrar padrões. São processadas as entradas e, detetando as suas regularidades, o algoritmo tenta progressivamente estabelecer representações internas para codificar características e classificá-las automaticamente (Ferneda, 2006) sendo por isso também designado por associação.

Em *data mining*, ambos os tipos de aprendizagem são muito importantes e já existem variações entre as duas categorias, sendo essas propostas designadas como semi-supervisionada. Tarefas de agrupamento e associação são considerados como não supervisionada. As tarefas mais comuns de aprendizagem supervisionada são a classificação (que também pode ser não supervisionada) e a regressão. De seguida, serão exploradas algumas técnicas de aprendizagem supervisionadas, algumas das quais usadas nesta dissertação.

2.4.1 Associação

As regras de associação permitem encontrar relacionamentos ou padrões frequentes entre conjuntos de dados. A Mineração de Regras de Associação – MBA, também conhecido como análise de cesta de mercado foi introduzida por Agrawal & Srikan (1994) como uma maneira de encontrar padrões associativos a partir de dados de um cesto de supermercado (Anselmo, 2017). Esta abordagem tem como base a teoria de que os clientes que compram determinado produto têm maior probabilidade em adquirir um outro item específico. As regras de associação não extraem a preferência de um individuo, mas localizam padrões entre conjuntos de elementos de cada transação distinta. As informações obtidas

a partir da análise podem ser usadas com finalidade de estratégias de marketing, vendas, serviços e operações (Chen, Tang, Shen, & Hu, 2005). Dada a aplicabilidade desta metodologia, tem vindo ao longo do tempo a ser utilizada para as mais diversas finalidades, desde auxílio nos diagnósticos médicos, análise de dados genéticos, sistemas de recomendação com base nas preferências do utilizador. O relacionamento é modelado na forma de um algoritmo condicional:

$$Pão => Manteiga\ [Suporte = 20\%, Confiança = 80\%]$$

A regra acima indica, com um suporte de 20% e uma confiança de 80%, que uma pessoa que compre pão provavelmente também irá comprar manteiga. Uma regra de associação é composta por dois conjuntos: o antecedente (do lado esquerdo) – LHS e o consequente (do lado direito) – RHS e regra lê-se da seguinte forma: "se antecedente então consequente" (Anselmo, 2017).

O modelo de regras de associação consiste em encontrar todas as regras que contenham suporte e confiança iguais ou superiores a um suporte mínimo e uma confiança mínima, previamente estipulados.

Suporte: número de transações que incluem todos os itens na parte antecedente e consequente da regra. Para a regra representada na Figura 8, o suporte da regra mede o número total de registos de transação que contêm simultaneamente os conjuntos de itens X e Y. O suporte serve para garantir que uma transação pertence a um padrão, ao ocorrer frequentemente. Uma transação com um suporte muito baixo pode não pertencer a um padrão e ser apenas uma ocorrência pontual.

Confiança: mede a probabilidade condicional de ocorrer Y dado que ocorreu X. Tendo como base a regra da Figura 8, uma confiança de 80% significa que em 80% das vezes que X ocorre, Y também ocorre. A confiança é útil para provar a fidedignidade de uma dada regra

Lift: também designado por coeficiente de interesse, mede a dependência entre o antecedente e o consequente. Na regra da Figura 8, interpreta-se como a indicação de quanto mais frequente se torna Y quando X ocorre. Um lift=1 significa que X e Y são independentes. O lift>1 indica que X e Y são

positivamente dependentes. Para um lift < 1, X e Y são negativamente dependentes. Ou seja, quanto maior o valor do lift, mais interessante é a regra, pois maior é a dependência entre os itens que a constituem.

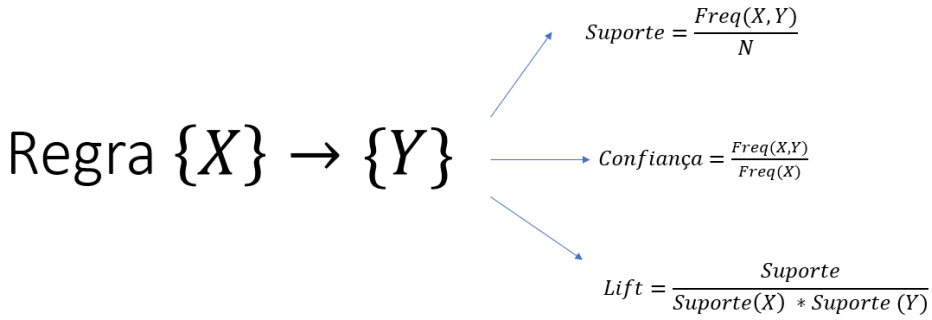


Figura 8 - Equações para o cálculo do suporte, confiança e lift.

2.4.2 Classificação

Classificação tem como objetivo a previsão de variáveis categóricas. Esta tarefa pode ser explicada como a procura de uma função que permita associar cada registo de uma base de dados a uma única categoria, também designada por classe. Depois de ter sido identificada, a função é posteriormente usada para fazer a previsão da classe em que os tais registos se enquadram (Kelly, 2014).

Um exemplo prático de aplicação de tal modelo é para previsão da doença da diabetes. Dado um conjunto de dados de utentes e cujo diagnóstico da doença é conhecido, é possível gerar um modelo para prever o diagnostico em novos dados em que a classificação da doença seja desconhecida. São vários os algoritmos de classificação, alguns dos mais conhecidos são os seguintes:

• **Árvores de decisão:**

São constituídas por um conjunto de elementos que contêm um teste num atributo, chamados de nós e cada ramo contém um possível valor deste atributo. Nesta técnica, uma decisão é tomada através de um percurso com origem no nó raiz, ou ponto de partida (maior nível hierárquico) até ao nó terminal, que corresponde a uma classe. No exemplo da Figura 9 a árvore de decisão mostra 4 variáveis: at1, at2, at3 e at4 numa hierarquia de decisão. Cada ramo desta árvore corresponde aos possíveis valores de cada variável. As "folhas" da árvore (elipses) correspondem aos valores de classe para as instâncias. Neste problema específico cada instância pode pretender a uma de duas classes "Yes" ou "No". Portanto esta árvore implementa 7 regras distintas. As mais curtas dizem que se o valor de at1 for b1 ou c1 a classe deve ser "No" (lado mais direito da árvore). (Maglogiannis, Karpouzis, Wallace, & Soldatos, 2007) . Em cada ramificação, os limites de recursos que melhor dividem as amostras localmente são encontradas. As métricas mais comuns para definir qual a melhor divisão são a impureza de gini e ganho de informação para tarefas de classificação.

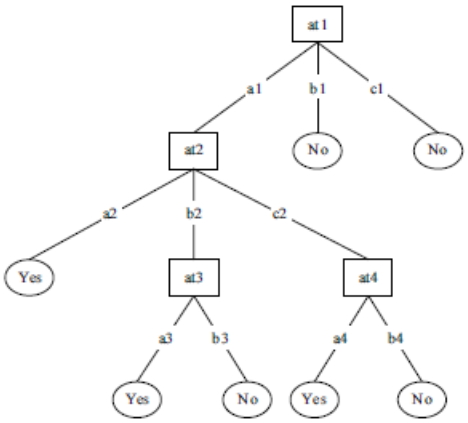


Figura 9 - Exemplo de árvore de decisão (Maglogiannis et al., 2007).

• **Floresta Aleatória:**

Tal como o próprio nome indica, consiste em agrupar um grande número de árvores de decisão individuais que operam como um conjunto. Cada árvore individual na floresta aleatória prevê a classe e a classe com mais resultados obtidos nas diversas árvores torna-se a previsão do modelo. Para o exemplo da Figura 10, o modelo de floresta aleatória prevê 1 dado que a maioria das árvores de decisão individuais preveem este resultado (Yiu, 2020). As florestas aleatórias estão a tornar-se cada vez mais populares por lidarem muito bem com interações complexas (Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008).

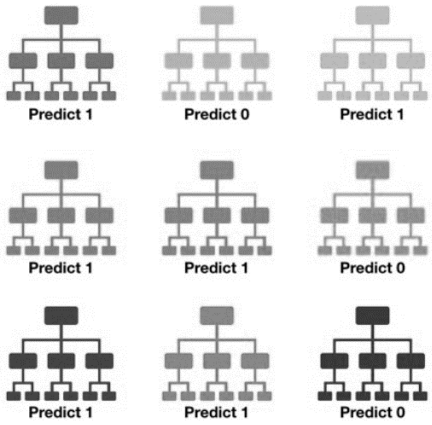


Figura 10 - Exemplo de previsão de floresta aleatória (Yiu, 2020).

• **Redes Neuronais Artificiais:**

São modelos simplificados do sistema nervoso central do ser humano (Cortez & Neves, 2000). São compostas por diversas unidades computacionais paralelas, interconectadas. As conexões entre os nós, também designadas por ligações ou sinapses entre nodo, contêm o conhecimento da rede, uma vez que possuem um valor associado – peso. Cada um dos neurónios artificiais, efetua um certo número de operações simples e transmite os seus resultados às unidades vizinhas com as quais possui conexão. Quando um nó recebe um determinado input, esses dados vão sofrer uma transformação matemática, por exemplo uma multiplicação com o peso, sendo depois os dados transmitidos para o nó

seguinte. Quando for ultrapassado um valor limite, é produzido o resultado final - Através de um processo de aprendizagem, as redes neuronais passam a ser capazes de reconhecer padrões, mesmo que os dados em causa não sejam lineares, estejam incompletos ou mesmo contraditórios (Cortez & Neves, 2000).

- **Redes Bayesianas:**

São uma técnica que se baseia no teorema de *Bayes*, que descreve a probabilidade de um evento tendo em conta o conhecimento a priori que pode estar relacionado com o evento. As redes Bayseanas são grafos que representam relações de probabilidade condicional.

Este tipo de estrutura permite fazer uma observação causa-efeito, uma vez que permite alterar o valor de algumas variáveis e, visualizar o respetivo efeito provocado nas restantes variáveis devido à alteração. Vários estudos comparativos demostram que os algoritmos Bayesianos, também conhecidos como naive *Bayes*, alcançaram resultados compatíveis com os métodos de árvore de decisão e redes neuronais. É um dos algoritmos mais utilizados por ser bastante simples e ter associado um alto valor preditivo (Zhang, 2004).

- **Support Vector Machines (SVM):**

As máquinas de vetores de suporte são modelos lineares que permitem classificação ou regressão.

É uma técnica que utiliza a noção de "margem", ou linha de separação, visto que hiperplanos de margem máxima são construídos com o objetivo de separar as classes num dado conjunto de dados (Figura 11). Toma-se que quanto maior for a distância entre os hiperplanos paralelos, melhor será a previsão (Veiga & Ferreira, 2011).

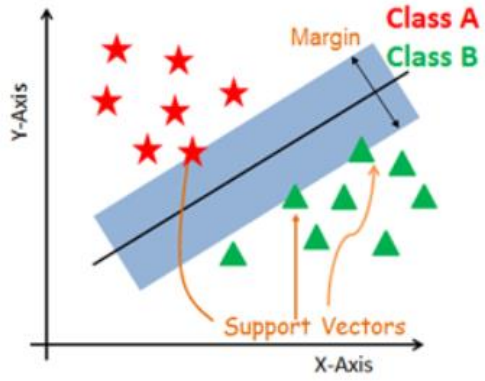


Figura 11 - Representação de hiperplano num dado conjunto de dados (DataCamp, 2019).

Validação de Modelos

Ao longo do processo de *data mining*, são vários os algoritmos e parâmetros que são testados por diversas vezes, o que leva à necessidade da existência de um modo de avaliação para facilitar a comparação e escolha do melhor modelo e parâmetros. É por isso necessário decidir o método de avaliação, que é a forma como os dados são divididos para os respetivos subgrupos de treino e de teste e qual o procedimento para estimar a precisão do modelo. A outra decisão importante são as métricas de avaliação do modelo. A matriz de confusão permite avaliar modelos de classificação, o número de falsos positivos (FP), falsos negativos (FN), verdadeiros positivos (VP) e verdadeiros negativos (VN), além de permitir o cálculo da *accuracy*, sensibilidade (taxa de verdadeiros positivos) e especificidade (taxa de falsos positivos) - Tabela 1 e Tabela 2.

Tabela 1 - Tabela de confusão.

	Condição Presente	**Condição Ausente**
Teste Positivo	VP	FP
Teste Negativo	FN	VN

Tabela 2 - Fórmulas para o cálculo de precisão, sensibilidade e especificidade.

Accuracy	$\frac{(TP+TN)}{(TP+FN+FP+TN)}$
Sensibilidade	$\frac{VP}{VP+FN}$
Especificidade	$\frac{VN}{VN+FP}$

Curva de *Receiver Operating Characteristic* (ROC), permite estabelecer a relação, para um modelo de classificação binária, entre a sensibilidade e 1 – especificidade.

Na figura 8 é possível observar duas curvas de ROC para dois modelos diferentes. A linha diagonal é correspondente à fronteira em que a probabilidade de encontrar um verdadeiro positivo é igual à de encontrar um falso positivo. A área a baixo da curva (AUC), fornece informação quanto ao melhor algoritmo a escolher, visto que quanto mais abrupta for a curva de ROC, isto é, quanto maior a área abaixo da curva, melhor é o modelo (Han et al., 2016).

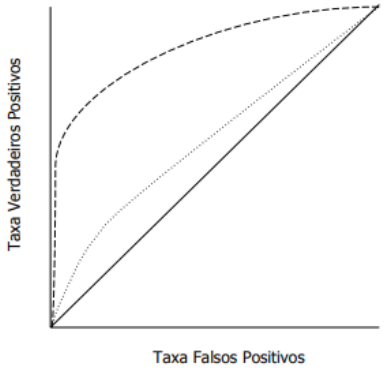


Figura 12 - Curva de ROC de dois modelos diferentes (Han et al., 2016).

As formas como as métricas descritas anteriormente irão ser utilizadas designam-se por métodos de validação. O objetivo crucial é a obtenção de uma avaliação que seja fidedigna, para que os resultados chegados na fase de avaliação sejam idênticos aos obtidos na previsão de novos casos (Veloso, 2000). Alguns dos métodos mais utilizados são:

- ✓ **Holdout**: o conjunto de dados é dividido em dois subgrupos. O modelo é construído com parte dos dados (grupo de treino) e depois, as métricas de avaliação são colocadas no subgrupo que não participou no processo de construção do modelo (grupo de teste). Desta forma o erro estimado é uma aproximação do erro real.

- ✓ **Cross-validation:** neste método, todos os dados são utilizados para aprendizagem do modelo. O conjunto de dados é dividido em k partes mutuamente exclusivas e o modelo é construído à custa de k-1 partes e testado no subgrupo sobrante. O processo é repetido k vezes e métricas de avaliação são aplicadas em cada uma das vezes, o valor da validação é dado pela média dos valores obtidos em todas as iterações. Existem estudos que sugerem que a validação seja obtida pelo cálculo específico a partir de TP, FP, TN e FN (Forman & Scholz, 2010).

- ✓ **Leave-1-out:** é um caso particular de validação cruzada. Com a validação cruzada calcula-se uma estatística na (s) amostra (s) deixada (s), enquanto com este método calcula-se uma estatística apenas das amostras mantidas. Nesta abordagem são calculados N cálculos de erro, sendo N o número total de dados.

2.4.3 Regressão

O método de regressão pode-se assimilar à classificação, no entanto, destina-se a situações em que o valor de uma variável é obtida a partir dos valores das

restantes variáveis, com a particularidade do registo ser identificado por um valor do tipo numérico. Pode ser utilizado, por exemplo, para estimar a probabilidade de sobrevivência, sabendo um conjunto de resultados de exames efetuados. As regressões podem ainda ser classificadas como lineares ou não-lineares.

Regressão linear: as regressões deste tipo são as mais simples e são utilizadas quando existe uma relação linear entre as variáveis preditoras (variáveis dependentes) e a resposta (variável independente). O número de varáveis envolvidas pode variar mediante os casos.

Regressão não-linear: são utilizadas em situações em que entre as variáveis preditoras e a resposta não existe uma relação linear. Um exemplo deste tipo de técnica é a regressão polinomial que consiste na adição de termos polinomiais ao modelo linear. O que, pela transformação das variáveis se consegue obter um modelo linear, que depois se consegue resolver pelas técnicas dos mínimos quadrados (Kelly, 2014).

3. Estado da Arte

Neste capítulo serão abordadas, de forma sucinta, diferentes técnicas utilizadas na atualidade sobre os temas que o presente estudo se debruça. Serão apresentados estudos relativos à previsão do número de dias de internamento de pacientes, em contexto hospitalar. Num primeiro momento serão apresentados os estudos relativos à abordagem como um problema de classificação e posteriormente à análise como um problema de regressão.

Em ambas as partes é de ressalvar, que se tentou dar importância não só aos estudos em que a população em estudo são pacientes diabéticos, por forma a conseguir estudar uma maior diversificação de técnicas e avaliar diferentes resultados, com o objetivo de auxiliar na escolha da melhor técnica a utilizar nesta dissertação.

3.1 Previsão de dias de internamento – Problema de Classificação

Os hospitais enfrentam diariamente problemas relacionados com organização dos leitos dos utentes, devido aos recursos limitados. Como tal, avaliar e prever o número de dias de internamento, mesmo sendo uma tarefa desafiadora, tem sido alvo de estudo ao longo dos tempos. Um estudo feito por (Walczak, Scorpio, & Pofahl, 1998) utilizou redes neuronais artificiais para avaliar o nível de doença dos pacientes com trauma pediátrico e, prever o número de dias de internamento. Foi estabelecida uma janela de tempo de 10 minutos, ou seja, os valores que não possam estar disponíveis dentro de 10 minutos após a chegada do utente no hospital, foram eliminados da amostra. Foram por isso utilizadas informações

dos pacientes que estão disponíveis no momento da admissão do paciente, ou logo de seguida. A especificidade da previsão foi definida em três categorias diferentes: curta (duração de internamento menor ou igual a 1 semana), média (mais de 1 semana, mas menor ou igual a 2 semanas) e longa (superior a 2 semanas). Foram aplicados quatro diferentes algoritmos nos dados em estudos e avaliados sob a métrica de comparação *accuracy* para as diferentes categorias de internamento. Detetou-se que a rede neuronal *fuzzy* ARTMAP deve servir como o principal preditor do número de dias de internamento, visto ter-se obtido uma *accuracy* de 100%, para os internamentos curtos. No entanto, o seu desempenho não foi tão positivo para os internamentos mais prolongados. Por esse facto, foi sugerida complementar com a rede neuronal de retro propagação de camada única, para os cuidados prolongados, pois este algoritmo obteve melhor desempenho para estas categorias de internamento. Esta pesquisa determinou, portanto, que a combinação de dois sistemas tem um melhor desempenho do que apenas um único sistema de rede neuronal.

Também (Hachesu, Ahmadi, Alizadeh, & Sadoughi, 2013) tentaram prever o número de dias de internamento hospitalar. A população alvo do estudo foram 4948 pacientes com doença arterial coronária (DAC), que foram admitidos no Hospital Académico e Educacional do Centro Médico e de Pesquisa Cardiovascular Rajaei em Teerão, Irão. O conjunto da amostra continha 36 atributos diferentes. A variável do período de internamento foi distinguida em três intervalos: [0,5], [6,9] e [10,[dias. Para o estudo foram utilizadas quatro técnicas diferentes: árvores de decisão, máquinas de vetores de suporte (SVM), rede neuronal artificial (RNA) e modelo de *ensembles*, isto é, um novo modelo combinando os três anteriores. Uma matriz de confusão foi obtida para calcular a *accuracy*, especificidade e sensibilidade. Os resultados demonstraram que o SVM foi o que obteve o melhor desempenho. Foi possível ainda apurar que existe uma tendência para o aumento do número de dias de internamento com pacientes que tenham doenças pulmonares, respiratórias ou pressão alta.

Com o mesmo objetivo de previsão de número de dias de internamento, (Morton et al., 2014) utilizaram diferentes técnicas - regressão linear múltipla (MLR), máquinas de vetor de suporte (SVM) e SVM+, aprendizagem de múltiplas tarefas (MTL) e florestas aleatórias (RF) em pacientes diabéticos. O SVM+ é uma extensão do SVM que normalmente é utilizado com o modelo LUPI – melhora o desempenho usando efetivamente o conhecimento disponível para treinamento, mas não para teste. Os dados da amostra contêm cerca de 8 milhões de registos de pacientes, mas apenas 10 000 registos foram utlizados para estudos que foram selecionados estrategicamente. Para o estudo foram

considerados a categoria etária, sexo, raça, tipo de pagamento (seguro privado/particular, etc.), tipo de admissão (emergência, eletivo, recém-nascido…) e APR-DRG (*All Patient Refined Diagnosis-Related Group*). A variável número de dias de internamento foi dividida em 2 categorias: curta e de longo prazo. Foi considerada uma estadia curta, aquela cujo número total de dias de internamento fosse menor que 3 dias. Dada a significância dos resultados, pode-se concluir que o SVM + alcançou a *accuracy* média mais alta e a AUC média mais alta (*Accuracy* de 68% e AUC de 76%), seguidas por RF, MTL e MLR.

Um outro estudo, (Turgeman, May, & Sciulli, 2017), avaliou a previsão do número de dias de internamento, com base em variáveis estáticas, isto é, que não mudam durante o período de internamento de pacientes com insuficiência cardíaca. Foi feita a comparação de diferentes métodos: redes neuronais, árvore de decisão CART (árvore de classificação e de regressão), árvore de decisão CHAID (Deteção automática de interação com qui-quadrado), modelo linear generalizado de *Poisson*, SVM e árvore de regressão cubista. Deste estudo resultou num melhor resultado para o método clubístico (com um R^2 de 0.79), seguido de rede neuronal, CART, CHAID, SVM e modelo de *Poisson*. A principal vantagem da árvore clubística é interpretabilidade das suas previsões, possibilitando a compreensão dos fatores subjacentes que podem afetar o período de internamento. Verificou-se que o erro de previsão é maior para os pacientes que tiverem mais internamentos num passado recente e para aqueles que tiveram estadias hospitalares anteriores mais longas. Os resultados indicam que o número de dias de internamento depende principalmente do número de internações anteriores, do número de consultas ambulatoriais anteriores, do tempo decorrido desde a última alta e do número de dias de leito que o paciente teve no ano anterior ao índice atual de admissão.

Ainda sobre a influência das doenças cardiovasculares no período de internamento dos pacientes, (Daghistani et al., 2019), utilizaram registos médicos eletrónicos de todos os pacientes que foram admitidos no serviço de cardiologia de adultos no *King Abdulaziz Cardiac Center* (KACC) – Arábia Saudita. Os pacientes foram divididos em três grupos com base na duração do internamento: curto (< 3 dias), intermédio (3-5 dias) e longo (>5 dias). Foram aplicadas quatro diferentes técnicas: florestas aleatórias, redes neuronais artificiais, máquinas de vetor de suporte e redes Bayseanas. O modelo de floresta aleatória foi o que obteve o melhor desempenho com uma *accuracy* de 80%, sendo que as redes neuronais tiveram o pior desempenho, com 45%.

Tendo como objetivo a previsão do número de dias de internamento de doentes com AVC, (Al Taleb, Hoque, Hasanat, & Khan, 2017) utilizou dados do

departamento de Neurologia do Hospital King Fahd Bin Abdul-Aziz, correspondente a 105 atributos e 866 pacientes com AVC e comparou diferentes técnicas de previsão (árvore de decisão e redes Bayseanas). Por forma a categorizar a variável do período de internamento, foi feita uma discretização dos dados por algoritmo de agrupamento (Expectation Maximization - EM), tendo daí surgido quatro grandes clusters. Os quatro valores discretos do atributo de período de internamento que representa esses clusters são: 0-2 dias, 3-7 dias, 8-16 dias e > 16 dias. A técnica de redes Bayseanas atingiu o melhor desempenho, com 81.3% de *accuracy*, ao passo que o algoritmo J48 obteve 77.1%. A rede Bayseana obteve valores de sensibilidade, especificidade e AUC superiores para todas as categorias da variável de previsão.

(Gentimis et al., 2018) exploraram o uso de redes neuronais para prever o período de internamento de pacientes com vários diagnósticos. A base de dados explorada foi o MIMC-III que contém cerca de 50 000 registos em unidades de UTI entre 2001 e 2012. Subdividiram a previsão em dois estados principais, curta e longa. Sendo que foi considerada uma estadia curta se o número total de dias foi igual ou inferior a 5 (valor mais próximo da média de dias de internamento). O modelo preditivo executa com precisão as estadias longas e curtas, com uma *accuracy* de aproximadamente 80%.

A partir de uma fonte de dados de internamentos hospitalares de pacientes diabéticos, extraído no Repositório de Aprendizado de Máquina da Universidade da Califórnia, Irvine, (Alahmar, Mohammed, & Benlamri, 2018) aplicaram diferentes técnicas de previsão algoritmo combinado, redes neuronais, floresta aleatória distribuída (DRF), modelo linear generalizado (GLM), máquina de aumento de gradiente (GBM) e redes Bayseanas. De todos os métodos, o modelo combinado foi o que obteve melhor desempenho com 81% de AUC, seguidos por GMB e floresta aleatória distribuída com 80%. Embora o modelo de redes Bayseanas tenha obtido o menor desempenho (AUC=74%), isso não significa que retirar esta técnica do modelo combinado irá aumentar o desempenho do mesmo, uma vez que a diversidade das técnicas é um dos fatores que fortalece o desempenho dos algoritmos combinados. Utilizaram a mediana do conjunto de dados para determinar o limite entre o tempo de permanência curto versus longo.

Sob uma perspetiva diferente (Livieris et al., 2018) aplicaram aprendizagem semi-supervisionada (SSL) para prever o tempo de permanência hospitalar. Este tipo de aprendizagem é uma extensão da aprendizagem supervisionada e não supervisionada, em que dados não rotulados são acrescentados ao conjunto de treinamento para aumentar a eficiência do classificador. Assim, o SSL utiliza uma grande quantidade de dados não rotuladas juntamente com dados rotuladas para

criar um classificador eficiente e preciso. Este estudo tentou avaliar diferentes algoritmos SSL – auto treinamento, co-treinamento e tri-treinamento. O primeiro é geralmente mais simples e visa aprender por si próprio, baseado nos dados rotulados. O classificador depois de treinado, classifica os restantes dados não rotulados. No co-treinamento, os dados são divididos em dois subconjuntos. Dois classificadores são treinados com os dados rotulados a partir dos dois respetivos subconjuntos. Os classificadores utilizam os dados não rotulados e ensinam o outro classificador com os exemplos não rotulados que foram classificados com um bom desempenho. O tri-treinamento consiste em três classificadores. Caso dois dos classificadores concordem com uma previsão, eles rotulam o exemplo não rotulado com essa mesma previsão e, aumentam o terceiro classificador o exemplo recém-rotulado. Cada um destes algoritmos SSL foi avaliado tendo por base as técnicas: redes Bayseanas, MLP, algoritmo de otimização do mínimo sequencial (SMO), algoritmo 3NN, árvore de decisão C4.5 e algoritmo PART. Neste estudo, os pacientes foram classificados segundo uma variável trinomial do tempo de internamento: 1-2 dias, 3-6 dias e mais de 6 dias. Os resultados demonstram que o tri-treinamento foi o método que exibiu melhores resultados, com proporções de dados rotulados de 20% e 40%. Já para proporções de 10% e 30% o co-treinamento e autotreinamento, respetivamente, demonstraram maior número de vitorias. No entanto, o tri-training foi o que relatou o melhor desempenho e precisão obtida pelo teste post hoc de Finner. Os resultados permitiram ainda verificar que os algoritmos SSL obtiveram resultados comparativamente melhores do que os respetivos algoritmos supervisionados.

(Andersson, 2019) tentou também avaliar diferentes métodos de machine learning (árvore de decisão, floresta aleatória, árvores com gradiente impulsionado, máquina de vetor de suporte, AdaBoost RF e redes neuronais) para prever o período de internamento dos pacientes que chegaram pelas urgências, pois estes são de particular importância devido à sua natureza não programada de internamento e também porque para estes doentes a percentagem de dados em falta era baixa. Foram estabelecidas duas abordagens diferentes: a dos pacientes que deram entradas pelas urgências, que estipula uma janela temporal até ao momento em que o paciente foi internado na enfermaria. A segunda abordagem englobava um período maior, até um pouco depois da admissão. O objetivo era perceber se a espera de um período maior para fazer a previsão do período de internamento total do paciente compensava pelo aumento da precisão. Os resultados demonstraram que ao adiar um pouco a previsão no tempo de admissão, pode ser feita uma melhor previsão. Todos os

modelos obtiveram desempenhos bastante bons, à exceção da árvore de decisão. Para a abordagem de urgências a precisão variou entre 70% a 72%. Para o estágio da admissão, aumentou para 74% a 75%. Tendo em conta a *accuracy*, velocidade de treinamento e interpretabilidade, o algoritmo de floresta aleatória foi considerado o método recomendado.

Artigo	Tipo de Previsão	Algoritmos Utilizados	Melhor Desempenho	População	Métrica de Comparação
(Walczak et al., 1998)	Trinomial	**Redes Neuronais:** Backpropagation (de 1 e 2 camadas);Radial-Basis-Function (RBF); Fuzzy ARTMAP	Combinação Backpropagation de 1 camada + fuzzy ARTMAP	Pacintes de trauma pediátrico	*Accuracy*
(Hachesu et al., 2013)	Trinomial	Árvore de Decisão (C5.0); Rede Neuronal; Máquina Vetor de Suporte; Algoritmo combinado.	Máquina Vetor de Suporte (SVM)	Pacientes com DAC	*Accuracy* (96.4%) Especificidade (97.3%) Sensibilidade (98.1%)
(Morton et al., 2014)	Binomial	Regressão Linear Múltipla (MLR); Máquinas de Vetor de Suporte (SVM); SVM+; Aprendizado de Múltiplas Tarefas (MTL); Florestas Aleatórias (RF)	Máquinas de Vetor de Suporte (SVM)	Pacientes Diabéticos	Accuracy (68%) F-score (65%) AUC (76%)
(Turgeman et al., 2017)	Binomial	Redes Neuronais; CART; CHAID; Modelo de Poisson; Máquina vetor de suporte; Árvore de Regressão Cubista	Árvore de Regressão Cubista	Pacientes com insuficiência cardíaca congestiva	MAE* =1 R^2=0.79 *Mean Absolute Error
(Al Taleb et al., 2017)	Quadrinomial	Decision Tree (J48) Bayesian network (BN)	Bayesian network (BN)	Doentes com AVC	Accuracy (81.3%)
(Gentimis et al., 2018)	Binomial	Redes Neuronais	Redes Neuronais	Pacientes no geral	Accuracy (80%)
(Alahmar et al., 2018)	Binomial	Stacked Ensemble Method; Deep Learning (DL) - Redes Neuronais; Distributed Random Forest (DRF); Generalized Linear Model (GLM); Gradient Boosting Machine (GBM); Naïve Bayes Classifier (NB)	Stacked Ensemble Method	Pacientes Diabeticos	AUC (81%)
(Livieris et al., 2018)	Trinomial	Semi Supervisioned Learning - SSL Self-training Co-training Tri-training	Tri-training (não foram encontradas diferenças entre os algoritmos)	Pacientes no geral	Accuracy aprox. 64%

		with: Naive Bays (NB); Multilayer Perceptron (MLP); Sequential Minimum Optimization (SMO); C4.5 decision tree PART; 3NN			
(Andersson, 2019)	Binomial	Árvores de Decisão (DT); Floresta Aleatória (RF); Gradient Boosting (GB) Máquina Vetor de Suporte (SVM); AdaBoosted DT e RF; Redes Neuronais Multi-Layer (MLP)	Floresta Aleatória	Pacientes no geral	*Accuracy*= 72%
(Daghistani et al., 2019)	Trinomial	Floresta Aleatória; Redes Neuronais Artificiais; Máquinas vetor de Suporte; Redes Bayseanas.	Floresta Aleatória	Pacientes cardiologia	*Accuracy*=80%

Tabela 3 -Sumarização do estado da arte, sob um problema de classificação.

3.2 Previsão de dias de internamento – Problema de Regressão

Deve-se ter em atenção que prever o número de dias de internamento como um problema de regressão é uma tarefa mais difícil em relação à abordagem de classificação, devido à complexidade associada na previsão do número exato de dias. (Cummings, 2018), fez um estudo para prever o tempo de permanência hospitalar no momento da admissão com a base de dados MIMC-III, utilizando um modelo de regressão. Para calcular o desempenho, foram utilizadas as métricas de RMSE e R^2.

O RMSE é o desvio padrão dos resíduos (erros de previsão), pelo que um modelo sem erros teria um RMSE de 0. A equação é apresentada a seguir, em que n é o número de registos, Y é a previsão do número de dias de internamento e y o número real (Equação 1).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - y_i)^2}{n}}$$

Equação 1 - Cálculo de RMSE.

O R^2 , ou coeficiente de determinação, é uma medida de ajuste de um modelo estatístico linear. Quanto maior este valor, mais ajustado é o modelo à amostra. Foram usados cinco modelos de regressão: regressão de floresta aleatória, regressão de k vizinhos, regressão linear, regressão de aumento de gradiente e regressão SGD. O modelo de regressão de aumento de gradiente foi o que obteve o melhor R^2, com 37%.

Este estudo conseguiu ainda verificar, a partir do cálculo de RMSE, que o modelo de aumento de gradiente é melhor em mais de 24% em relação aos modelos de estimação por média ou mediana.

Também (Combes, Kadri, & Chaabane, 2014) tentaram fazer a previsão do número de dias de internamento como um problema de regressão. Utilizaram diferentes modelos:

- LR: regressão linear;
- DS (*Decision Stump*): classe para construção e uso de um stump de decisão;
- Modelo de M5P: indução de árvores para prever classes contínuas;
- REPTree: constrói uma árvore de decisão usando remoção de erros;
- SVM;
- PRLM: Modelos lineares de regressão de ritmo;
- KStar: classificador de vizinhos K-mais próximos.

Dos resultados, pode-se concluir que as performances dos algoritmos foram bastante semelhantes. Tendo, no entanto, o modelo M5P obtido o melhor coeficiente de correlação, com 0.5893 ficando logo de seguida LR com 0.5823.

Tabela 4 - Sumarização do estado da arte, sob um problema de regressão.

Artigo	**Tipo de Previsão**	**Algoritmos Utilizados**	**Melhor Desempenho**	**População**	**Métrica de Comparação**
(Combes et al., 2014)	Regressão	LR, DS, M5P, REPTree: SVM, PRLM, KStar	M5P e LR	Departamento de Emergência	*Coeficiente de Correlação:* 0.5893 e 0.5823
(Cummings, 2018)	Regressão	Regressão de floresta aleatória; Regressão de k vizinhos; Regressão linear; Regressão de aumento de gradiente; Regressão SGD	Regressão de aumento de gradiente	Pacientes no geral	$R^2 = 37\%$

4. Análise e Processamento de Dados

Neste capítulo será apresentada a metodologia usada neste trabalho. Será apresentado e explicado o *dataset* utilizado, assim como as decisões tomadas para o pré-processamento dos dados passando pela escolha dos modelos de previsão utilizados.

4.1 O *data set*

Os dados utilizados para o estudo deste trabalho foram obtidos do UCI *Machine Learning Repository* (Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, 2014) e representa o atendimento clínico em 130 hospitais dos EUA, durante 10 anos (1999 - 2008), resultando em 101 766 episódios e 50 variáveis diferentes. Os episódios respeitam os seguintes critérios:

- São episódios de internamento;
- São episódios em que a diabetes foi inserida como diagnostico, podendo ser motivo principal ou secundário;
- O tempo de permanência no hospital esteve compreendido entre 1 e 14 dias;
- Foram realizados testes de laboratório;
- Foi administrada medicação ao utente durante o internamento.

Tabela 5 - Apresentação e descrição de todas as variáveis do *data set*.

Nome Variável	**Descrição e valores**	**Tipo**	**% *missing values***
Id episódio	Identificador do episódio	Numérico	0
Id paciente	Identificador do paciente	Numérico	0
Raça	Valores: Caucasiano, Asiático, Africano, Americano, Hispânico e outros.	Nominal	2
Género	Valores: masculino, feminino e desconhecido/inválido	Nominal	0
Idade	Agrupamento em grupos de 10 anos: [0,10), [10, 20), . . ., [90, 100)	Nominal	0
Peso	Peso em libras	Numérico	97
Tipo de Admissão	Identificador com 9 valores distintos, por exemplo, emergência, urgência, eletiva, recém-nascido e não disponível	Nominal	0
Destino após alta	Identificador com 29 valores distintos, por exemplo, enviado para casa, morreu ou não está disponível	Nominal	0
Fonte de admissão	Identificador com 21 valores distintos, por exemplo, referência médica, ponto de atendimento, transferência de outro hospital.	Nominal	0
Nº dias internamento	Número de dias entre a data de admissão e de alta	Numérico	0
Código de pagamento	Identificador com 23 valores distintos, por exemplo, Medicare, Blue Cross e autoapagamento	Nominal	40
Especialidade	Especialidade médica, por exemplo, medicina interna, clínica geral.	Nominal	49
Nº proc. laboratoriais	Número de exames de laboratórios realizado naquele episódio	Numérico	0
Nº procedimentos	Número de procedimentos (exceto testes de laboratório) realizados durante o encontro	Numérico	0
Nº medicamentos	Número de medicamentos diferentes administrados durante o internamento	Numérico	0
Nº visitas	Número de consultas ambulatoriais do paciente no ano anterior ao encontro	Numérico	0
Nº visitas emergência	Número de visitas de emergência do paciente no ano anterior ao encontro	Numérico	0
Nº visitas pacientes internados	Número de visitas de internamento do paciente no ano anterior ao encontro	Numérico	0
Diagnóstico 1	O diagnóstico primário de 848 valores distintos	Nominal	0
Diagnostico 2	Diagnóstico secundário de 923 valores distintos	Nominal	0
Diagnostico 3	Diagnóstico secundário adicional de 954 valores distintos	Nominal	1
Nº de diagnósticos	Número de diagnósticos inseridos no sistema	Numérico	0
Resultado glicose	Indica a faixa do resultado ou se o teste não foi realizado. Valores: "> 200", "> 300", "Normal" e "Nenhum" se não for medido	Nominal	0

Resultado A1c	Indica a faixa do resultado ou se o teste não foi realizado. Valores: "> 8", "> 7", "normal" se o resultado for menor que 7% e "nenhum" se não for medido.	Nominal	0
Mudança medicação diabetes	Indica se houve uma alteração nos medicamentos para diabéticos (dosagem ou nome genérico). Valores: "change" e "no change"	Nominal	0
Medicação diabetes	Nome genéricos, por exemplo, metformina, repaglinida, nateglinida	Nominal	0
Medicamentos	Este campo indica se o medicamento foi prescrito ou houve uma mudança na dose. Valores: "aumentou" se a dosagem aumentou durante o encontro, "baixo" se a dose foi diminuída, "constante" se não mudou e "não" se o medicamento não foi prescrito. No total existiam 24 medicamentos diferentes.	Nominal	0
Readmissão	Dias para readmissão hospitalar. Valores: "<30" se o paciente foi readmitido em menos de 30 dias, "> 30" se o paciente foi readmitido em mais de 30 dias e "Não" para nenhum registo de readmissão.	Nominal	0

Análise estatística básica

Para a amostra em estudo, existem vários episódios referentes ao mesmo utente. Existem, portanto, 101766 episódios diferentes, correspondentes a 71518 pacientes. Decidiu-se manter todos os episódios referentes ao mesmo utente, uma vez que é interessante estudar quais as variações que levaram a uma alteração no número de dias de internamento. Na Figura 13 é possível analisar de uma forma geral a distribuição da população em estudo por diferentes categorias: género, etnia, idade e especialidade médicas mais frequentes de internamento.

Há um equilíbrio na variável género, verificando-se apenas uma diferença de aproximadamente 6% de maior número de mulheres em relação aos homens. Este valor superior de mulheres em relações a homens na amostra, não é sustentada pela informação atual de distribuição da diabetes no mundo. Segundo (Atlas, 2019), a prevalência estimada de diabetes em mulheres com idade entre 20-79 anos é ligeiramente menor do que nos homens (9,0% vs 9,6%). Uma minoria de apenas 3 indivíduos foi contabilizada como de sexo desconhecido ou inválido. No que se refere à etnia da população, e como já seria de se prever uma vez que os dados foram recolhidos em hospitais dos EUA, a etnia caucasiana é a mais evidenciada de entre todas as restantes, com 53491 indivíduos correspondente a aproximadamente 75% da população em estudo. De seguida, segue-se a etnia Afro-Americana, que representa 18%, seguindo-se a hispânica e

a asiática com aproximadamente 2% e 1% respetivamente. A restante população, cerca de 4%, foi categorizada como 'Outros', ou de valor desconhecido '?'.

Quanto à distribuição da população por idades, verifica-se um aumento crescente desde os 0 anos até aos 70-80 anos, sendo que nesta categoria, atinge o pico de incidência, abrangendo 18210 indivíduos, correspondente a 25% da população. A partir deste intervalo de idades, é visível uma diminuição acentuada. No entanto, tendo em conta que nos EUA no ano de 2008 a esperança média de vida era de 78.04 anos, pode-se constatar que nos intervalos de idades, 80-90 e 90-100, a frequência de internamentos é diminuída por consequência da diminuição da população correspondente a estes intervalos de idades. Pelo que, se pode constatar que existe um aumento de internamentos com o envelhecimento da população.

Quanto à especialidade médica de internamento, existe uma grande diversidade com cerca de 73 especialidades diferentes. O gráfico da Figura 11 imagem D, representa apenas as 5 mais influentes. A medicina interna ocupa o maior lugar, com cerca de 14635 internamentos, correspondente a 20%. Segue-se emergência/trauma, clínica geral e familiar, cardiologia e cirurgia geral que ocupam cerca de 33%. Os restantes 47% estão dispersos por diversas especialidades que vão desde endocrinologia, neurologia, radiologia, pneumologia, entre uma grande diversidade de outras especialidades médicas.

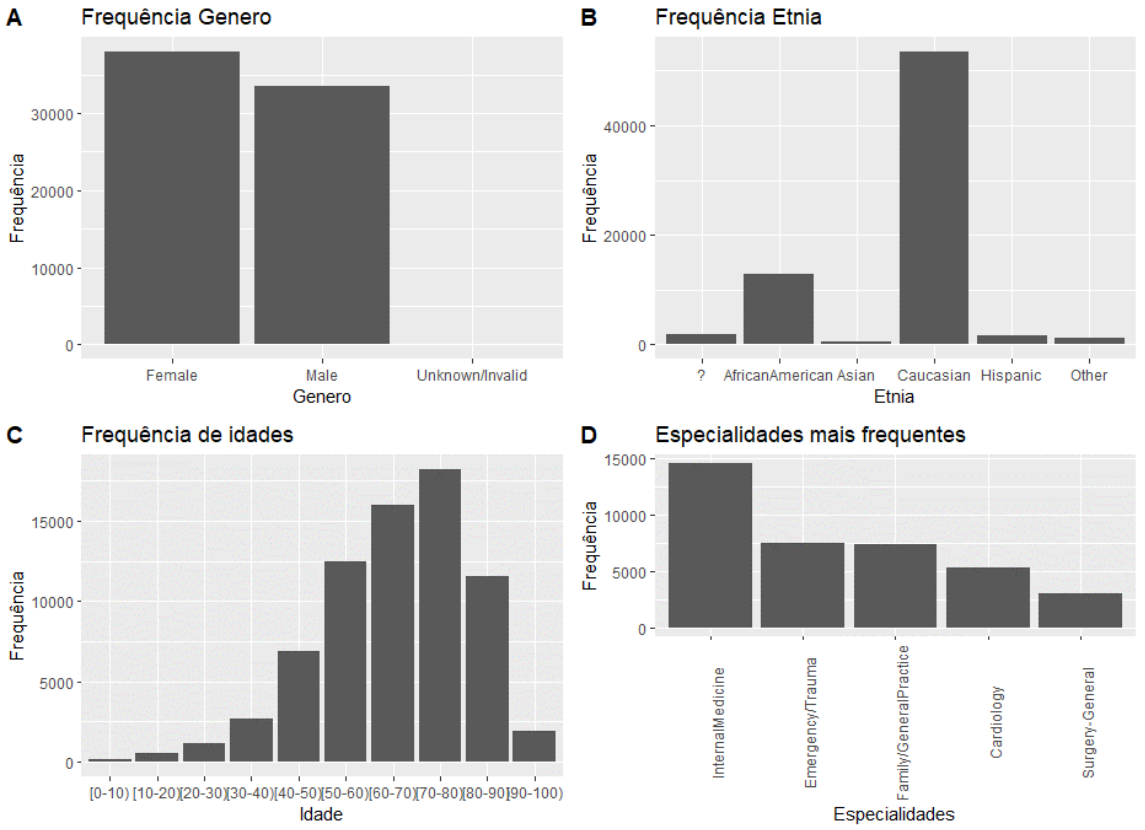


Figura 13 - Frequência da população em estudo por género, etnia, idades e especialidade de internamento mais frequentes.

4.2 Limpeza de dados

Embora o conjunto de dados em estudo tenha apenas uma pequena minoria com valores em falta, é necessário analisá-los e perceber o que fazer em cada situação. No gráfico da Figura 14 está representado percentualmente o nível de valores em falta para todas as variáveis onde algum valor é desconhecido. Os autores desta base de dados diferenciaram dois modos diferentes de valores em falta. No caso das variáveis representadas na Figura 14, de que é exemplo o peso, os autores assinalaram com "?" todos os valores que eram desconhecidos. Assumiu-se que aquela variável não foi medida e por isso é de valor desconhecido. Existem, no entanto, outras variáveis que embora em determinadas situações se desconheça o valor, os autores assinalaram como '*None*', no caso por exemplo da glicose não ter sido medida.

Por se considerar que uma ausência de medição é diferente da assinalação de que aquele atributo não foi medido, nesta secção apenas trataremos dos casos em que houve apenas ausência de medição. Esses casos ocorreram para as variáveis raça, peso, forma de pagamento, especialidade médica e os diagnósticos 1, 2 e 3.

Na Figura 14, pode ver-se uma linha horizontal que representa o limite percentual permitido de valores em falta considerado para este estudo (50%). Verifica-se, portanto, que por esta análise a variável peso foi desconsiderada para este estudo com uma percentagem de 97% de valores ausentes. Para as variáveis que se mantiveram no estudo, substitui-se os valores em falta por "Desconhecido".

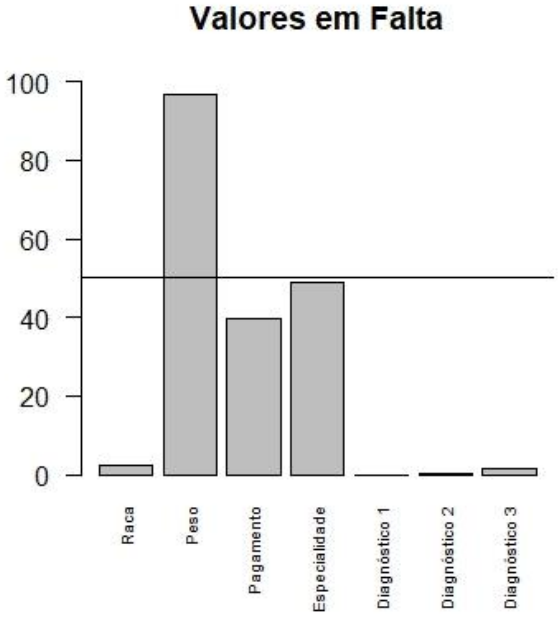


Figura 14 - Percentagens de valores em falta.

Identificação de *outliers*

Para estudar a incidência de *outliers*, fez-se um boxplot para as variáveis numéricas. No entanto, variáveis numéricas cujos valores representam codificação foram tratadas como categóricas. O resultado obtido é apresentado no gráfico da Figura 15.

Pode-se verificar que o bloxplot para as visitas anteriores ambulatoriais e de emergência, têm uma amplitude interquartil praticamente nula, o que significa que a variação é pouco significativa e que grande parte dos pacientes não continham visitas anteriores seja em ambulatório seja em emergência, no ano anterior ao episódio. É por isso, também que nestas variáveis se verifica uma maior incidência de *outliers*. Já a variável de número de visitas de internamento no ano anterior ao episódio, tem uma distância interquartil ligeiramente superior, demonstrando por isso um aumento de dispersão dos resultados para esta

variável, no entanto, o número de *outliers* embora seja menor, ainda é bastante alto.

As variáveis tempo de internamento e número de procedimentos têm amplitudes maiores, o que significa que existe uma grande variação dos dados para ambas. Os *outliers* são pouco evidenciados, e a distribuição é praticamente simétrica. É ainda visível que o número de dias mínimo de internamento foi de 1 dia, o que já seria de esperar dado que é este o limite mínimo de dias de internamento para os episódios deste estudo. Já o número máximo para a distribuição foi de 12 dias, verificando-se *outliers* para os 13 e 14 dias de internamento. Por fim, o número de procedimentos laboratoriais e número de medicamentos administrados, têm um aumento de *outliers* face às últimas duas variáveis anteriores. De referir ainda que para o número de procedimentos laboratoriais, os valores mínimo e máximo são mais distantes e a sua dispersão é mais significativa. No entanto, é na variável número de medicação administrada que se verifica um maior número de *outliers*.

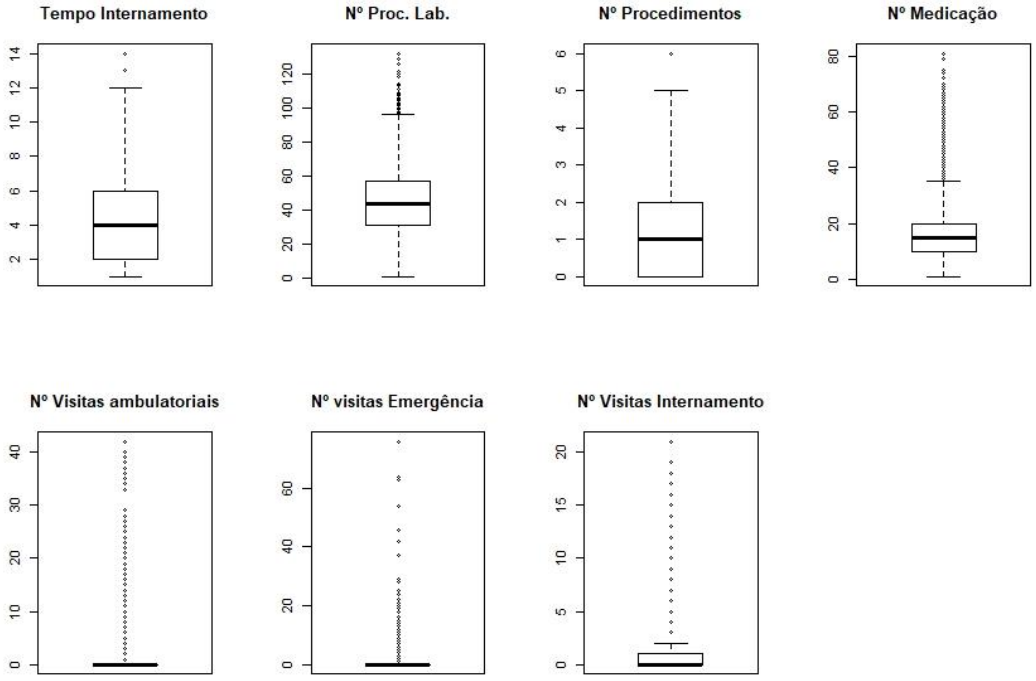


Figura 15 - Análise de *outliers* para variáveis numéricas.

Na área médica manter os *outliers* na amostra pode ser importante na medida em que estes podem conter informação essencial relativa a algum grupo específico de pacientes. Por este facto, decidiu-se avaliar se o número de visitas anteriores teria influência no prolongamento do internamento de um paciente.

Na Figura 16 é possível analisar a variação do número de dias de internamento com o número de visitas anteriores (ambulatória, de emergência e de internamento) no ano anterior ao encontro. Observa-se que a curva mais estável é a do número de visitas anteriores ambulatórias. Verifica-se que independentemente do número de internamentos ambulatórios anteriores ao encontro, o número de dias de internamento varia próximo dos 3 a 5 dias. A variação é ligeira e é atingido o máximo para um número total de visitas anteriores próximo dos 15.

Quanto à variação do prolongamento hospitalar tendo em conta o número de visitas anteriores de emergência, observa-se uma curva mais oscilatória. Observa-se um aumento do número de dias de internamento com o aumento de

episódios anteriores de emergência, atingindo o pico de 5 dias de internamento para aproximadamente 15 visitas anteriores de emergência. A partir daí, observa-se uma diminuição crescente do período médio de internamento. Poderá indicar que indivíduos que têm contactos hospitalares de emergência mais frequentes, têm internamentos mais curtos.

Verifica-se ainda que o número de dias de internamento médio para episódios que tenham tido internamentos no ano anterior ao encontro é próximo dos episódios que não tiveram contacto de internamento anterior, estando bastante próximo dos 4/5 dias de internamento. No entanto, é de realçar que para números muito elevados de internamentos anteriores (>15) ao encontro a média de dias de internamento reduz para próximo dos 2/3 dias. Esta informação pode-se traduzir por pessoas que tenham tido muitos internamentos anteriores, estivessem mais vulneráveis e por isso o desfecho tenha sido por exemplo morte ou transferência para outro hospital (estes casos serão tratados no ponto 4.4 - Seleção). Por se ter identificado alguma padronização do período de internamento com as visitas anteriores, decidiu-se por isso manter estas variáveis na nossa amostra, por se puderem traduzir em variáveis influentes no número de dias de internamento. Mais à frente, será estudado o fator de correlação destas variáveis com o período de internamento.

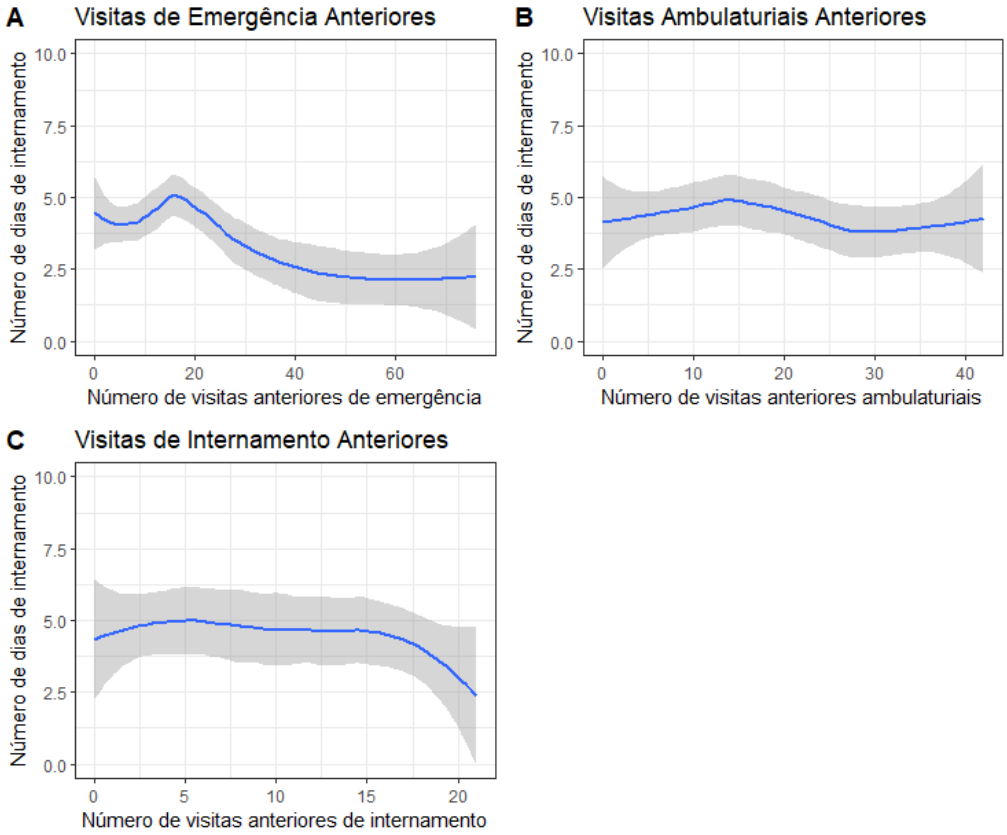


Figura 16 - Variação do número de dias de internamento médio com o número de visitas anteriores no ano anterior ao encontro.

4.3 Seleção

Antes da seleção, pode ser necessário fazer a integração de dados que visa combinar dados com origem em fontes distintas. Neste caso em particular, para o presente estudo, esta etapa não foi necessária visto que os dados estão relacionados entre si e num único formato (.csv).

Nesta etapa é importante perceber quais são os atributos que são de facto relevantes para o presente objetivo de estudo, uma vez que dependendo do objetivo do estudo nem todos os dados podem ter a mesma relevância. Esta é etapa é importante, uma vez que informação adicional pouco importante para o

estudo, poderá exigir um maior esforço computacional e criação de novas regras nos algoritmos utilizados, de forma desnecessária. Detetou-se que algumas categorias se traduzem em pouca informação útil para o nosso estudo dado terem uma variabilidade muito reduzida, ou nula. As variáveis referentes ao recurso de medicamentos indicam regra geral pouca variabilidade com 99% ou mais com a categoria 'No', indicando que esse medicamento não foi prescrito durante o internamento. Dada esta informação, optou-se por não considerar estas variáveis no estudo.

Dado que o principal objetivo desta análise é estudar o número de dias de internamento, decidiu-se extrair da nossa amostra todos os casos cujo desfecho terminou em morte ou transferência para um hospício, uma vez que poderia influenciar o nosso estudo. A referida filtragem resultou em 99 351 episódios, após a filtragem.

Pelo facto de apenas existirem 3 elementos que o género é desconhecido ou inválido, optou-se por retirar estes casos da nossa amostra, perfazendo um total de 99 348 episódios.

4.4 Transformação

Este passo serve para transformar os valores existentes em valores que possam ser utilizados da forma mais conveniente. Em relação ao diagnóstico de internamento, existem 3 variáveis diferentes, referentes ao diagnóstico 1, 2 e 3. Cada uma destas variáveis está codificada sob a Classificação Internacional de Doenças 9ª Revisão Modificação Clínica – CID-9-CM, que é bastante especifica, o que resultou em mais de 700 níveis diferentes para cada variável. Desta forma, tentou-se diminuir o número de níveis, diminuindo a especificidade da codificação. A codificação CID-9-CM segue uma sintaxe especifica, em que as principais categorias são agrupadas de forma a organizar e incluir um conjunto de doenças semelhantes. Desta forma e tendo por base a tabela 2 de (Strack et al., 2014), fez-se o mapeamento dos códigos descrito na Tabela 6. Isto é, agrupou-se os diferentes códigos pelas suas categorias principais, tendo em conta

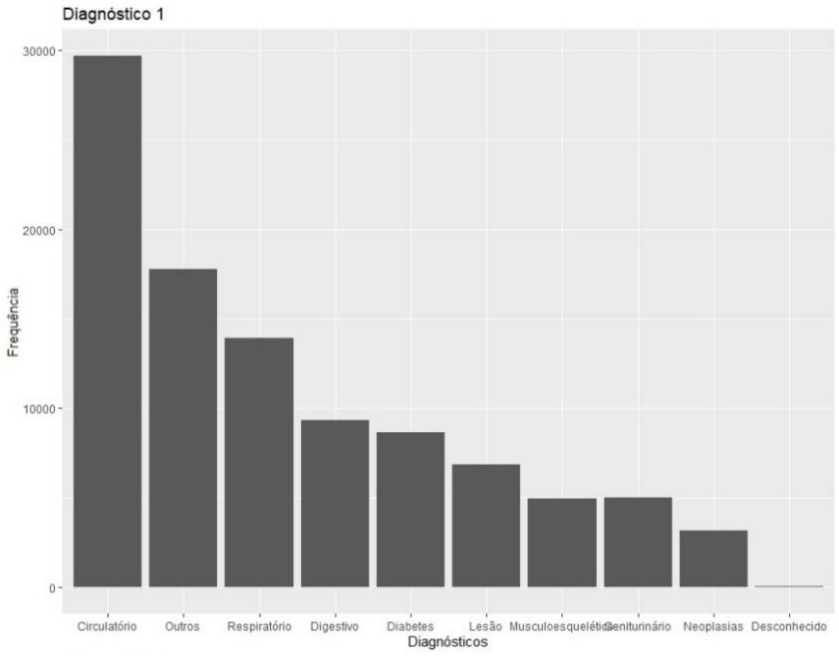
que todas as categorias que obtiveram menos de 3.5% de ocorrências foram registadas na categoria 'Outros'.

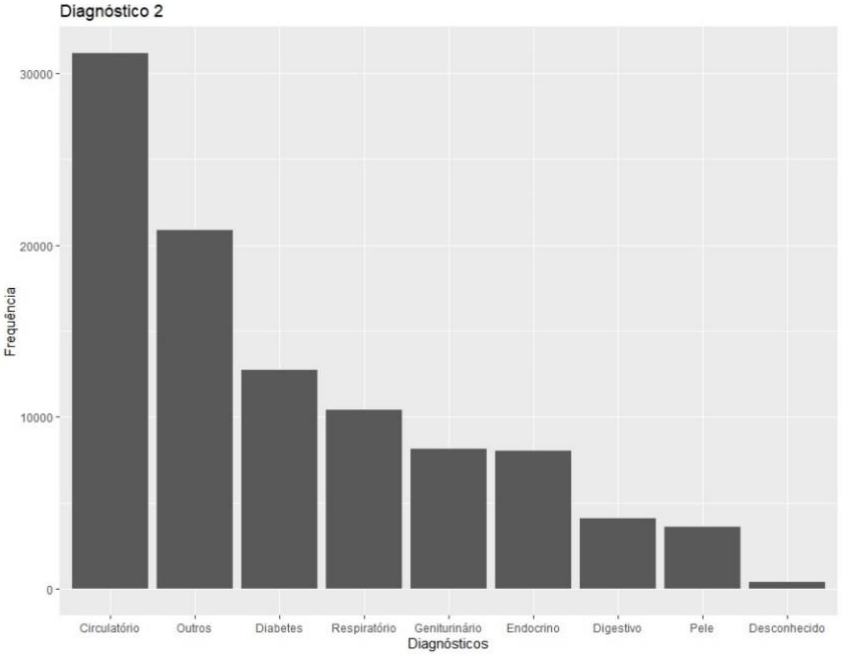
Tabela 6 - Categorização da variáveis diagnóstico 1, 2 e 3.

Código ICD9	**Nome do Grupo**	**Descrição**
390–459, 785	Circulatório	Doenças do sistema circulatório
460–519, 786	Respiratório	Doenças do sistema respiratório
520–579, 787	Digestivo	Doenças do sistema digestivo
250.xx	Diabetes	Diabetes mellitus
800–999	Lesões	Lesões e envenenamentos
710–739	Musculoesquelético	Doenças do sistema músculo-esquelético e do tecido conjuntivo
580–629, 788	Geniturinário	Doenças do aparelho geniturinário
140–239	Neoplasias	Neoplasias
780, 781, 784, 790–799	Outros sintomas	Outros sintomas, sinais e condições mal definidas
240–279, exceto 250	Endócrino	Doenças endócrinas, nutricionais e metabólicas e distúrbios da imunidade, excluindo a diabetes.
680–709, 782	Pele	Doenças da pele e tecido subcutâneo
001–139	Infeciosas e Parasitárias	Doenças infeciosas e parasitárias
290–319	Mental	Doenças mentais
E–V	Causas Externas	Causas externas de lesão e classificação suplementar
280–289	Sangue	Doenças do sangue e órgãos formadores de sangue
320–359	Sistema Nervoso	Doenças do sistema nervoso
630–679	Complicações gravidez	Complicações da gravidez, parto e puerpério

360–389	Órgãos Sensoriais	Doenças dos órgãos sensoriais
740–759	Congênitas	Anomalias congênitas

Na Figura 17 é possível observar o resultado da categorização das referidas variáveis. Em todos os diferentes diagnósticos, o grupo das doenças circulatórias foi o mais evidenciado. Sendo que o conjunto de todas as outras categorias com uma percentagem mais reduzida, agrupada na categoria 'Outros', ficou em segundo lugar. No entanto, a terceira categoria mais evidente difere do diagnostico principal para os secundários. No diagnóstico principal, segue-se a categoria das doenças do foro respiratório. Nos diagnósticos secundários, observa-se uma maior incidência da doença da diabetes. De referir ainda que no diagnóstico principal se verifica uma presença mais significativa da categoria 'Lesão', 'Musculosquelética' e 'Neoplasias', face aos restantes diagnósticos em que estas categorias ficaram agrupadas na categoria 'Outros', por apresentarem menor número. As doenças endócrinas, nutricionais e metabólicas e distúrbios da imunidade, excluindo a diabetes, foram codificadas em maior número como diagnóstico secundário e terciário. As doenças relacionadas com a pele tiveram uma maior evidência no diagnóstico secundário e as causas externas no diagnóstico terciário.


Diagnóstico 1
30000
20000
10000
0
Frequência
Circulatório
Outros
Respiratório
Digestivo
Diabetes
Lesão
Musculoesquelético
Geniturinário
Neoplasias
Desconhecido
Diagnósticos


Diagnóstico 2
30000
20000
10000
0
Frequência
Circulatório
Outros
Diabetes
Respiratório
Geniturinário
Endocrino
Digestivo
Pele
Desconhecido
Diagnósticos

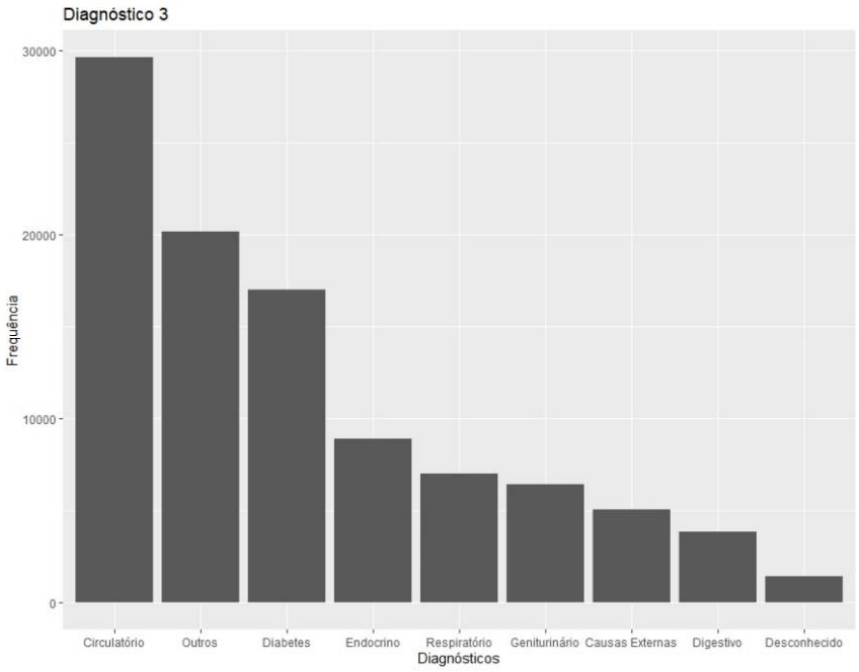


Figura 17 - Frequência de diagnósticos 1, 2 e 3 categorizados.

De ressalvar que em todos os tipos de diagnóstico, a diabetes é das doenças mais predominantes. No diagnóstico principal, a doença da diabetes teve uma frequência percentual de 8,7%. Já no diagnóstico secundário e terciário obteve respetivamente 12,8% e 17,1%. O que se verifica que houve mais casos de internamento em que a doença da diabetes não foi o motivo principal do episódio.

A variável especialidade médica para além de, como mencionado anteriormente, ter alguma percentagem relevante de dados desconhecidos, tem muitos níveis associados (73). Posto isto, optou-se por agrupar as subcategorias em categorias mais gerais de forma a diminuir o número de níveis diferentes. Por exemplo, inicialmente o *dataset* estava composto com a distinção da categoria cardiologia, com cardiologia-pediátrica. No entanto, a especialidade médica, embora para doentes alvo diferentes, é a mesma. Para além de que conseguimos distinguir ambos os grupos pelo fator idade. Por este motivo, todas as categorias da especialidade pediatria foram agrupadas nos grupos superiores. Além disso, existia muita especificidade nas especialidades médicas, de que é a cirurgia que

estava composta por cirurgia cardiovascular, torácica, maxilofacial, vascular, etc. em que se agrupou tudo numa única categoria – cirurgia. Após este tratamento de dados, a variável especialidade médica ficou reduzida a 45 níveis diferentes.

4.5 Avaliação de Padrões

Na Figura 18 está representada a distribuição do número de dias de internamento. Verifica-se que existe um aumento crescente de episódios em que a duração do internamento vai desde 1 aos 3 dias, e a partir daqui existe uma diminuição acentuada até aos 9 dias, estabilizando depois até aos 14 dias – limite máximo de dias de internamento da amostra em estudo. A média e a mediana de dias de internamento, é de aproximadamente 4 dias.

Relativamente ao número médio de internamentos por paciente, é de 1,4 vezes. Sendo que os pacientes com maior número de internamentos teve 40 episódios diferentes.

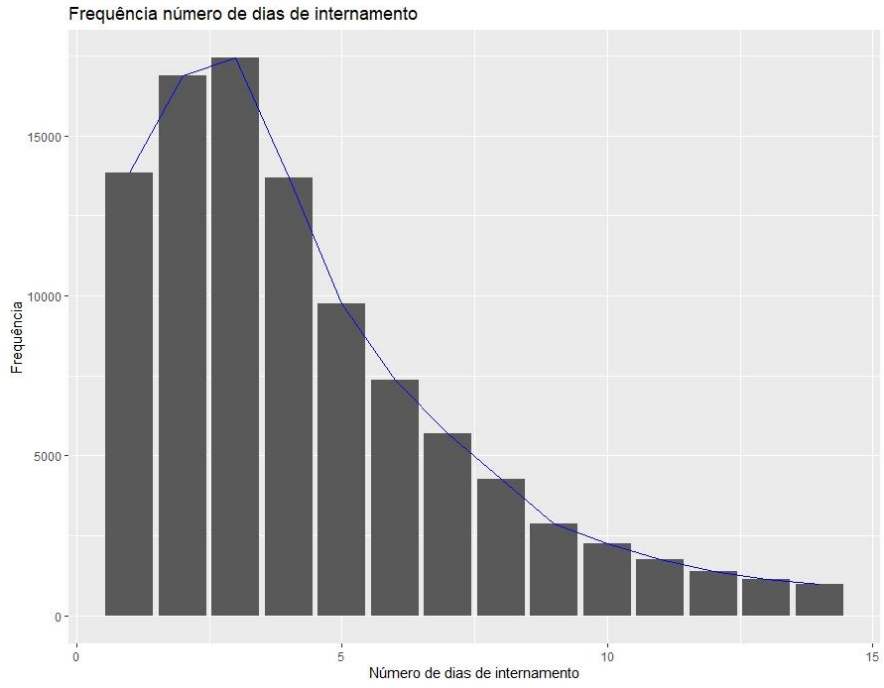


Figura 18 - Frequência número de dias de internamento.

Na Figura 19 é possível analisar a distribuição do número de dias de internamento sob diversas categorias: idade, sexo e etnia. Verifica-se que a distribuição do sexo feminino é semelhante à do sexo masculino. É possível ainda verificar uma diminuição do número de dias de internamento com a diminuição da idade dos pacientes.

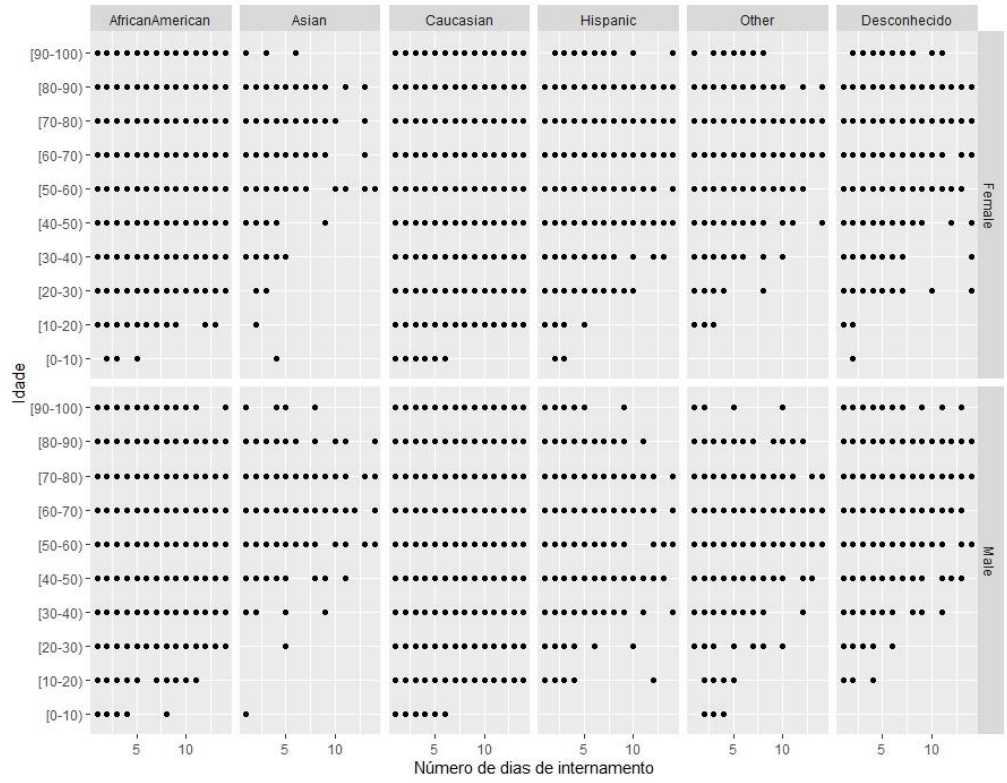


Figura 19 - Análise do número de dias de internamento por idade, sexo e etnia.

Em aproximadamente 95% dos casos de internamento em estudo, não foi registado o valor da medição da glicose. Com o objetivo de perceber se a medição da glicose aquando o internamento do paciente, teve influência no número de dias de internamento, foi produzido o gráfico da Figura 20. Embora com ligeiras diferenças, a distribuição do número de dias de internamento de ambos os grupos

é semelhante. No grupo em que foi medida a glicose, observa-se um aumento mais acentuado da frequência dos dois para três dias.

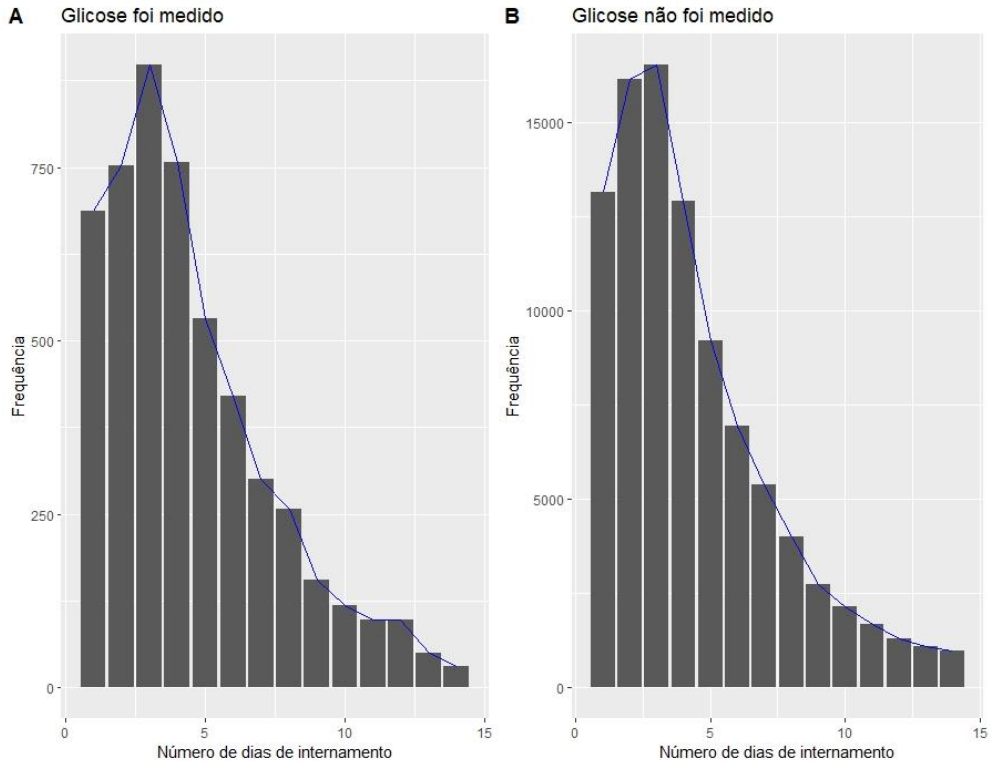


Figura 20 -Distribuição do número de dias de internamento. Internamentos em que a glicose foi medida versus não foi medida.

O gráfico da Figura 21 mostra a distribuição do número de dias de internamento dos episódios em que a glicose foi medida no primeiro contacto com o paciente, comparativamente aos episódios em que este parâmetro não foi registado. Verifica-se que em metade dos casos em que este parâmetro foi registado, o valor identificado foi classificado como normal. Cerca de 28% foi classificado como > 200 mg/dL e os restantes, 23%, como > 300 mg/dL. Sendo que, as pessoas sem diabetes devem ter um valor de glicemia entre 80 e 110 mg/dL antes das refeições e até 140 mg/dL depois das refeições (APDP –

Associação Protectora dos Diabéticos de Portugal, 2020). Pelo que, as duas categorias fora do âmbito considerado normal, pode significar que a pessoa é diabética.

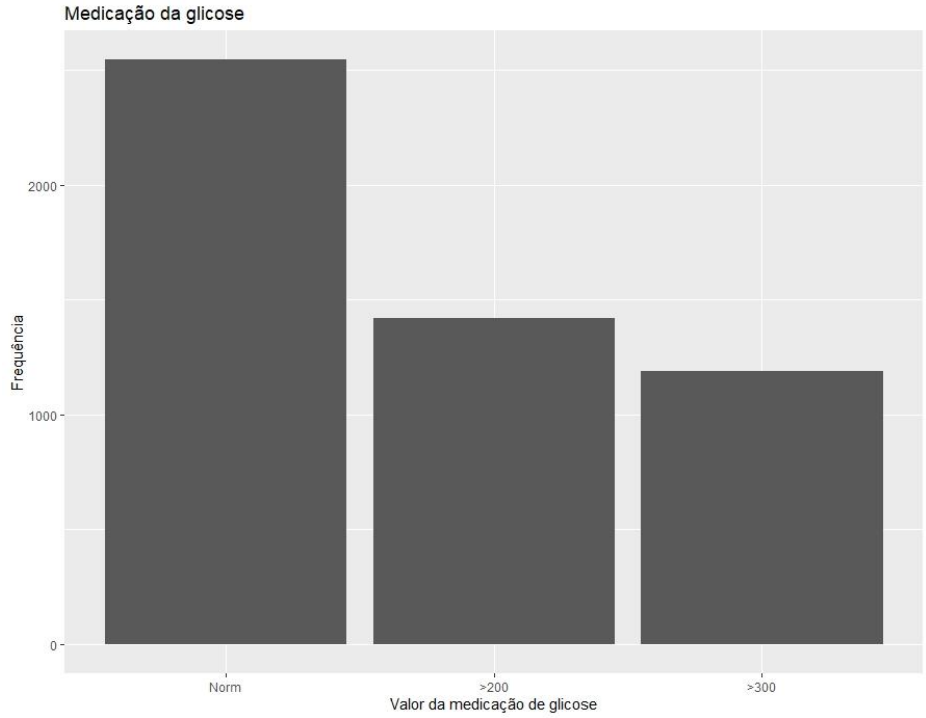


Figura 21 -Frequência dos valores de glicose medidos.

Com o objetivo de saber a distribuição do número de dias de internamento, tendo em consideração o valor da medição da glicose, é apresentado o gráfico da Figura 22. Percebe-se que há uma maior irregularidade no período de internamento, quando a medição foi registada como normal. Para as três categorias, a maior predominância de internamentos encontra-se abaixo dos 5 dias. No entanto, é ainda possível constatar que com o aumento da concentração da glicose no sangue, há também um aumento de concentração de internamentos para períodos mais prolongados.

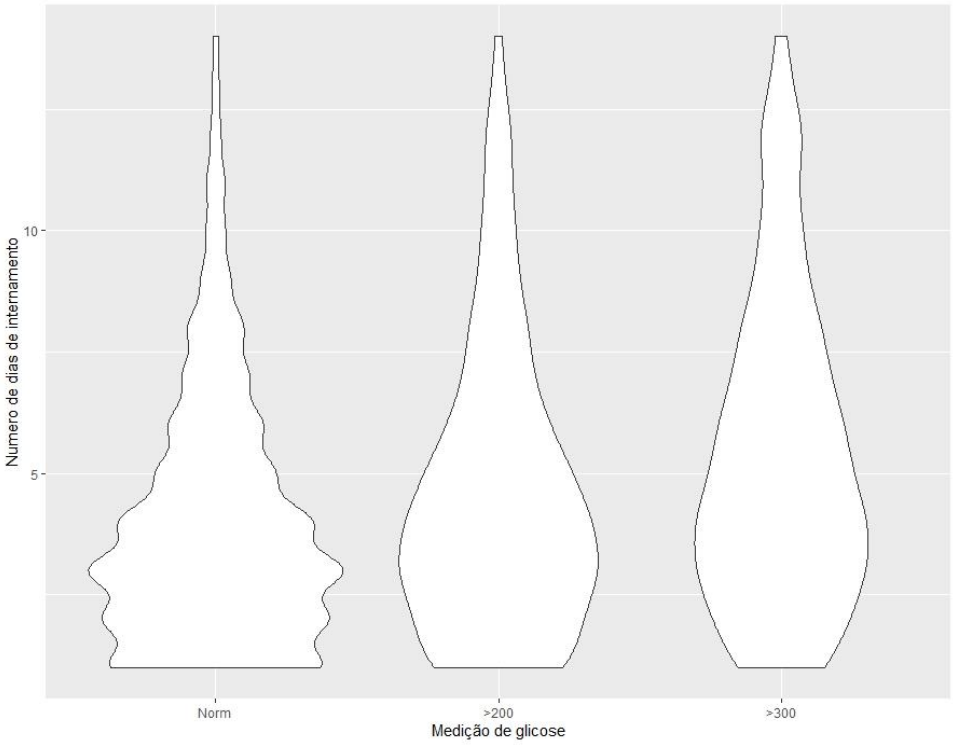


Figura 22 - Distribuição do número de dias de internamento, por cada categoria da medição da glicose.

- **Correlação Linear**

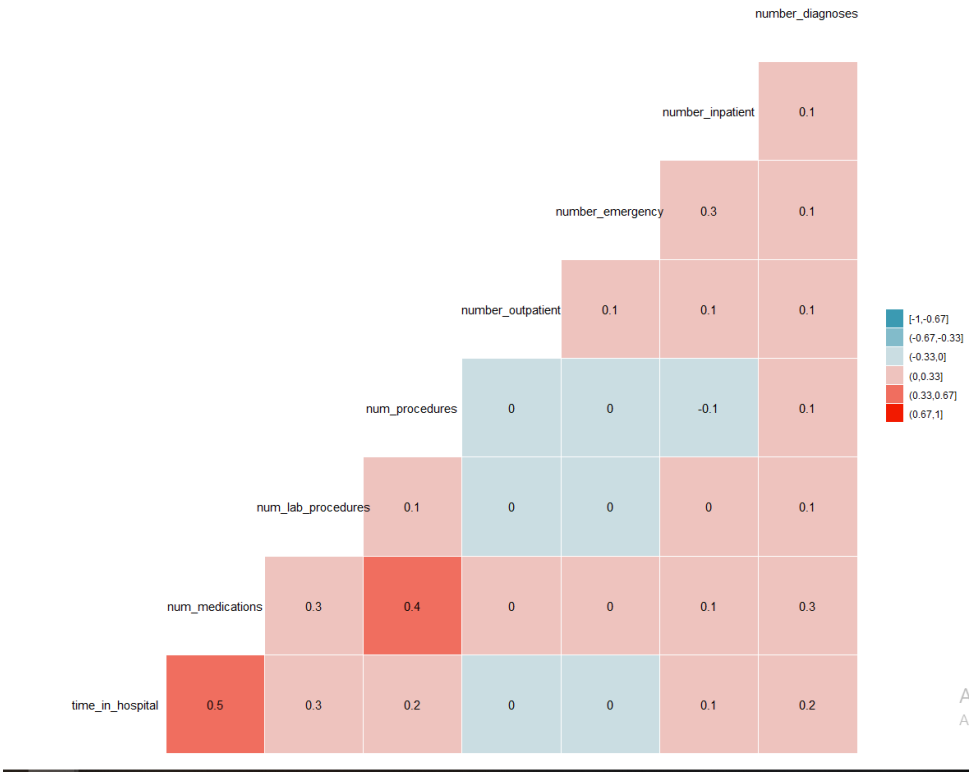
Na imagem da Figura 23 é possível observar uma matriz de correlação das variáveis numéricas. Este gráfico permite visualizar quais as variáveis que têm correlação positiva, negativa, fraca ou forte com as restantes variáveis.

Do resultado apresentado pode-se constatar que as correlações mais fortes se dão entre as variáveis tempo de internamento/número de medicamentos e número de procedimentos/ número de medicamentos, com correlações de 0.5 e 0.4, respetivamente. Ambas as correlações são fortes e positivas, o que significa que à medida que o número de dias de internamento aumenta, há também um aumento do número de medicamentos administrados e, à medida que o número

de procedimentos aplicados no internamento aumenta também aumenta o número de medicamentos administrados nesse mesmo internamento.

No que se refere às visitas no ano anterior ao encontro, mas relativas apenas às visitas de internamento, verifica-se uma correlação com o número de dias de internamento, positiva e fraca (0.1). Quanto às visitas anteriores ambulatórias e de emergência, não se observa qualquer corelacionamento (0). As variáveis número de procedimentos e número de diagnósticos, têm uma correlação ligeiramente superior (0.2), predizendo que internamentos mais prolongados têm tendência a ter um maior número total de doenças/problemas diagnosticados e um maior número de procedimentos. Existe ainda uma observação pertinente relativamente aos resultados obtidos, uma vez que a correlação com número de procedimentos laboratoriais é superior (0.3), o que significa que internamentos hospitalares mais prolongados, em geral têm um maior número de procedimentos laboratoriais aplicados do que os não laboratoriais.

Verifica-se uma correlação mais forte do número de medicamentos diferentes administrados com o período total de internamento (0.5). O que já seria de esperar, porque episódios com internamentos mais prolongados, poderão estar associado a indivíduos com maiores problemáticas associadas e que por isso necessitam de um maior número de medicamentos diferentes administrados.



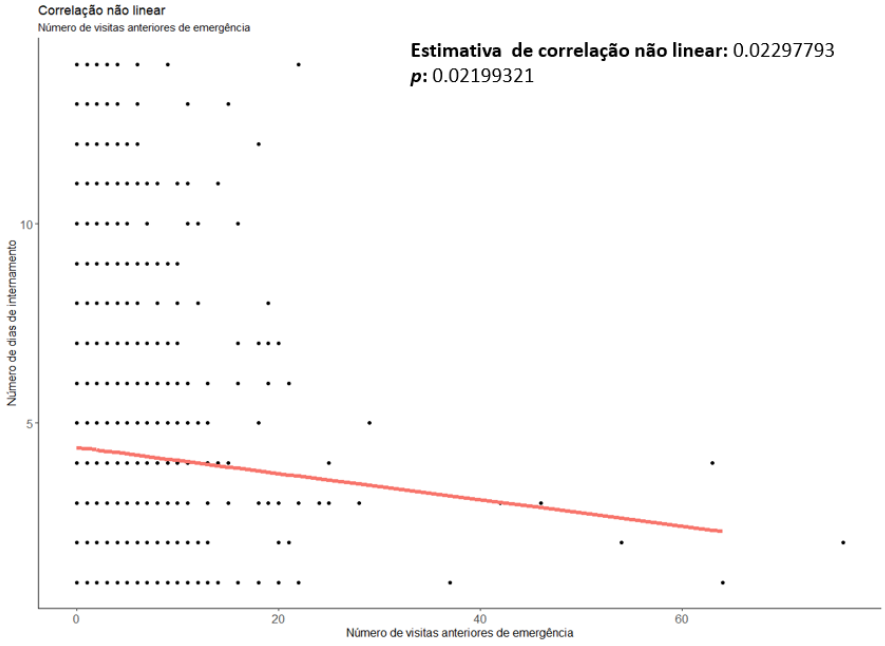
Legenda: Time_in_hospital – Número de dias de internamento; **Num_lab_procedures –** Número de procedimento laboratoriais; **Num_procedures –** Número de procedimentos; **Num_medications** – Número de medicação administrada; **Number_outpatient -** número de consultas ambulatoriais do paciente no ano anterior ao internamento; **Number_emergency** – número de visitas de emergência do paciente no ano anterior ao encontro; **Number_inpatient-** Número de visitas de internamento do paciente no ano anterior ao encontro. **Number_diagnosis** -Número de diagnósticos.

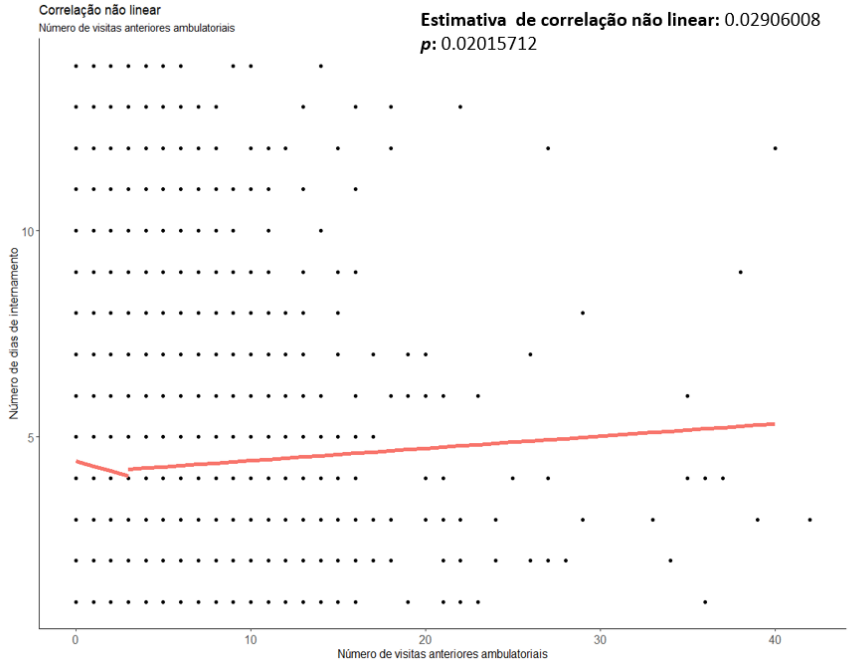
Figura 23 - Matriz de correlação das variáveis numéricas.

- **Correlação não linear**

No método anterior estudou-se as correlações lineares, no entanto, os dados podem ter uma correlação não linear, mas pouca ou nenhuma correlação linear. Por vezes, variáveis não correlacionadas linearmente são negligenciadas durante o tratamento de dados. Para evitar isso, decidiu-se avaliar a existência de correlação das visitas anteriores relativamente à variável número de dias de internamento, dado que se verificou correlação baixa (visitas de internamento) ou nenhuma (visitas de emergência e ambulatórias).

Para esse fim, utilizou-se um estimador de correção não linear do R – nlcor. Esta função devolve a estimativa da correlação não linear, o valor de p correspondente e um gráfico de visualização do corelacionamento não linear. Se $p>0.05$, a estimativa de correlação não linear pode ser considerada como ruído, ou estatisticamente não significativa. Na Figura 24 é possível observar as correlações não lineares da variável número de dias de internamento com o número de visitas anteriores. Para todas as representações, obtiveram-se resultados estatisticamente significativos ($p < 0.05$). Para as visitas anteriores de internamento prevê-se uma relação não linear positiva e fraca, tal como a identificada no corelacionamento linear ($\cong 0.1$). Para ambos os corelacionamentos de emergência e ambulatórias, verifica-se uma estimativa de correlação não linear bastante baixa ($\cong 0.02$ e $\cong 0.03$, respetivamente). No entanto, por existir algo tipo de relacionamento e serem estatisticamente significativos, decidiu-se manter estas variáveis em estudo, por poderem conter informações importantes.


Correlação não linear
Número de visitas anteriores de emergência
Estimativa de correlação não linear: 0.02297793
p: 0.02199321
Número de dias de internamento
10
5
0
20
40
60
Número de visitas anteriores de emergência


Correlação não linear
Número de visitas anteriores ambulatoriais
Estimativa de correlação não linear: 0.02906008
p: 0.02015712
Número de dias de internamento
10
5
0
10
20
30
40
Número de visitas anteriores ambulatoriais

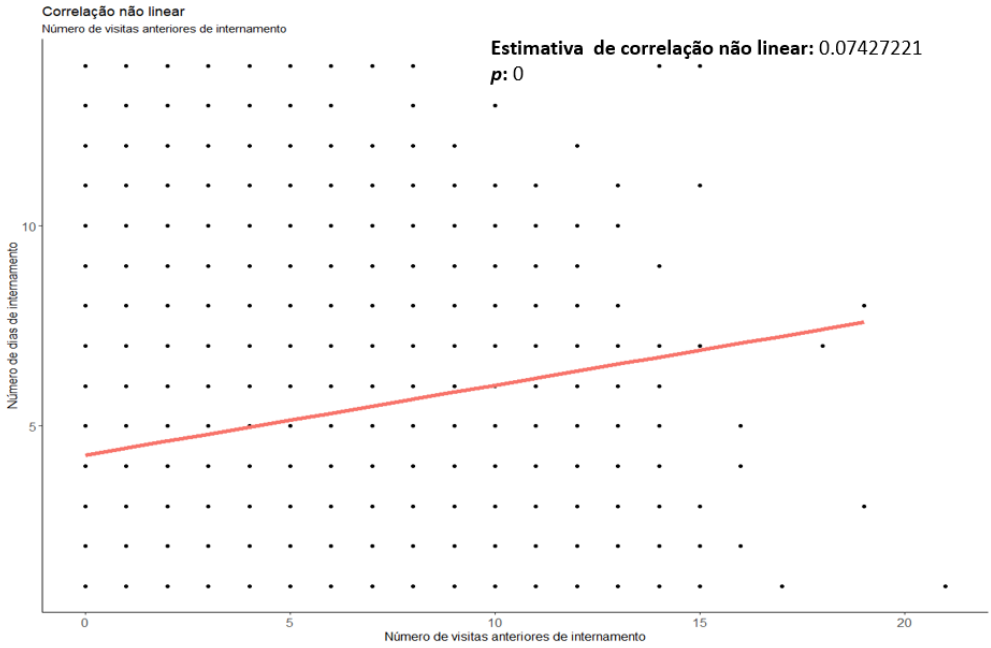


Figura 24 - Representação das correlações não lineares do número de dias de internamento com o número de visitas anteriores.

4.6 Seleção de Variáveis Conhecidas

Como já mencionado anteriormente, o objetivo principal deste trabalho é perceber e estudar o que leva indivíduos cujo diagnóstico hospitalar seja a doença da diabetes a ter um período de internamento total mais prolongado. E, por conseguinte, conseguir desenvolver um método eficaz que possa de futuro auxiliar os prestadores de cuidados de saúde a prever o período de internamento de um paciente. Desta forma, é importante que as variáveis em estudo sejam aquelas em que estejam disponíveis logo no primeiro contacto com o paciente de forma a que a previsão possa ser feita logo no início do período de internamento aumentando a vantagem da sua utilização.

Existem, por isso, algumas variáveis na nossa amostra que foram desconsideradas por apenas se poderem obter essa informação no final do internamento. Num primeiro contacto com o paciente, o prestador de cuidados de saúde desconhece o número de medicamentos totais diferentes que irão ser administrados, o número de procedimentos médicos aplicados, ou o destino após alta. Selecionou-se por isso apenas as variáveis que são conhecidas na triagem com paciente, que estão representadas a seguir:

Obtiveram-se 15 variáveis diferentes que serão usadas daqui em diante no estudo para previsão do número de dias de internamento.

5. Resultados e Discussão

Até ao momento foram aplicados alguns métodos de processamento dos dados e realizadas algumas análises estatísticas e avaliação de padrões. Pretende-se, neste capítulo, fazer a exploração dos dados com técnicas mais avançadas. Decidiu-se abordar o problema como classificação, reduzindo os níveis diferentes da variável número de dias de internamento uma vez que a análise por regressão poderia fornecer mais informação do que o necessário. Do ponto de vista clínico, o excesso de informação poderia aumentar a dúvida sobre as medidas a adotar para cada caso e as informações obtidas na tarefa de classificação seriam suficientes para os prestadores de cuidados de saúde obterem uma indicação de que se o paciente estaria ou não na zona de risco, transmitindo-lhes a informação da existência de algo potencialmente problemático para o paciente, caso existisse (Andersson, 2019). Neste sentido e dada a abordagem de classificação serão realizadas duas abordagens diferentes:

- Regras de associação;
- Algoritmo de Random Forest.

5.1 Estrutura de Classificação

Com o objetivo de identificar as categorias principais do número de dias de internamento, tentou-se identificar os potenciais pontos de mudança principais desta variável. Os pontos de mudança referem-se às variações nos dados de séries temporais. Esta metodologia é utilizada em diversas áreas, desde deteção de mudanças climáticas, reconhecimento da fala, análise de imagem e de voz (Daniel Harris, BA, Lynn McNicoll, MD, Gary Epstein-Lubow, MD, and Kali S. Thomas, 2017).

5.1.1 Ponto de Mudança – *Change Point*

Por forma a encontrar os pontos de mudança da curva de distribuição do número de dias de internamento dos episódios em estudo, aplicaram-se dois algoritmos diferentes – `cpt.mean` e `cpt.meanvar`, da biblioteca *changepoint*. Ambas as funções têm o objetivo de identificar pontos de mudança, com a particularidade de que a função cpt.mean identifica alterações tendo em conta a média dos dados, ao passo que a função cpt.meanvar deteta as alterações nos dados tendo em conta a média e a variância dos mesmos. Em ambas as funções as alterações são encontradas usando o método fornecido, que pode ser único ponto de mudança (AMOC) ou vários pontos de mudança usando métodos exatos (PELT ou SegNeigh) ou aproximados (BinSeg). Um ponto de mudança é indicado como a primeira observação do novo segmento / regime.

Na Figura 25 é possível observar 4 formas diferentes de distribuição da variável número de dias de internamento, usando as duas diferentes funções e com métodos e/ou testes estatísticos diferentes. O primeiro gráfico da Figura 25 apresenta alterações na média dos dados, a partir de um único ponto (método AMOC), com um teste estatístico normal. A amostra ficou dividida no 5º dia, tendo consequentemente gerado dois grupos: período de internamento curto (1-5 dias) e período de internamento longo (6-14 dias). O segundo gráfico foi também gerado olhando à variação na média dos dados, no entanto, com um método diferente, BinSeg. Foram obtidos mais pontos de mudança relativamente à simulação anterior: 3, 4, 5, 7 e 8. Desse modo, considerou-se dividir a amostra em 4 diferentes categorias – internamento curto (1-3 dias), internamento médio (4-5 dias), internamento médio-longo (6-7dias) e internamento longo (8-14 dias).

A segunda função foi utilizada com o método BinSeg com uma distribuição estatística normal, que implementa o método de segmentação binária para identificar pontos de mudança. O resultado foi a obtenção de 2 pontos de mudança diferentes (4 e 8), o que resultou na divisão em 3 grupos diferentes: internamento curto (1-4 dias), internamento médio (5-8 dias) e internamento longo (9-14 dias).

Na quarta e última representação, embora tenha sido usada a mesma função e o mesmo método anterior (BinSeg), foi assumida uma distribuição estatística do teste diferente – Poisson. Chegou-se ao resultado de 3 pontos de mudança diferentes (4, 6 e 9), o que resultou na divisão em 4 grupos diferentes:

internamento curto (1-4 dias), internamento médio (5-6), internamento médio-longo (7-9) e internamento longo (10-14 dias).

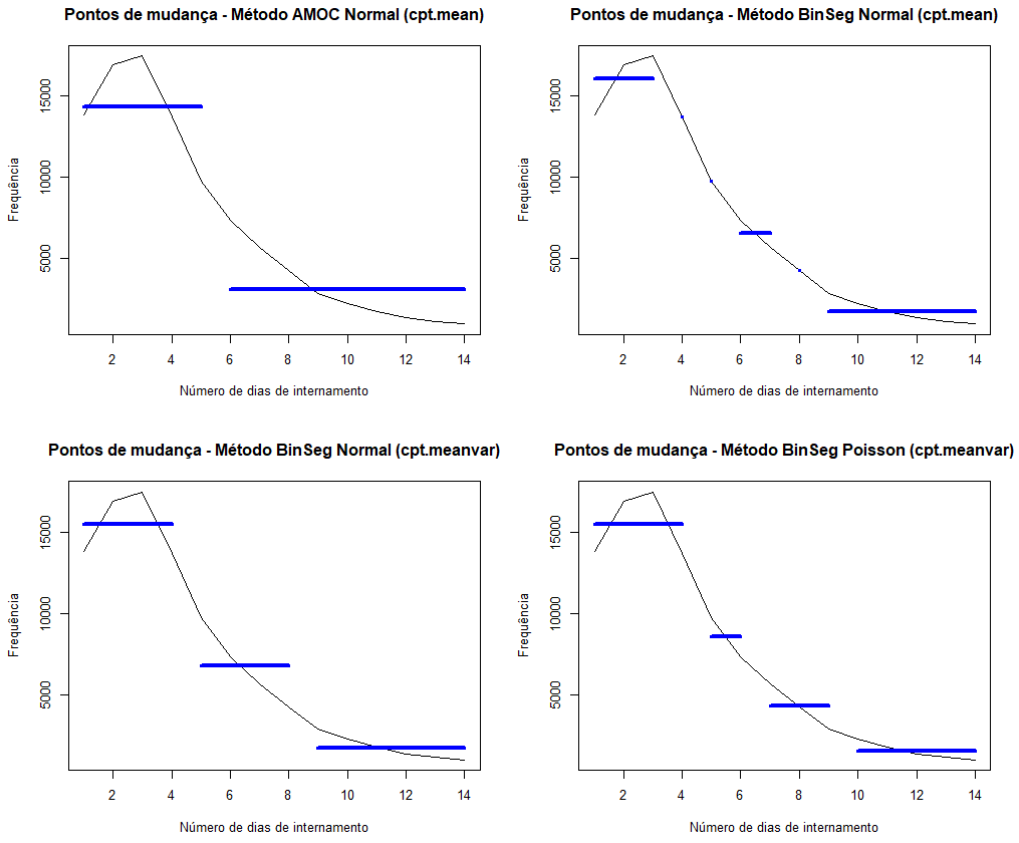


Figura 25 - Análise dos pontos de mudança da curva de distribuição do número de dias de internamento, por diversas técnicas.

A Tabela 7 apresenta a distribuição de cada grupo de acordo com os pontos de mudança para cada método/distribuição estatística. Verifica-se que o grupo de internamento curto é, para todos os métodos, o que tem mais peso percentual na nossa amostra. O método AMOC, foi o que apresentou uma maior discrepância entre internamento curto versus restantes períodos de internamento. No entanto, também foi para este método que se observou um maior peso percentual dos internamentos longos. O método BinSeg Poisson

obteve alguns grupos com peso percentual bastante reduzido, de que é exemplo o internamento longo que apenas atingiu 8% da amostra. Perante isto, o método BinSeg, com teste estatístico normal, pela função `cpt.mean`, foi a que obteve os grupos mais equilibrados percentualmente. Desta forma, decidiu-se prosseguir com esta distribuição categórica e com o método de AMOC, por se observar uma maior incidência de internamentos longos. No entanto, para este último método foi necessário fazer uma seleção aleatória de internamentos curtos para equilibrar ambos os grupos, explicado mais à frente - 5.1.1 Discretização de variáveis.

Tabela 7 - Distribuição percentual de cada um dos grupos para os diversos métodos de análise de pontos de mudança.

	Internamento Curto	**Internamento Médio**	**Internamento Médio-Longo**	**Internamento Longo**
Método AMOC (cpt.mean)	72%	-	-	28%
Método BinSeg Normal (cpt.mean)	48%	23%	13%	15%
Método BinSeg Normal (cpt.meanvar)	62%	27%	-	11%
Método BinSeg Poisson (cpt.meanvar)	62%	17%	13%	8%

5.1.1 Discretização de variáveis

Com a finalidade de extrair informações úteis à cerca da amostra em estudo, decidiu-se aplicar o método de regras de associação. Por este motivo foi necessário converter as variáveis numéricas em variáveis categóricas, utilizando para isso a mesmas metodologias selecionadas para divisão da variável número de dias de internamento – Método AMOC (cpt.mean) e Método BinSeg Normal (cpt.mean). Na Figura 26 e na Tabela 8 é possível observar os intervalos obtidos e as categorizações atribuídas, respetivamente pelo método BinSeg Normal

(cpt.mean) em que a variável número de dias de internamento contem quatro níveis diferentes. Na tabela 8 pode ver-se entradas com valores arredondados à unidade e, por isso, alguns valores tomam valor de 0. No entanto, com mais casas decimais, pode-se verificar que esses valores são ligeiramente superiores a 0.

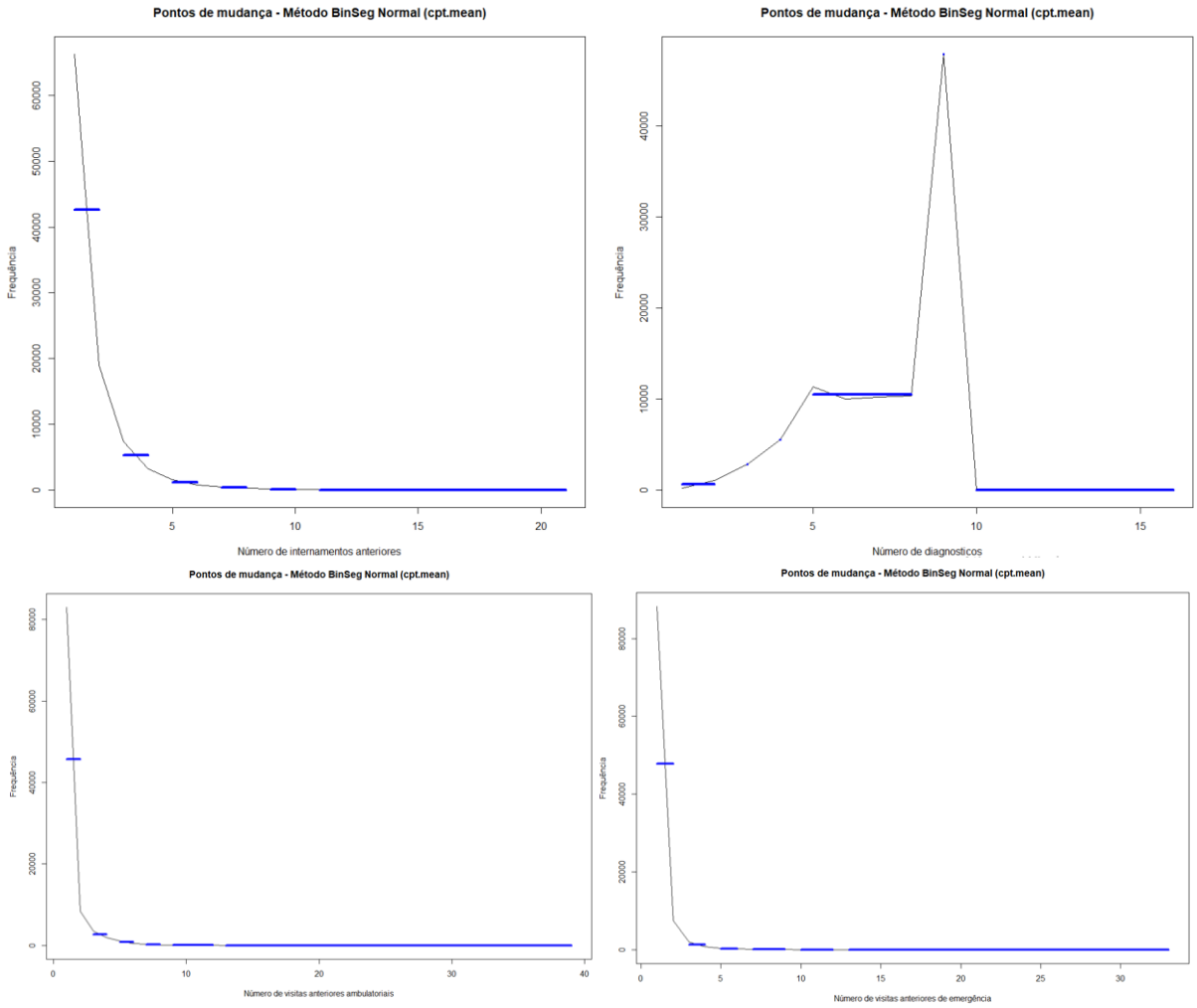
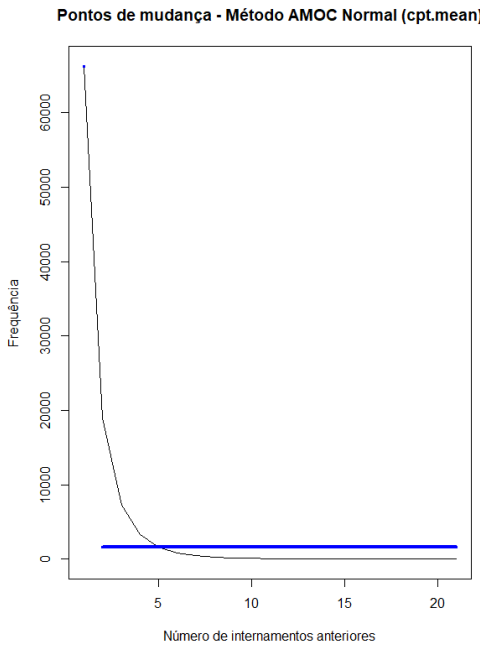


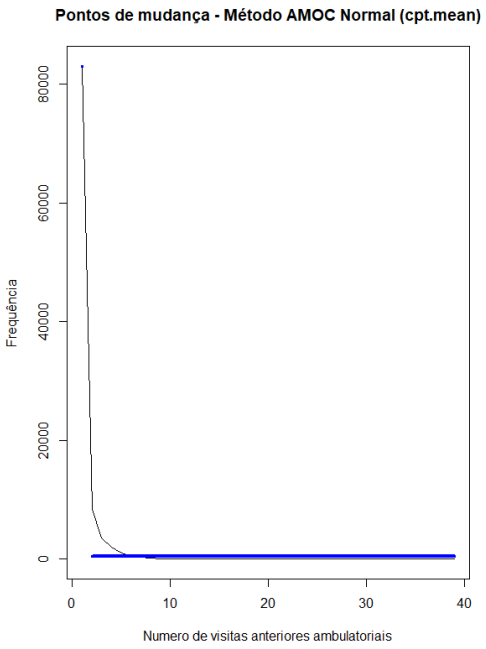
Figura 26 - Discretização das variáveis numéricas – Método BinSeg Normal (cpt.mean).

Tabela 8 - Categorização dos intervalos das variáveis – Método BinSeg Normal (cpt.mean).

Número de internamentos anteriores	0-2	93%	Baixo
	3-4	5%	Intermédio
	5-6	1%	Médio
	7-8	0%	Alto
	9-10	0%	Elevado
	>=11	0%	Muito elevado
Número de diagnósticos	0-2	1%	Reduzido
	3	3%	Intermédio
	4	6%	Médio
	5-8	42%	Alto
	>=9	48%	Elevado
Número de visitas anteriores ambulatoriais	0-2	96%	Baixo
	3-4	3%	Intermédio
	5-6	1%	Médio
	7-8	0%	Alto
	>= 9	0%	Elevado
Número de visitas anteriores de emergência	0-2	98%	Baixo
	3-4	2%	Intermédio
	5-6	0%	Médio
	7-9	0%	Alto
	>= 10	0%	Elevado

Na Figura 27 e na Tabela 9 é possível observar os intervalos obtidos e as categorizações atribuídas, respetivamente pelo método AMOC Normal (cpt.mean) em que a variável número de dias de internamento foi tratada como um problema binomial. Com este método e como já tinha sido mencionado anteriormente, existia um grande desequilíbrio percentual entre o número de internamentos curtos e longo. Pelo que foi necessário primeiramente equilibrar a amostra. Existiam 78941 episódios curtos para 20407 internamentos longos. Pelo que se decidiu selecionar de forma aleatória, pelo método `sample_n` do pacote `dplyr` do R, apenas 25 000 amostras de internamento curto, dado que o desequilíbrio de classes é problemático com classificadores e métricas de classificação.





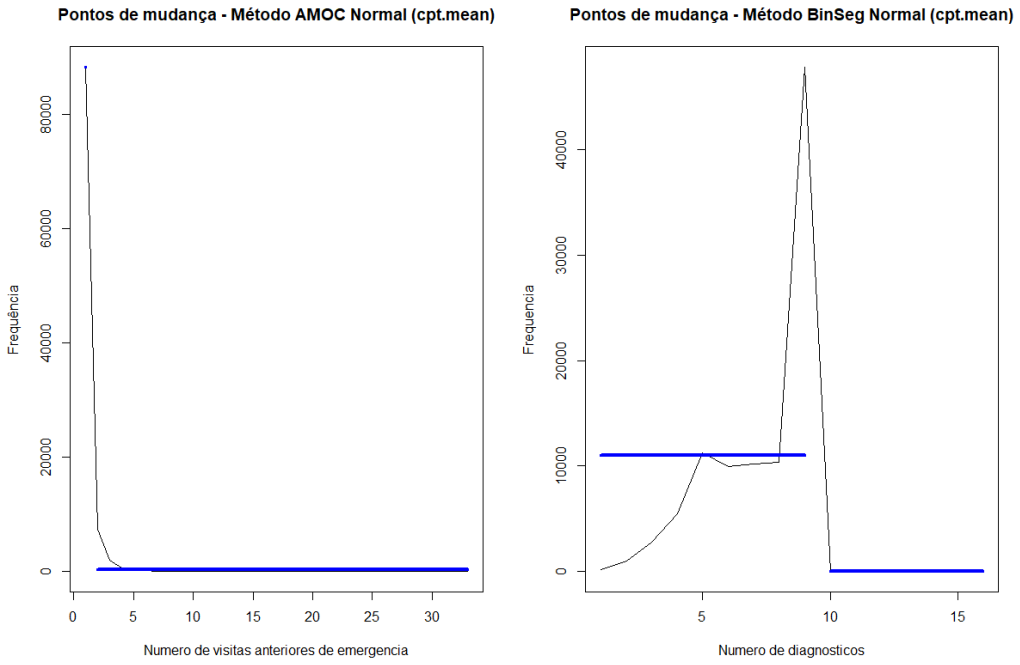


Figura 27 - Discretização das variáveis numéricas – Método AMOC Normal (cpt.mean).

Tabela 9 - Categorização dos intervalos das variáveis - Método AMOC Normal (cpt.mean).

Número de internamentos anteriores	0-1	85%	Reduzido
	>=2	15%	Alto
Número de diagnósticos	0-9	100%	Reduzido
	>=10	0%	Alto
Número de visitas anteriores ambulatoriais	0-1	92%	Reduzido
	>=2	8%	Alto
Número de visitas anteriores de emergência	0-1	96%	Reduzido
	>=2	4%	Alto

5.2 Estudo das regras de associação

De forma a identificar as regras de associação da amostra em estudo, aplicou-se o algoritmo apriori disponível no R, do package adicional arules. O processo do R ressume-se no seguinte processo: o algoritmo percorre cada transação do conjunto de dados, isto é, cada linha do conjunto. De seguida, gera as regras que respeitem os limites mínimos definidos de suporte e confiança. As regras obtidas neste trabalho têm como consequentes as diversas variáveis da amostra. Embora sejam todas relevantes por nos fornecer informações úteis da amostra, é mais relevante para o estudo em questão perceber quais os antecedentes que levam a que o consequente seja um período de internamento maior ou menor. Fez-se, portanto, subconjuntos das regras em que o consequente fosse a variável tempo de internamento. De forma a determinar os melhores parâmetros de confiança e suporte de forma a identificar as regras mais importantes da amostra, fez-se um conjunto de testes que permitiram avaliar o número de regras obtidas para um conjunto de valores possíveis de suporte e confiança.

Nas tabelas e figuras seguintes é possível verificar para cada categoria e para cada metodologia de divisão, o número de regras obtidas para vários possíveis valores de suporte e de confiança. Subdividiu-se o número de regras para valor total e filtrado. Os valores totais são referentes ao número total de regras obtidas sem qualquer seleção. Os valores filtrados correspondem ao número de regras obtidas após a remoção de regras redundantes e/ ou regras estatisticamente não significativas. Regras redundantes são regras que não trazem mais informação em relação à já existente noutras regras. Ou seja, são regras que são iguais ou que têm os mesmos itens do lado direito, mas um ou mais itens do lado esquerdo. Para deteção das regras redundantes utilizou-se a função `is.redundant()`. Quanto às regras estatisticamente não significativas, utilizou-se a função `is.significant()` que usa o teste de Fisher com a finalidade de identificar as regras em que a probabilidade de ocorrência de itens à esquerda e à direita é estatisticamente dependente.

- **Método BinSeg Normal (cpt.mean) – Variável número de dias de internamento como um problema quadrinomial.**

Tabela 10 - Número de regras obtidas (total e após a filtragem), para diferentes valores de suporte e confiança - Classificação da variável número de dias de internamento em 4 categorias.

Valores limite de suporte e confiança	**Número de regras –** Internamento Curto		**Número de regras –** Internamento Médio		**Número de regras –** Internamento Médio- Longo		**Número de regras –** Internamento Longo	
	Total	Filtrado	Total	Filtrado	Total	Filtrado	Total	Filtrado
Sup. 0.1 Conf. 0.6	0	0	0	0	0	0	0	0
Sup.0.01 Conf. 0.6	14635	0	0	0	0	0	0	0
Sup.0.01 Conf.0.5	76878	0	0	0	0	0	0	0
Sup.0.001 Conf 0.5	1834615	53847	0	0	0	0	2208	0
Sup.0.001 Conf.0.4	2955040	54377	15	0	0	0	4034	0
Sup.0.001 Conf.0.3	3322934	54377	49515	6557	0	0	13751	0
Sup.0.001 Conf. 0.20	3334205	54377	1245593	63104	4255	24	147447	2463

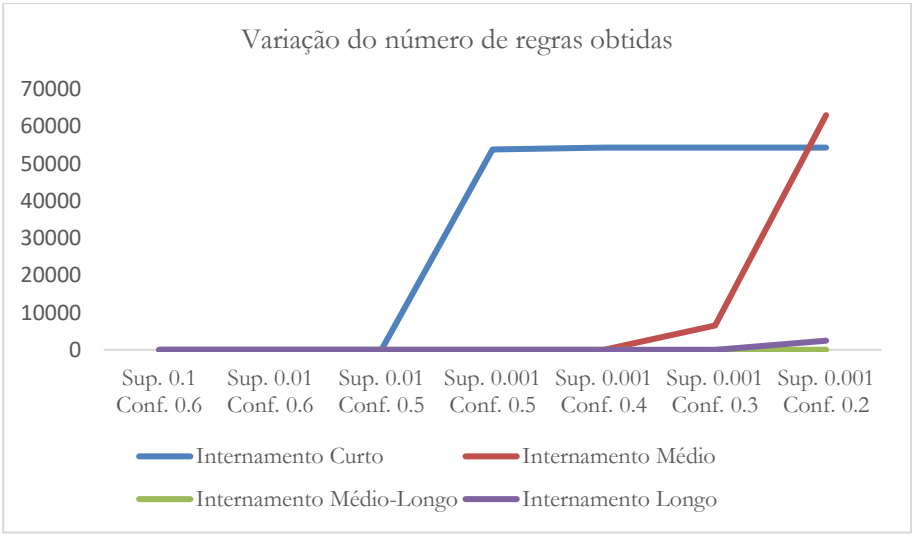


Figura 28- Número de regras obtidas (após a filtragem), para diferentes valores de suporte e confiança - Classificação da variável número de dias de internamento em 4 categorias.

- **Método AMOC Normal (cpt.mean) – Variável número de dias de internamento como um problema binomial**

Tabela 11- Número de regras obtidas (total e após a filtragem), para diferentes valores de suporte e confiança - Classificação da variável número de dias de internamento em 2 categorias.

Valores limite de suporte e confiança	**Número de regras –** Internamento Curto		**Número de regras –** Internamento Longo	
	Total	Filtrado	Total	Filtrado
Sup.0.1 Conf.0.6	0	0	0	0
Sup. 0.01 Conf.0.6	8867	334	288	0
Sup. 0.01 Conf. 0.5	31106	1138	4156	66
Sup.0.001 Conf. 0.5	474800	22233	180642	12615
Sup. 0.001 Conf. 0.4	573347	22233	376866	14456
Sup. 0.001 Conf. 0.3	587723	22233	459586	14456
Sup. 0.001 Conf. 0.20	588919	22233	472006	14456

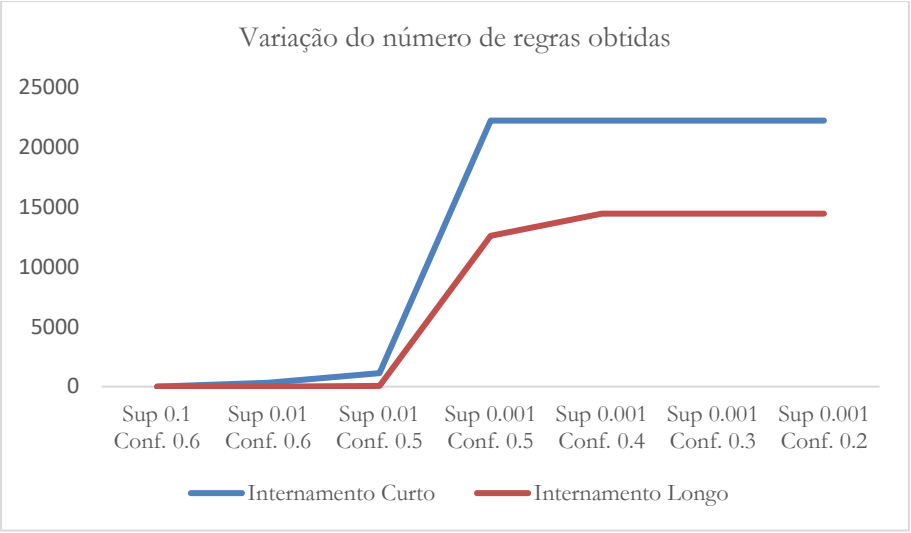


Figura 29 - Número de regras obtidas (após a filtragem), para diferentes valores de suporte e confiança - Classificação da variável número de dias de internamento em 2 categorias.

5.2.1 Estudo das regras de associação frequentes

Para divisão das variáveis numéricas em 2 categorias distintas, obteve-se mais resultados para melhores valores de suporte e confiança. Enquanto que para a divisão em 4 categorias só se obtiveram regras para todas as categorias para valor de confiança de 0.2 e de suporte de 0.001, na divisão em apenas 2 categorias foi possível obter regras de associação para um valor de suporte mais alto 0.01 e um nível mais elevado de confiança, 0.5 (Tabela 10, Figura 28 e Tabela 11, Figura 29). Por este motivo, decidiu-se prosseguir o estudo com esta metodologia de classificação.

O valor do suporte permite identificar as regras que se devem considerar para uma análise mais aprofundada. Neste caso, um valor de suporte de 0.01 significa que as regras geradas dizem respeito a transações que correspondem a pelo menos 1% das transações totais (45407). É importante ressalvar que o valor de suporte é baixo, mas o facto de existirem regras com valor de suporte baixo podem levar a encontrar padrões raros de pacientes, mas não menos importantes.

A confiança de 0.5 significa que para um dado antecedente, o consequente aparece pelo menos 50% das vezes.

Regras de associação obtidas – Consequente Internamento Curto

Nas Tabela 12 e na Tabela 13 são apresentadas as 5 regras obtidas com valores mais altos de suporte e confiança, respetivamente.

Tabela 12 - 5 regras com valores mais altos de suporte para o consequente internamento curto.

[1] {race=Caucasian} => {time_in_hospital=Short internment}

Support: 0.6295
Confidence: 0.5547
Lift: 1.00

[2] {number_diagnoses=Reduced} => {time_in_hospital=Short internment}

Support: 0.5499
Confidence: 0.5506
Lift: 1.00

[3] {admission_source_id=Emergency Room} => {time_in_hospital=Short internment}

Support: 0.3514
Confidence: 0.5955
Lift: 1.00

[4] {admission_type_id=Emergency} => {time_in_hospital=Short internment}

Support: 0.3300
Confidence: 0.5981
Lift: 1.01

[5] {medical_specialty=Desconhecido} => {time_in_hospital=Short internment}

Support: 0.3064
Confidence: 0.5817
Lift: 1.00

Tabela 13 - 5 regras com valores mais altos de confiança para o consequente internamento curto.

[6] {race=Caucasian,
age=[60-70),
admission_type_id=Elective,
diag_1=Circulatório,
number_outpatient=Reduced} => {time_in_hospital=Short internment}

Support: 0.0100
Confidence: 0.7197
Lift: 1.14

[7] {race=Caucasian,
gender=Male,
age=[60-70),
admission_type_id=Emergency,
diag_1=Circulatório,
number_outpatient=Reduced} => {time_in_hospital=Short internment}

Support: 0.0100
Confidence: 0.7193
Lift: 1.14

[8] {admission_type_id=Emergency,
admission_source_id=Physician Referral,
number_inpatient=Reduced,
number_emergency=Reduced} => {time_in_hospital=Short internment}

Support: 0.0106
Confidence: 0.7176
Lift: 1.15

[9] {race=Caucasian,
age=[40-50),
admission_source_id=Physician Referral} => {time_in_hospital=Short internment}

Support: 0.0108
Confidence: 0.6280
Lift: 1.14

[10] {gender=Male,
admission_type_id=Elective,
diag_1=Circulatório,
diag_2=Circulatório} => {time_in_hospital=Short internment}

Support: 0.0109
Confidence: 0.6279
Lift: 1.14

A análise de regras com valores altos de suporte é importante na medida em que nos permite analisar os itens (antecedentes) mais frequentes, mas nem sempre isso pode significar que sejam mais influentes para o consequente em estudo, daí que a análise das regras só faz sentido quando existe a avaliação simultânea da confiança e do lift. Da amostra de regras selecionadas com valores mais altos de suporte obtido verificou-se a aquisição de regras mais curtas, isto é, com apenas um item no antecedente. Para cada uma destas regras o lift obtido foi igual ou aproximadamente igual a 1. Isto significa que o antecedente é independente do consequente. Vejamos o exemplo da regra [1], diz-nos que aproximadamente 63% dos internamentos da amostra em estudo tiveram internamento curto e diziam respeito a indivíduos da raça caucasiana. A confiança da regra indica que a probabilidade de um individuo da raça caucasiana ter tido um internamento curto é de 55%. Já o valor de lift, por ser igual a 1, indica que o antecedente é independente do consequente da regra. Por outras palavras, pode dizer-se que, para a amostra em estudo, ser da raça caucasiana não influência se o internamento hospitalar será curto. Inferir que dado antecedente é independente para determinado consequente pode ser bastante importante para perceber não só o que pode levar ao internamento curto/longo, como também os fatores que não influenciam esse final. Pelo que, e para a amostra em estudo pode-se inferir que ser da raça caucasiana ou, o número de diagnósticos registados ter sido reduzido ou, o tipo/modo de admissão ter sido de emergência, ou a especialidade médica ser desconhecida, não é dependente do consequente internamento curto.

Para o conjunto de regras obtido com maiores valores de confiança, verifica-se um aumento da complexidade das regras, isto é, no número de itens do antecedente. É igualmente notório um aumento no valor de lift para este conjunto de regras selecionado. Analisaremos a regra com maior valor obtido de confiança, a regra [6]. Pode dizer-se que para a amostra em estudo, apenas 1% dos internamentos dizem respeito a indivíduos da raça caucasiana, no intervalo de idades dos 60 aos 70 anos, em que o tipo de admissão foi eletiva, o diagnóstico principal foi do sistema circulatório, o número de consultas ambulatoriais do paciente no ano anterior ao internamento foi reduzido e o internamento hospitalar foi curto. No entanto existe uma probabilidade de 72% (valor de confiança) de que estes fatores ocorram em simultâneo. Existe ainda a indicação de que o internamento curto é 1.14 vezes superior nos indivíduos que cumpram aqueles requisitos referidos anteriormente. O maior valor de lift obtido está incluído no conjunto de regras com maior valor de confiança – regra [8]. Esta

regra indica que o internamento curto é 1.15 vezes superior em internamentos cujo tipo de admissão tenha sido de emergência, a fonte de admissão tenha surgido por referência médica e o número de visitas de internamento e de emergência do paciente no ano anterior ao encontro tenha sido reduzido.

Regras de associação obtidas – Consequente Internamento Longo

A seguir na Tabela 14 e Tabela 15 estão apresentadas as 5 regras obtidas com valores mais altos de suporte e confiança, respetivamente para o conjunto de regras obtidas em que o consequente foi o internamento longo.

Tabela 14 - 5 regras com valores mais altos de suporte para o consequente internamento longo.

[11] { diag_2=Digestivo,
number_emergency=Reduced,
number_outpatient=Reduced } => {time_in_hospital=Long internment}

Support: 0.0183
Confidence: 0.5000
Lift: 1.12

[12] {gender=Female,
age=[70-80),
diag_2=Outros,
number_diagnoses=Reduced} => {time_in_hospital=Long internment}

Support: 0.0148
Confidence: 0.5052
Lift: 1.12

[13] {gender=Female,
admission_type_id=Urgent,
medical_specialty=Desconhecido} => {time_in_hospital=Long internment}

Support: 0.0148
Confidence: 0.5064
Lift: 1.13

[14] {gender=Female,
admission_type_id=Urgent,
medical_specialty=Desconhecido,
number_diagnoses=Reduced} => {time_in_hospital=Long internment}

Support: 0.0147
Confidence: 0.5067
Lift: 1.13

[15] {race=Caucasian,
diag_2=Digestivo,
number_emergency=Reduced} => {time_in_hospital=Long internment}

Support: 0.0147
Confidence: 0.5018
Lift: 1.12

Tabela 15 - 5 regras com valores mais altos de confiança para o consequente internamento longo.

[16] {diag_1=Circulatório,
diag_2=Geniturinário,
number_outpatient=Reduced} => {time_in_hospital=Long internment}

Support: 0.0102
Confidence: 0.5195
Lift: 1.16

[17] {gender=Female,
age=[60-70),
admission_type_id=Urgent,
number_diagnoses=Reduced} => {time_in_hospital=Long internment}

Support: 0.0107
Confidence: 0.5174
Lift: 1.15

[18] {admission_type_id=Emergency,
diag_2=Digestivo,
number_diagnoses=Reduced,
number_emergency=Reduced} => {time_in_hospital=Long internment}

Support: 0.0108
Confidence: 0.5174
Lift: 1.15

[19] {admission_type_id=Emergency,
diag_2=Digestivo,
number_emergency=Reduced} => {time_in_hospital=Long internment}

Support: 0.0108
Confidence: 0.5173
Lift: 1.15

[20] {gender=Female,
age=[60-70),
admission_type_id=Urgent} => {time_in_hospital=Long internment}

Support: 0.0107
Confidence: 0.5169
Lift: 1.15

No que se refere às regras cujo consequente é o internamento longo, obtiveram-se valores de suporte e confiança bastante semelhantes em ambos os grupos. Em todas as regras apresentadas o valor de lift foi superior a 1, indicando a dependência entre as variáveis. A regra com maior lift de entre todas as regras foi também a que obteve o melhor valor de confiança, regra [16]. Esta regra indica que a probabilidade de um paciente que tenha um internamento cujo diagnóstico principal seja do foro circulatório, o diagnóstico secundário do foro geniturinário e o número de visitas ambulatoriais do paciente no ano anterior ao internamento seja reduzido, ter um internamento longo é 52%. O internamento longo é 1.16 vezes superior em internamentos que obedeçam a estes critérios.

Observa-se que de entre as regras apresentadas existem antecedentes que se repetem dando indicação de que, quando presentes, influenciam um internamento mais prolongado. É o caso do sexo feminino e da idade compreendida entre os 60 e 70 anos. Outra questão importante identificada é que o segundo diagnóstico (diag_2) nas regras para internamento curto aparece apenas uma vez, ao passo que nas regras em que o consequente é o internamento longo o diagnóstico secundário aparece repetidamente em quase todas as regras, podendo isto indicar que o segundo diagnóstico tem uma maior influência na previsão de internamentos mais prolongados.

5.2.2 Estudo das regras de associação raras

Umas das principais limitações levantadas aos algoritmos que têm por base a geração de regras de associação baseada num único valor mínimo de suporte tem a ver com a pesquisa de itens frequentes na base de dados. Uma vez que para base de dados que apresentem itens com frequências diferentes, ao ser definido apenas um valor de suporte mínimo, caso esse limite seja demasiado elevado o número de regras obtidas é reduzido substancialmente. Pelo contrário, caso o valor de suporte mínimo seja definido como um valor muito baixo, são retornadas uma grande quantidade de regras o que torna difícil a sua interpretação e análise (Anselmo, 2017). Os objetos raros são, regra geral, mais difíceis de identificar e generalizar do que os objetos frequentes. Mas é extremamente relevante identificá-los. São diversos os exemplos da importância desta identificação. Para este estudo em particular torna-se relevante identificar padrões que à partida não seriam previsíveis e que podem levar ao prolongamento da estadia hospitalar dos doentes diabéticos.

Ao contrário das regras de associação com itens frequentes, as regras de associação com itens raros têm valores de suporte baixo e de confiança alta (Anselmo, 2017).

(Koh, Rountree, & O'keefe, 2006) propõem a definição de dois valores de suporte para estudar regras que tenham suportes baixos, mas valores de confiança altos. Para isso estabeleceu-se o intervalo de suporte de 0.0001 a 0.01 e o valor de confiança 1. Obtiveram-se 19 305 regras em que o consequente foi o internamento curto e 11 410 regras para o internamento longo. Selecionaram-se algumas das regras com valor de lift superior que estão apresentadas na Tabela 16 e na Tabela 17.

Tabela 16 - 5 regras raras com melhores valores de lift para o consequente internamento curto.

[21] {admission_type_id=Newborn} => {time_in_hospital=Short internment}

Support: 0.0001
Confidence: 1
Lift: 1.82

[22] {medical_specialty=Anesthesiology, number_inpatient=Reduced} => {time_in_hospital = Short internment}

Support: 0.0001
Confidence: 1
Lift: 1.82

[23] {medical_specialty=Osteopath,diag_2=Circulatório} => {time_in_hospital=Short internment}

Support: 0.0001
Confidence: 1
Lift: 1.82

[24] {gender=Male,medical_specialty=Osteopath}=> {time_in_hospital=Short internment}

Support: 0.0001
Confidence: 1
Lift: 1.82

[25] {age=[40-50),medical_specialty=Gynecology}=> {time_in_hospital=Short internment}

Support: 0.0001
Confidence: 1
Lift: 1.82

Tabela 17 - 5 regras raras com melhores valores de lift para o consequente internamento longo.

[26] {age=[70-80),
medical_specialty=PhysicalMedicineandRehabilitation,
diag_3=Geniturinário } => {time_in_hospital=Long internment}

Support: 0.0001
Confidence: 1
Lift: 4.87

[27] {age=[10-20),
medical_specialty=Pulmonology,
diag_3=Diabetes} => {time_in_hospital=Long internment}

Support: 0.0001
Confidence: 1
Lift: 4.86

[28] {age=[70-80),medical_specialty=Pathology}=>{time_in_hospital=Long internment}

Support: 0.0001
Confidence: 1
Lift: 3.23

[29] {age=[20-30),medical_specialty=Psychology} =>{time_in_hospital=Long internment}

Support: 0.0001
Confidence: 1
Lift: 2.23

[30] {age=[50-60),number_diagnoses=High}=> {time_in_hospital=Long internment}

Support: 0.0001
Confidence: 1
Lift: 2.13

Das regras raras apresentadas verificou-se que o lift aumentou comparativamente ao lift obtido para o conjunto de regras frequentes. Esta análise vem confirmar a importância de regras que, embora possam aparecer menos vezes, tenham uma influência mais significativa no período total de internamento. Pode-se afirmar, à luz dos dados em estudo, que a probabilidade de um internamento cujo tipo de admissão tenha sido recém-nascido ter um internamento curto é de 100%. O internamento curto é 1.82 vezes superior entre os internamentos que a admissão tenha sido "recém nascido". Suportado pelo mesmo valor de confiança e de lift pode-se concluir que o internamento curto ocorre aproximadamente o dobro das vezes entre internamentos que a especialidade médica seja anestesiologia e o número de visitas de internamento do paciente no ano anterior ao encontro tenha sido reduzido. O mesmo se pode concluir sobre os antecedentes especialidade médica de osteopatia/ diagnóstico secundário ser do foro circulatório quando ocorrem em conjunto e, ao que parece, ser do sexo masculino e a especialidade médica do internamento ser osteopatia ou então ter uma idade compreendida entre os 40 e 50 anos e a especialidade médica do internamento ser de ginecologia tem mais probabilidade do desfecho do internamento ser curto.

Das regras obtidas, cujo consequente é o internamento mais prolongado chegou-se a algumas previsões que já seriam à partida de esperar. É o exemplo da regra [26] que indica que o internamento longo é 4.87 vezes superior em internamentos cuja especialidade médica seja medicina física e de reabilitação, a idade do paciente esteja compreendida entre os 70 e os 80 anos e o diagnóstico terciário seja do foro geniturinário. A idade avançada de um paciente ajuda por si só a prever a ocorrência de um internamento mais prolongado, causado pelas co-morbilidades associadas ao envelhecimento (Marques & Ferreira, 2010). Para além de que se está a tratar de episódios em que a diabetes foi inserida num dos

diagnósticos e, por isso, sabe-se também que a idade pode acelerar o aparecimento de complicações relacionadas à doença da diabetes. Algumas dessas possíveis complicações são doenças musculoesqueléticas (Silva & Skare, 2012) e doenças do trato urinário (de Oliveira, Marinheiro, & da Silva, 2011). Daí que seja plausível que internamentos de pacientes com estes problemas associados possam ter mais complicações ao longo do internamento e, por isso, levar a um internamento mais prolongado. Outro exemplo que já seria expectável é a regra [30], dado que se um paciente for diagnosticado com um elevado número de diagnósticos deve, à partida, necessitar de mais cuidados. No entanto, chegou-se a resultados menos expectáveis à primeira vista e que podem dar informações importantes ao estudo. Vejamos o exemplo da regra [27], aparentemente o internamento hospitalar longo é 4.86 vezes superior em crianças/adolescentes com idade compreendida entre os 10 e os 20 anos que tenham sido internadas pela especialidade de pneumologia e em que a diabetes foi inserida como um diagnóstico terciário. Outro contorno interessante é perceber pela regra [29] que jovens com idade compreendida entre os 20 e 30 anos que tenham sido internados pela especialidade médica de psicologia estejam mais relacionados com internamentos mais prolongados.

5.3 Previsão do número de dias de internamento

5.3.1 Algoritmo Random Forest

Com as regras de associação obtidas no ponto 5.2 deste trabalho, foi possível tirar algumas conclusões e informações úteis à cerca da amostra em estudo. No entanto, sentiu-se a necessidade de se chegar a uma perspetiva mais integrada da dependência entre as variáveis. Neste sentido, optou-se por aplicar o algoritmo Random Forest, dado que este método gera várias árvores de decisão e uma árvore de decisão, por sua vez, gera várias regras de uma só vez relacionando as diferentes variáveis. E, por fim, estas regras obtidas estão também relacionadas entre si através dos valores das variáveis.

Conforme explorado no capítulo de estado da arte desta dissertação, a algoritmo de floresta aleatória é muito utilizado pois, para além de ser um método simples de compreender também se obtém, regra geral, bons resultados. Por este motivo, selecionou-se o package `RandomForest` do R para prever o tempo de internamento hospitalar. Dois parâmetros são muito importantes por terem muita influência na previsão final no algoritmo de Random Forest: `ntree` e `mtry`. O parâmetro `ntree` corresponde ao número de árvores usadas no algoritmo e não deve ser definido para um número muito pequeno, para garantir que cada linha de entrada seja prevista pelo menos algumas vezes. O parâmetro `mtry` diz respeito ao número de variáveis aleatórias usadas para divisão em cada nó. O algoritmo Random Forest segue o seguinte procedimento:

1. Seleção aleatória de `ntree` amostras de treino dos dados originais. A seleção é realizada por *bootstrap* – método de reamostragem, em que as amostras selecionadas podem ser repetidas na seleção;
2. Para cada umas das amostras de *bootstrap* é criada uma árvore de classificação;
3. Em cada nó interno é selecionado aleatoriamente `mtry` dos N preditores (variáveis) e determinada a melhor divisão usando apenas esses preditores. A melhor divisão é selecionada com base no cálculo do índice de Gini. Para escolher as variáveis do próximo nó, serão escolhidas novamente `mtry` variáveis, excluindo as já selecionadas anteriormente e o processo prossegue até ao último nó.
4. A nova árvore criada é salva ao lado das construídas até agora.
5. Quando todas as `ntree` árvores estiverem criadas é possível fazer a previsão de novos dados. (Zawbaa, Hazman, Abbass, & Hassanien, 2014).

Reduzir o `mtry` tende a reduzir consequentemente a correlação. Pelo contrário, o aumento deste parâmetro aumenta a sua correlação. É, por isso, necessário encontrar a faixa "ideal" de `mtry` que pode ser encontrada usando a taxa de erro OOB – Out of Bag. Cada árvore é treinada com aproximadamente 2/3 do total de dados de treinamento, pelo que os dados são selecionados de forma aleatória com substituição dos dados originais. Para cada uma das árvores, a partir dos dados restantes (1/3), é calculada a taxa de erro de classificação – OOB. A taxa média de erro de todas as árvores é calculada para determinar o erro geral obtido para a respetiva classificação. Optou-se por utilizar o método de validação *Holdout* na proporção 70%/30%, para garantir de forma quantitativa a capacidade de generalização do modelo.

Inicialmente selecionou-se o conjunto de treino (70%) e separou-se novamente em conjunto de teste e de treino. Testou-se o modelo de Random Forest para diferentes valores de `ntree` e `mtry` e selecionou-se os parâmetros que deram os melhores resultados no conjunto de testes interno. Posteriormente, testou-se nos 70% dos dados totais e fez-se a validação no conjunto de teste (30%) que ficaram de fora. No momento da seleção dos parâmetros `ntree` e `mtry,` primeiramente criou-se o modelo de Random Forest com os parâmetros padrão e, posteriormente tentou-se chegar aos melhores resultados ajustando os parâmetros. Para modelos de classificação, o padrão de `mtry` é a raiz quadrada do número de variáveis preditoras (arredondado para baixo), o que neste caso corresponde a 4 e o valor padrão de `ntree` é de 500 (Cutler, Beard, Cutler, & Gibson, 2007).

Fixou-se o valor de `ntree` em 500 e tentou-se chegar ao melhor valor de `mtry`, ou seja, o valor de `mtry` com o qual se obteve o melhor valor de *accuracy* no *dataset* de validação. Os resultados obtidos estão apresentados na Figura 30 e pode-se constatar que o melhor valor obtido de *accuracy* foi para o `ntry`=5. A partir deste valor de `ntry`, o valor de *accuracy* estabilizou.

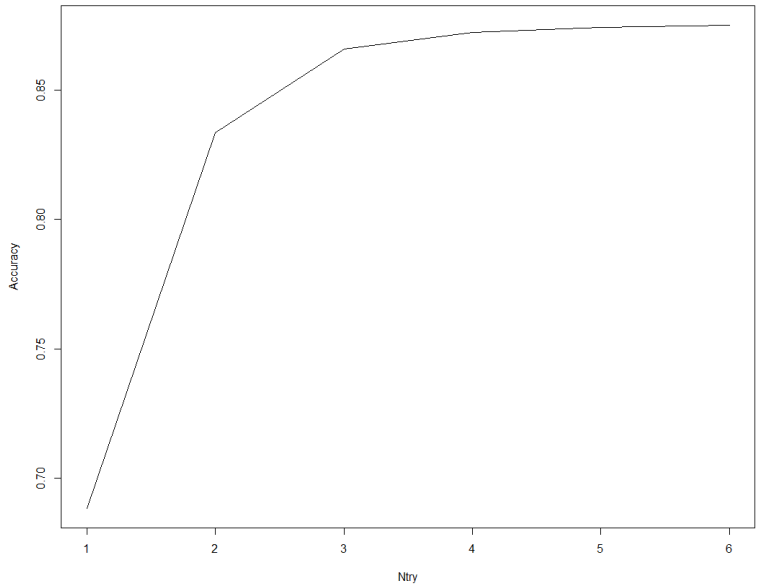


Figura 30 - Variação da a*ccuracy* para diferentes valores de `ntry`.

Definiu-se o valor de `ntry` a 5 e variou-se o `mtree`, entre 400 e 800, para chegar ao melhor valor possível. Concluiu-se que o melhor valor de precisão

obtido foi para `mtree`=600. Utilizou-se, por isso, estes parâmetros para gerar o modelo de previsão de número de dias de internamento. A matriz de confusão para o conjunto de validação está apresentada na Tabela 18. Obteve-se uma *accuracy* de 81%. A curva de ROC obtida para o internamento curto e longo está apresentada na Figura 31.

Tabela 18 - Matriz de confusão dados de validação. *Accuracy* dados de validação: 81%, sensibilidade: 72%, especificidade: 89%

Observed / **Predicted**	**Long internment**	**Short internment**
Long internment	4439 VP	845 FP
Short internment	1717 FN	6621 VN
Error (%)	27,89	11,32

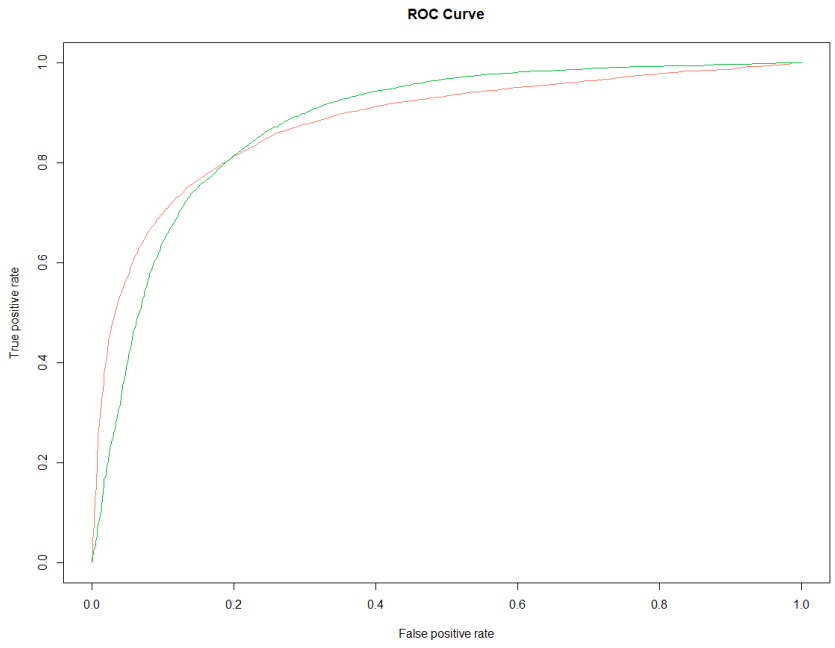


Figura 31 - Curva de ROC obtida para internamento longo (vermelho) e internamento curto (verde).

A biblioteca Random Forest possui uma funcionalidade bastante útil que é a função `importance()` que permite identificar as variáveis que tiveram mais influência na previsão do número de dias de internamento. Este é um conceito difícil de definir dado que a importância de uma variável pode ser devido à interação com outras variáveis (Liaw & Wiener, 2014). O resultado está apresentado na Figura 32. O parâmetro *Mean Drecrease Accuracy* representa o quanto a *accuracy* do modelo diminui se essa variável for eliminada.

Verifica-se que o número de diagnósticos toma a maior importância para a previsão dos dados em estudo seguido da especialidade médica e o diagnóstico principal. Já anteriormente, pela análise das regras de associação, tinha sido possível perceber que o número de diagnósticos tem influência no período de internamento. Foram encontradas diversas regras em cujo antecedente estava incluída esta variável - regras [2], [12], [17], [18] e [30].

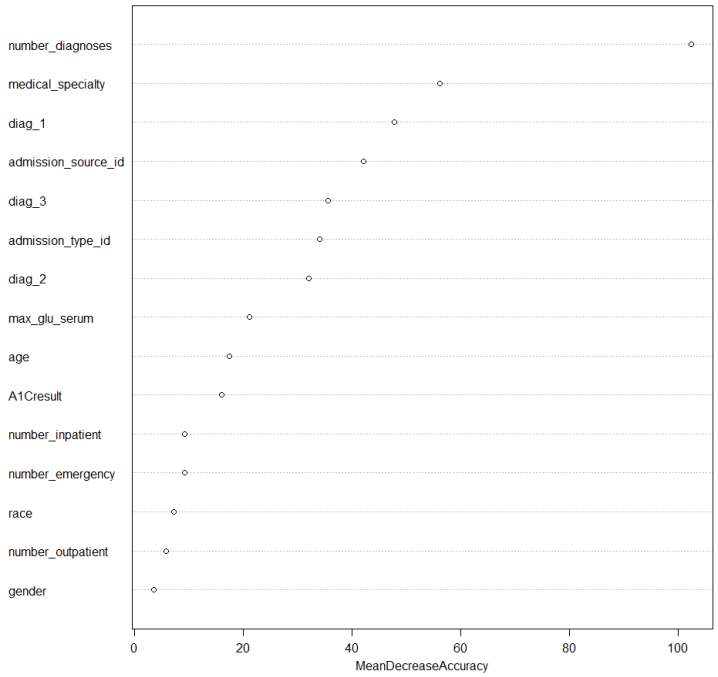


Figura 32 - Representação da importância das variáveis em estudo.

5.3.2 Aplicação para previsão do prolongamento hospitalar

De forma a aplicar o modelo de Random Forest obtido, desenvolveu-se uma aplicação de interface com o utilizador com a biblioteca *shiny* do R. A interface foi desenvolvida com o intuito de ser possível, de forma rápida e intuitiva introduzir as informações necessárias relativas ao paciente e ao internamento e obter de imediato a previsão estimada do prolongamento hospitalar para aquele paciente e sob as condições associadas. O exemplo está demonstrado na Figura 33, Figura 34 e Figura 35:

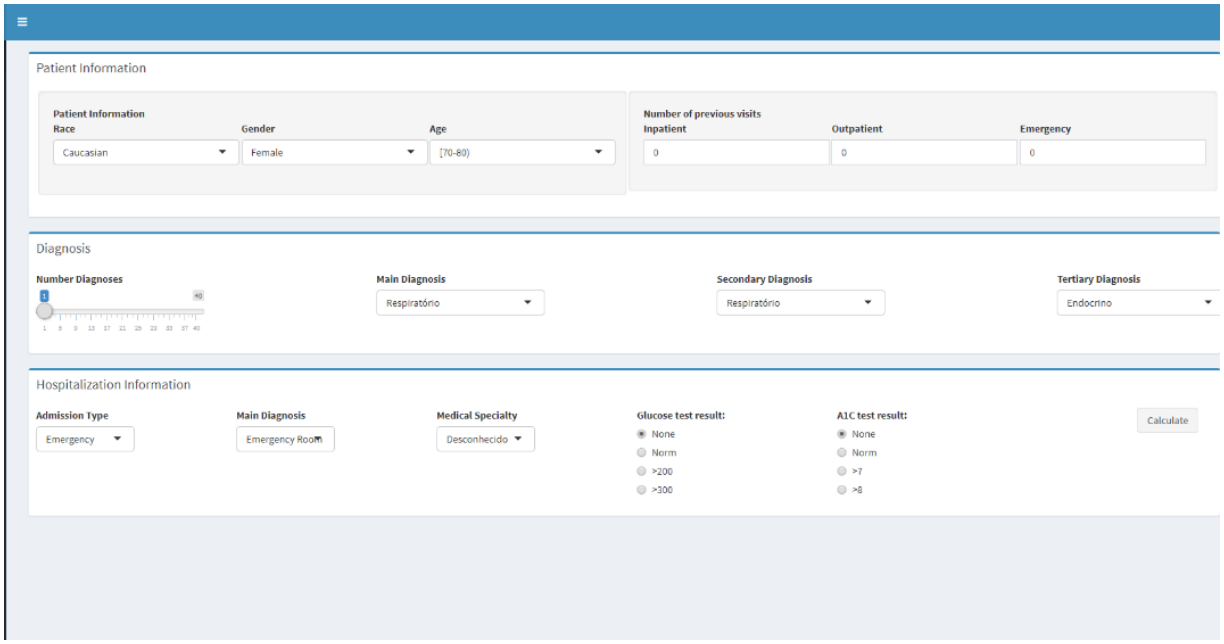


Figura 33 - Ilustração da interface com o utilizador para introdução dos dados relativos ao paciente e ao internamento.

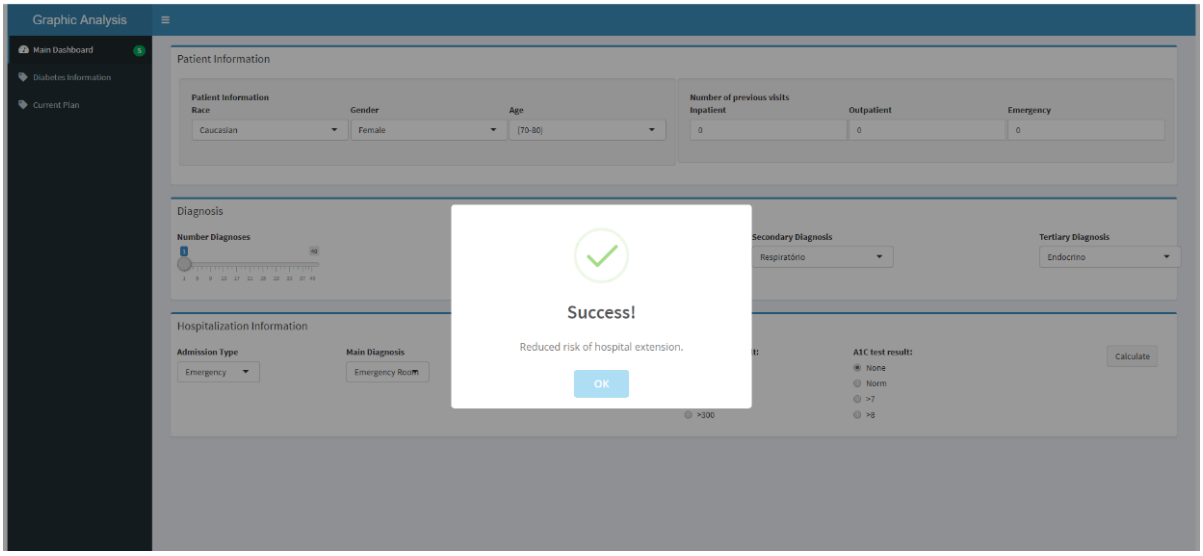


Figura 34 - Exemplificação de uma mensagem de sucesso, no caso de os dados introduzidos indicarem uma probabilidade acrescida de internamento curto para os dados introduzidos.

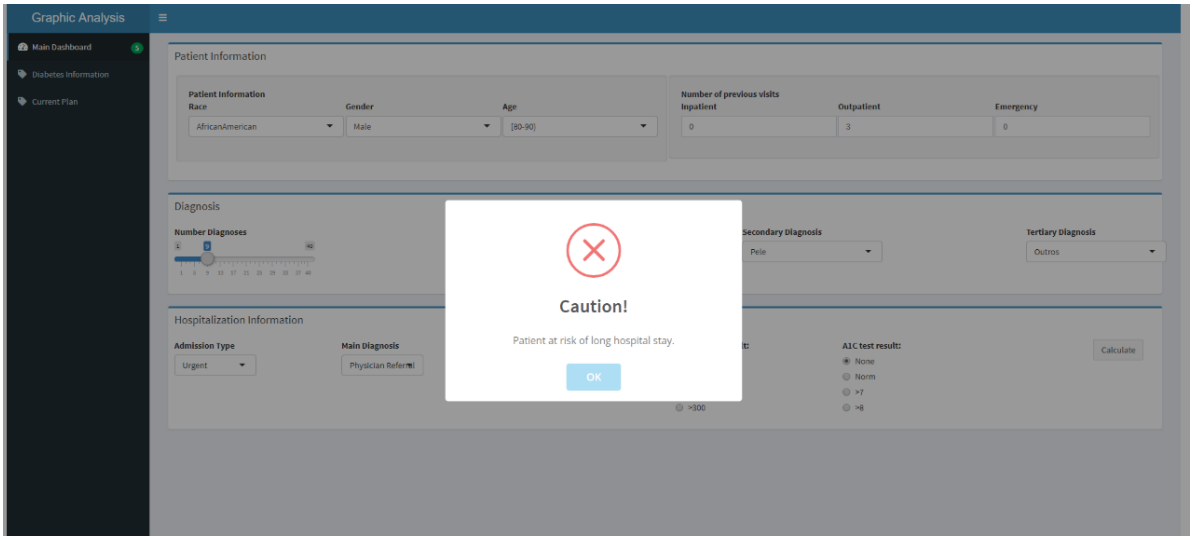


Figura 35 -Exemplificação de uma mensagem de risco elevado, no caso de os dados introduzidos indicarem uma probabilidade acrescida de internamento longo para os dados introduzidos.

6. Conclusões e Trabalho Futuro

Pode-se constatar que as regras de associação representam uma técnica promissora para encontrar padrões ocultos em conjuntos de dados médicos. Permitem perceber e analisar de que forma e que determinadas condições ou variáveis podem levar a determinado desfecho. Neste caso em particular o que pode levar um paciente diabético a ter um período de internamento mais curto ou mais prolongado. O principal problema encontrado sobre regras de associação foi o grande número de regras que são obtidas, a maioria das quais irrelevantes. Tal número de regras torna a pesquisa lenta e dificulta a interpretação. Conseguiu-se, no entanto, chegar a alguns padrões interessantes sustentados com algum suporte e confiança.

Conseguiu-se ainda a partir de 15 variáveis diferentes relativas ao paciente e ao internamento hospitalar e com a utilização do algoritmo Random Forest obter uma *accuracy* de 81%, o que comparativamente ao estado da arte analisado nos pareceu bastante positivo. Foi ainda possível concluir que tanto as regras de associação quanto o algoritmo de Random Forest apontaram para um subconjunto comum de variáveis importantes na previsão de internamento curto e longo.

Constatou-se ainda que a biblioteca `shiny()` do R tem potencialidade para ser implementada como uma ferramenta auxiliar na previsão no número de dias de internamento.

6.1 Trabalhos Futuros

Como trabalho futuro, pretende-se explorar o algoritmo com um maior volume de dados e sem a limitação de 14 dias máximo de internamento, de forma a permitir explorar padrões relativos a internamentos prolongados para além dos já estudados neste trabalho. É também objetivo num futuro trabalho, explorar dados relativos aos internamentos de doentes diabéticos, mas sem a limitação de que nesses internamentos tenham obrigatoriamente sido administrados medicação e realizados testes de laboratório. Ter esta limitação pode traduzir-se em perda de informação importante de padrões que fazem parte do contexto de internamento hospitalar de doentes diabéticos, como foi observado neste trabalho.

Pretende-se também ampliar os fatores de determinadas variáveis incluídas neste estudo. Exemplo disso é a variável do valor de glicose. Neste estudo apenas tínhamos informação quando o resultado obtido no teste era normal ou superior ao considerado normal. No entanto, estudos realizados (Nirantharakumar et al., 2012) revelam que a hipoglicemia, ou seja, valores de glicose a baixo dos padrões considerados normais, estão associados a um prolongamento hospitalar.

Pretende-se ainda num futuro trabalho estabelecer a comparação de outros algoritmos de previsão para além dos incluídos neste trabalho, tais como redes neuronais artificiais e SVM. É ainda objetivo comparar a viabilidade de estudar a variável período de internamento não só como um *output* binomial, conforme apresentado neste trabalho, mas também trinomial e/ou quadrinomial por se traduzir em informação mais especifica para o auxiliar de saúde.

Outra questão importante a explorar de futuro seria a melhoria da aplicação em `Shiny()` desenvolvida neste trabalho. Pretende-se melhorar a forma de obtenção de dados. Atualmente cada um dos campos de *input* tem de ser introduzido manualmente na aplicação e não ficam armazenados em memória. Seria bastante benéfico que existisse uma integração com o(s) sistema(s) de informação dos hospitais de saúde, de forma a que informações já contidas nesses sistemas fossem importadas para a aplicação, para reduzir a quantidade de informação necessária a introduzir pelo auxiliar de saúde. Exemplos de campos que poderiam ser automaticamente preenchidos seriam o número de internamentos anteriores ao encontro e dados relativos ao paciente como a idade, etnia e sexo. Para além disso, o resultado obtido deveria ser registado na base de dados do sistema de saúde hospitalar para análise e histórico futuro. Ainda em relação à aplicação, era vantajoso que para além da mensagem de

feedback relativo ao prolongamento hospitalar dos pacientes, se se conseguisse obter a probabilidade relativa àquele feedback e a margem de erro associada.

7. Referências

Al Taleb, A. R., Hoque, M., Hasanat, A., & Khan, M. B. (2017). Application of data mining techniques to predict length of stay of stroke patients. *2017 International Conference on Informatics, Health and Technology, ICIHT 2017*, 1–5. https://doi.org/10.1109/ICIHT.2017.7899004

Alahmar, A., Mohammed, E. A., & Benlamri, R. (2018). Application of data mining techniques to predict the length of stay of hospitalized patients with diabetes. *Proceedings - 2018 International Conference on Big Data Innovations and Applications, Innovate-Data 2018*, 38–43. https://doi.org/10.1109/Innovate-Data.2018.00013

Ali, S., & Dornhorst, A. (2011). Diabetes in pregnancy: Health risks and management. *Postgraduate Medical Journal*, *87*(1028), 417–427. https://doi.org/10.1136/pgmj.2010.109157

American Diabetes Association. (2014). Diagnosis and Classification of Diabetes Mellitus Definition and Description Of Diabetes Mellitus, *37*(January), 81–90. https://doi.org/10.2337/dc14-S081

Andersson. (2019). Predicting Patient Length Of Stay at Time of Admission Using Machine Learning. *KTH Royal Institute of Technology School Of Engineering Sciences In Chemistry, Biotechnology And Health*.

Anselmo. (2017). Regras de Associação – Market Basket Analysis itens frequentes e itens raros.

APDP – Associação Protectora dos Diabéticos de Portugal. (2020). Hiperglicemia. Retrieved from https://apdp.pt/diabetes/a-pessoa-com-diabetes/hiperglicemia/

Aragão, C. (2016). Dia Mundial da Saúde: "Diabetes é a doença do século XXI." Retrieved from https://jpn.up.pt/2016/04/07/diabetes-e-a-doenca-do-seculo-xxi/

Atlas, I. D. F. D. (2019). *Idf diabetes atlas*.

Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and J. N. C. (2014). Diabetes 130-US hospitals for years 1999-2008 Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008

Camilo, C., & Silva, J. (2009). Mineração de Dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFC)*, 29. https://doi.org/10.1007/978-3-319-18032-8_50

Carral, F., A, G. O., C, Salas, J., B A. (2001). Care resource utilization and direct costs incurred by peoplewith diabetes in a Spanish hospital.

Chen, Y. L., Tang, K., Shen, R. J., & Hu, Y. H. (2005). Market basket analysis in

a multiple store environment. *Decision Support Systems*, *40*(2), 339–354. https://doi.org/10.1016/j.dss.2004.04.009

Combes, C., Kadri, F., & Chaabane, S. (2014). Predicting Hospital Length of Stay Using Regression Models: Application To Emergency Department. *10ème Conférence Francophone de Modélisation, Optimisation et Simulation- MOSIM'14*. Retrieved from https://hal.archives-ouvertes.fr/hal-01081557/

Comino, E. J., Harris, M. F., Islam, F., Tran, D. T., Jalaludin, B., Jorm, L., … Haas, M. (2015). Impact of diabetes on hospital admission and length of stay among a general population aged 45 year or more : a record linkage study, 1–13. https://doi.org/10.1186/s12913-014-0666-2

Côrtes, S. D. C., Porcaro, R. M., & Lifschitz, S. (2002). Mineração de Dados – Funcionalidades, Técnicas e Abordagens. *PUC-Rio Informática*, 35. Retrieved from ftp://ftp.inf.puc-rio.br/pub/docs/techreports/02_10_cortes.pdf

Cortez, P., & Neves, J. (2000). Redes Neuronais Artificiais.

Cummings, D. (2018). Predicting hospital length-of-stay at time of admission. Retrieved from https://towardsdatascience.com/predicting-hospital-length-of-stay-at-time-of-admission-55dfdfe69598

Cutler, D. R., Beard, K. H., Cutler, A., & Gibson, J. (2007). Random Forests for Classification in Ecology, (December). https://doi.org/10.1890/07-0539.1

Daghistani, T. A., Elshawi, R., Sakr, S., Ahmed, A. M., Al-Thwayee, A., & Al-Mallah, M. H. (2019). Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. *International Journal of Cardiology*, *288*(xxxx), 140–147. https://doi.org/10.1016/j.ijcard.2019.01.046

Daniel Harris, BA, Lynn McNicoll, MD, Gary Epstein-Lubow, MD, and Kali S. Thomas, P. (2017). A Survey of Methods for Time Series Change Point Detection. *Physiology & Behavior*, *176*(1), 139–148. https://doi.org/10.1016/j.physbeh.2017.03.040

DataCamp. (2019). Support Vector Machines with Scikit-learn. Retrieved from https://www.datacamp.com/community/tutorials/

Oliveira, E. G., Marinheiro, L. P. F., & da Silva, K. S. (2011). Diabetes melito como fator associado às disfunções do trato urinário inferior em mulheres atendidas em serviço de referência. *Revista Brasileira de Ginecologia e Obstetricia*, *33*(12), 414–420. https://doi.org/10.1590/S0100-72032011001200007

Deshpande, D., Harris-Hayes, M., & Schootman, M. (2008). Diabetes-Related Complications. *Diabetes Special Issue*, *88*(11), 1254–1264.

Ferneda, E. (2006). Redes neurais e sua aplicação em sistemas de recuperação de informação, 25–30.

Forman, G., & Scholz, M. (2010). Apples-to-Apples in Cross-Validation Studies : Pitfalls in Classifier Performance Measurement, *12*(1), 49–57.

Gentimis, T., Alnaser, A. J., Durante, A., Cook, K., & Steele, R. (2018). Predicting hospital length of stay using neural networks on MIMIC III data. *Proceedings - 2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 2017 IEEE 15th International Conference on Pervasive Intelligence and*

Computing, 2017 IEEE 3rd International Conference on Big Data Intelligence and Compu, *2018-Janua*, 1194–1201. https://doi.org/10.1109/DASC-PICom-DataCom-CyberSciTec.2017.191

Hachesu, P. R., Ahmadi, M., Alizadeh, S., & Sadoughi, F. (2013). Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthcare Informatics Research*, *19*(2), 121–129. https://doi.org/10.4258/hir.2013.19.2.121

Han, J., Kamber, M., & Pei, J. (2016). *Data Mining, Third Edition: Practical Machine Learning Tools and Techniques*. *Complementary literature None*. https://doi.org/0120884070, 9780120884070

Health at a Glance 2017. (2017). *Revista de Investigacion Clinica* (Vol. 49).

IJsselmuiden, C. B., & Faden, R. R. (1992). The New England Journal of Medicine Downloaded from nejm.org on January 31, 2011. For personal use only. No other uses without permission. Copyright © 1992 Massachusetts Medical Society. All rights reserved., *326*.

International Diabetes Federation. (2003). *DIABETES*.

Kelly, L. (2014). *Mining data. Northern Ontario Business*.

Knecht, L. A. D., Gauthier, S. M., Castro, J. C., Schmidt, R. E., Whitaker, M. D., Zimmerman, R. S., … Cook, C. B. (2006). Diabetes care in the hospital: is there clinical inertia? *Journal of Hospital Medicine (Online)*, *1*(3), 151–160. https://doi.org/10.1002/jhm.94

Koh, Y. S., Rountree, N., & O'keefe, R. (2006). Finding Non-Coincidental Sporadic Rules Using Apriori-Inverse. *International Journal of Data Warehousing and Mining (IJDWM)*, *2*(2), 38–54. https://doi.org/10.4018/jdwm.2006040102

Liaw, A., & Wiener, M. (2014). Classification and Regression by randomForest, (November 2001).

Livieris, I. E., Kotsilieris, T., Dimopoulos, I. F., & Pintelas, P. (2018). Predicting length of stay in hospitalized patients using SSL algorithms. *ACM International Conference Proceeding Series*, 16–22. https://doi.org/10.1145/3218585.3218588

Maglogiannis, I., Karpouzis, K., Wallace, M., & Soldatos, J. (2007). *Emerging Artificil Intelligence Applications In Computer Engineering*. Retrieved from http://www.ghbook.ir/index.php?name=فرهنگ و رسانه های نوین&option=com_dbook&task=readonline&book_id=13650&page=73&chkhashk=ED9C9491B4&Itemid=218&lang=fa&tmpl=component

Marques, C., & Ferreira, J. (2010). UNIVERSIDADE DA BEIRA INTERIOR Causas do Prolongamento do Internamento: O caso de um serviço de Medicina Interna.

Menzin, J., Korn, J. R., Cohen, J., Lobo, F., Zhang, B., Friedman, M., & Neumann, P. J. (2010). Relationship between glycemic control and diabetes-related hospital costs in patients with type 1 or type 2 diabetes mellitus. *Journal of Managed Care Pharmacy*, *16*(4), 264–275.

https://doi.org/10.18553/jmcp.2010.16.4.264

Morton, A., Marzban, E., Giannoulis, G., Patel, A., Aparasu, R., & Kakadiaris, I. A. (2014). A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients. *Proceedings - 2014 13th International Conference on Machine Learning and Applications, ICMLA 2014*, 428–431. https://doi.org/10.1109/ICMLA.2014.76

Nirantharakumar, K., Marshall, T., Kennedy, A., Narendran, P., Hemming, K., & Coleman, J. J. (2012). Short Report : Treatment Hypoglycaemia is associated with increased length of stay and mortality in people with diabetes who are hospitalized, 445–448. https://doi.org/10.1111/dme.12002

Observatório da diabetes. (2016). *Diabetes: Factos e Numeros ano 2015*. *Relatório anula do observatório nacional de diabetes*.

Observatório Nacional da Diabetes. (2010). *Diabetes: factos e números 2009 - Relatório Anual do Observatório Nacional de Diabetes*. Lisboa: Sociedade Portuguesa de Diabetologia.

Paiva, A. (2016). Melhorias no controlo da glicemia através de data mining Melhorias no controlo da glicemia através de data mining.

Palaniappan Sellappan, A. R. (1991). Intelligent Heart Disease Prediction System Using Data Mining Techniques Sellappan. *Anasthesiologie Und Intensivmedizin*, *32*(2), 52–54.

Silva, M. B. G., & Skare, T. L. (2012). Manifestações musculoesqueléticas em diabetes mellitus. *Revista Brasileira de Reumatologia*, *52*(4), 601–609. https://doi.org/10.1590/s0482-50042012000400010

Strack, B., Deshazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, *2014*. https://doi.org/10.1155/2014/781670

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*, 1–11. https://doi.org/10.1186/1471-2105-9-307

Tripathi, B. K., & Srivastava, A. K. (2006). Diabetes mellitus: Complications and therapeutics. *Medical Science Monitor*, *12*(7), 130–147.

Turgeman, L., May, J. H., & Sciulli, R. (2017). Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission. *Expert Systems with Applications*, *78*, 376–385. https://doi.org/10.1016/j.eswa.2017.02.023

Veiga, D. M., & Ferreira, D. (2011). Será Possível Melhorar O Diagnóstico Da Icterícia Neonatal? Aplicação De Técnicas De Data Mining.

Veloso, F. (2000). Um Modelo para Previsã de Churn na Área do Retalho. *Journal of Political Economy*.

Walczak, S., Scorpio, R. J., & Pofahl, W. E. (1998). Predicting Hospital Length

of Stay with Neural Networks. *Proceedings of the Eleventh International FLAIRS Conference*, 333–337. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/citations;jsessionid=B65EA46BD014209948E015D231FB84BB?doi=10.1.1.534.507

World Health Organization. (2016). Global Report on Diabetes. *Isbn*, *978*, 6–86. Retrieved from http://www.who.int/about/licensing/

Yiu, T. (2020). Understanding Random Forest How the Algorithm Works and Why it Is So Effective.

Zawbaa, H. M., Hazman, M., Abbass, M., & Hassanien, A. E. (2014). Automatic fruit classification using random forest algorithm, 164–168.

Zhang, H. (2004). The optimality of Naive Bayes. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004*, *2*, 562–567.