INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Previsão do Pagamento em Atraso de Faturas

PEDRO JOSÉ GONÇALVES PORTELA FALHAS DA COSTA
Outubro de 2020

POLITÉCNICO
DO PORTO

Instituto Superior de Engenharia do Porto

Previsão do Pagamento em Atraso de Faturas

Pedro José Gonçalves Portela Falhas da Costa

Dissertação para obtenção do Grau de Mestre em Engenharia Informática, Área de Especialização em Informação e Conhecimento

Orientador: Carlos Manuel Abreu Gomes Ferreira

Porto, outubro 2020

Resumo

O pagamento em atraso das faturas é um dos principais desafios das operações de uma empresa. Com uma gestão inadequada do processo de cobrança de faturas, os pagamentos em atraso podem-se acumular e causar problemas no negócio. Por outras palavras, o aumento do número de faturas não pagas pode levar a problemas de fluxo de caixa na empresa. Nesta dissertação, é desenvolvido um sistema automático de treino de modelos de previsão do pagamento de faturas.

Na realização da solução, são criados modelos de aprendizagem automática supervisionada para identificar antecipadamente as faturas que serão pagas em atraso. Será seguida a metodologia CRISP-DM, onde os principais procedimentos abordados pelo trabalho são a limpeza e pré-processamento de dados, construção de modelos de aprendizagem automática e avaliação do desempenho do modelo.

A solução pretendida pela CPCIT4all fornecerá aos seus clientes do setor público, como a Goldenergy, fornecedor dos dados para o projeto, a possibilidade de prever que clientes seus têm maior probabilidade de não pagar dentro do limite. Podendo assim, atuar sobre estes para que realizem os pagamentos dentro o limite, melhorando consequentemente o fluxo de caixa da empresa.

Palavras-chave: Pagamento de faturas em atraso, Inteligência Artificial, Aprendizagem Automática, Deteção de Anomalias, Análise Preditiva, Balanceamento de dados.

Palavras-chave (tecnologia): Python, Pandas, Numpy, Sklearn, Power BI.

Abstract

The late payment invoices is one of the main challenges of a company's operations. With a inadequate management of the invoice collection process, the late payments can accumulate and cause business problems. In other words, increasing the number of unpaid invoices can lead to cash flow problems in the company. In this dissertation, a proactive approach was directed to improve the management of invoice payment collection, using a predictive model generated by an automated mechanism.

In the realization of the solution, supervised machine learning models were created to identify in advance the invoices that will be paid late. The CRISP-DM methodology will be followed, where the main procedures covered by the work are data cleaning and pre-processing, construction of automatic learning models and model performance evaluation.

The solution intended by CPCIT4all, will provide its public sector customers, such as Goldenergy, provider of the data for this project, with the possibility of predicting which of their customers are more likely to not pay within the limit, acting on them to make payments within the limits, thus improving their invoice collection process and consequently increasing the company's cash flow.

Keywords: Late payment invoices, artificial intelligence, machine learning, anomaly detection, predictive analysis, unbalanced data.

Keywords(technology): Python, Pandas, Numpy, Sklearn, Power BI

Agradecimentos

À minha família por todo o apoio dado durante este percurso.

Índice

Lista de Figuras

Lista de Tabelas

Acrónimos e Símbolos

Lista de Acrónimos

CPCIT4all	Companhia Portuguesa de Computadores, Inovação Tecnológica, Lda
CRISP-DM	Cross-Industry Standard Process for Data Mining
KDD	Knowledge Discovery in Databases
Python	*Linguagem de Programação Python*
IA	Inteligência Artificial
ISEP	Instituto Superior de Energia do Porto
	Vector Space Model

1 Introdução

O presente capítulo apresenta o contexto do projeto, o problema abordado, os objetivos a alcançar, a análise de valor da solução oferecida, a metodologia aplicada na criação do modelo de aprendizagem automática e por fim uma descrição da estrutura do documento, onde é descrito brevemente o que é abordado em cada capítulo.

1.1 Contexto

O projeto realizado na CPCIT4all, assenta no desenvolvimento de algoritmos de aprendizagem automática capazes de prever se os pagamentos das novas faturas dos clientes serão realizados dentro ou fora do tempo limite.

Os pagamentos das faturas em atraso é um dos principais desafios das operações de uma empresa, uma vez que a sua acumulação poderá criar problemas no seu fluxo de caixa da empresa.

A CPCIT4all tem um conjunto de clientes do setor de serviços públicos, como a GoldEnergy que pretende um produto de análise e previsão do pagamento em atraso de faturas.

Devido a este requisito, no projeto desenvolvido foi estudado o comportamento dos clientes quanto ao pagamento de faturas de uma empresa do setor de serviços públicos. Em vez de estudarmos este comportamento de uma forma tradicional, onde os contabilistas analisam o histórico de cada cliente classificando-o apenas como sendo de risco ou não e consumindo uma enorme quantidade de tempo, visto que os registos serão analisados manualmente, estamos interessados em detetar antecipadamente e de forma automática as potenciais faturas atrasadas usando métodos de análise inovadores.

Os resultados do sistema pretendido serão depois apresentados através de um relatório dinâmico do Power BI [1], para ser fornecido ao cliente uma forma intuitiva e informativa.

Todo o processo de desenvolvimento da funcionalidade é acompanhado por outros processos de engenharia, de onde se destaca a análise dos dados fornecidos, bem como a elaboração do design de toda a funcionalidade desenvolvida.

O projeto foi realizado apenas por um membro, encarregado por todas as partes necessárias para o seu desenvolvimento: análise, design, desenvolvimento e documentação técnica.

1.2 Problema

Os pagamentos em atraso das faturas é um dos principais desafios das operações de uma empresa. Com uma gestão inadequada do processo de cobrança de faturas, os pagamentos em atraso podem-se acumular e causar problemas nos negócios. Em gestão comercial, a cobrança de faturas não pagas é um trabalho muito tedioso e importante para a maioria das empresas.

Devido à grande quantidade de dados existentes a abordagem tradicional para extração de informação sobre os comportamentos dos clientes não é possível, como explicado anteriormente, por isso apenas através da utilização de um modelo de aprendizagem automática será possível analisar e prever o comportamento dos clientes.

Sendo uma base de dados real, os dados não se encontram tratados e dados de baixa qualidade geram modelos de baixa qualidade, sendo necessário um processo de tratamento dos dados.

Existem dois tipos de clientes, os clientes novos e os clientes recorrentes. Os clientes novos são os clientes que realizam o primeiro pagamento e os clientes recorrentes são clientes com mais de um pagamento já realizado, sendo necessário a criação de um modelo para cada tipo de cliente.

A CPCIT4all tendo vários clientes industriais e pretende fornecer um serviço para lidar com estes problemas, sendo necessário uma ferramenta genérica que resolva os problemas de previsão e análise de dados específicos de cada empresa cliente da CPCIT4all, sendo necessário a construção de um sistema automático capaz de lidar com estes problemas.

1.3 Objetivo

O objetivo deste projeto é construir mecanismo automático de criação de modelos de aprendizagem supervisionada para prever as faturas que serão pagas em atraso com antecedência para melhorar o processo de cobrança destas.

Este mecanismo será um mecanismo automático, onde para cada conjunto de dados de cada empresa, apreenda qual o modelo mais indicado, criando assim o modelo de previsão para essa empresa.

Os resultados da previsão destes modelos serão depois disponibilizados através de um relatório dinâmico de Power BI, fornecendo uma análise estatística sobre os conjuntos de dados fornecidos pelas empresas clientes e sobre os resultados de previsão.

1.4 Análise de Valor

A CPCIT4all, através do seu serviço de previsão de faturas pagas em atraso apresenta uma solução para a área contabilística, mais especificamente o processo de cobrança de faturas para empresas do setor de serviços públicos de fornecimento de gás e de eletricidade. Em empresas como a GoldEnergy, fornecedores dos dados para este projeto, devido ao seu grande número de clientes, é importante identificar os clientes que possivelmente terão um comportamento negativo, ou seja, mais suscetíveis a realizar o pagamento em atraso.

No caso da GoldEnergy mais de 38 milhões de euros em média por ano eram cobrados em atraso, provocando problemas no fluxo de caixa da empresa.

Através do uso do modelo de previsão, as empresas conseguirão identificar quais são estes clientes, podendo assim interagir com eles de forma a minimizar o impacto no fluxo de caixa da empresa.

1.5 Metodologia

O mecanismo automático a construir, necessita de criar e treinar modelos de previsão. Este treino de modelos implementa a metodologia do processo CRISP-DM. Este processo é composto por seis etapas:

- **Compreensão do negócio** (Business Understanding): Compreensão dos objetivos do projeto e os seus requisitos;

- **Compreensão dos dados** (Data Understanding): Familiarização com o conjunto de dados iniciais, identificação da qualidade dos dados;

- **Preparação dos dados** (Data Preparation): Limpeza dos dados, seleção e extração de atributos;

- **Mineração dos dados** (Modeling): Utilização de algoritmos de aprendizagem automática para criação de modelos;

- **Avaliação** (Evaluation): Avaliação dos modelos criados através do uso de métricas, identificação de problemas não identificados nas etapas anteriores e a verificação se os resultados atendem os objetivos do negócio;

- **Implantação** (Deployment): Colocação dos modelos resultantes em produção e configurar o sistema para realizar uma mineração contínua dos dados.

Os modelos a criar são modelos de classificação binária, com os valores "pago dentro do limite" e "pago depois do limite". Na etapa de mineração de dados, são usados vários algoritmos de aprendizagem supervisionada em conjunto com diferentes técnicas de balanceamento de dados para serem criados modelos de aprendizagem automática.

Estes modelos são avaliados segundo a métrica precisão geral (accuracy) e revocação (recall), dando prioridade à métrica de recall uma vez que o custo de errar a classificação de uma fatura paga em atraso é muito superior do que o custo de errar a classificação de uma fatura paga dentro do limite de tempo.

1.6 Estrutura do Documento

Este documento apresenta uma estrutura dividida em 6 capítulos gerais, sendo cada um deles representado por um título: Introdução, Contexto, Fundamentos e Estado de Arte, Descrição Técnica, Experiências e Conclusão.

- A Introdução é composta por uma curta apresentação do contexto, do problema, dos objetivos a realizar com esta dissertação, da análise de valor e a metodologia aplicada.

- O Contexto apresenta o problema em questão, a solução encontrada e a análise de valor da solução.

- O capítulo de Fundamentos e Estado de Arte, apresenta os fundamentos necessários para o leitor compreender a informação do documento, apresenta uma análise sobre trabalho relacionado, onde são estudados três estudos sobre o mesmo problema que enfrentamos, é também apresentado

as tecnologias utilizadas e termina com uma análise comparativa do trabalho relacionado.

- A Descrição Técnica apresenta a engenharia de requisitos, o caso de estudo, seguido da análise e design arquitetural da solução.

- O capítulo de Experiências apresenta as experiências da criação dos modelos de aprendizagem automática, apresentando para cada experiência, a sua configuração e os seus resultados. E concluindo com uma análise comparativa entre as experiências.

- No capítulo Conclusão, são referidas as considerações finais do projeto, apresentado um balanço crítico, bem como algumas possíveis abordagens futuras e são discutidos os resultados.

2 Contexto de Negócio

O setor de serviços públicos está a estabelecer-se como um dos setores mais inovadores a utilizar tecnologias emergentes e avançadas, como automação, inteligência artificial e análises de dados avançadas [2]. Neste capítulo é descrito o problema em questão, a sua solução, a área de negócio enquadrada no projeto e a proposta de valor da solução desenvolvida.

2.1 Problema

A CPCIT4all deseja fornecer um serviço de previsão para os seus clientes no setor de serviços públicos, para prever se os clientes destas empresas irão pagar em atraso as suas faturas de energia.

Os pagamentos em atraso das faturas é um dos principais desafios das operações de uma empresa. Com uma gestão inadequada do processo de cobrança de faturas, os pagamentos em atraso podem-se acumular e causar problemas nos negócios. Por outras palavras, o aumento do número de faturas não pagas pode levar a problemas de fluxo de caixa na empresa.

A Figura 1 e 2 ilustram o problema descrito para o caso da Goldenergy, que contêm mais de 450 mil clientes industriais e domésticos.

Na Figura 1 é apresentado a percentagem de faturas pagas dentro e fora do limite, onde é possível verificar que 16% das faturas anuais são pagas em atraso.

Na Figura 2, é apresentada a percentagem do valor da soma das faturas pagas dentro e fora do limite ao longo dos anos, onde é possível verificar uma média de 25% dos valores faturados ao longo do ano são realizados em atraso, isto significa, que para o caso da Goldenergy, que em média, mais de 38 milhões de euros por mês são cobrados depois da data limite de pagamento das faturas.

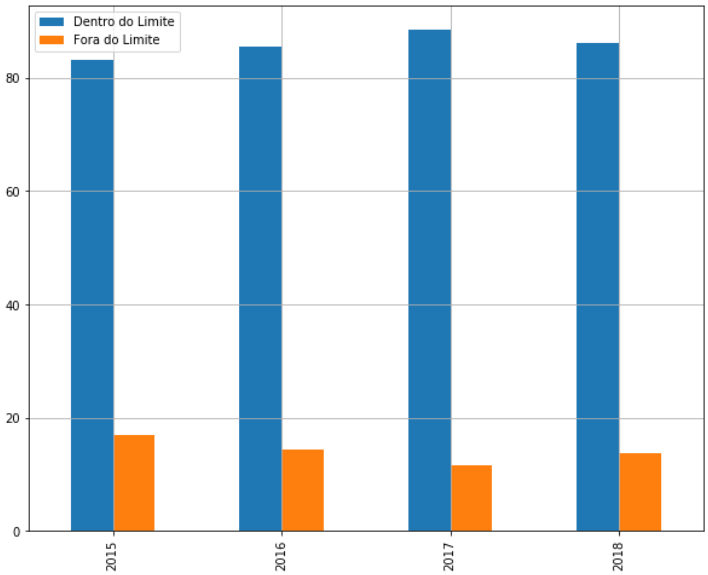

Figura 1 – Percentagem do número de faturas pagas dentro e fora no período 2015-2018

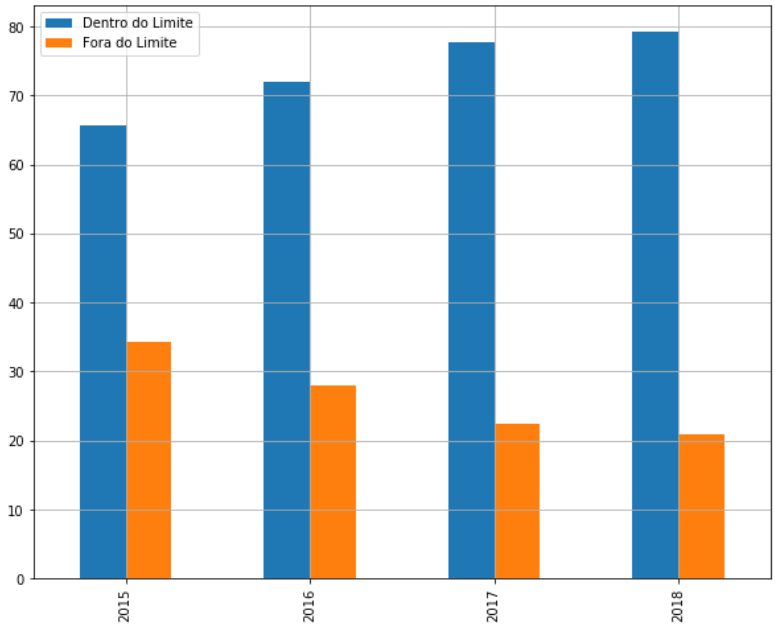

Figura 2 – Percentagem da soma de valores das faturas dentro e fora do limite no período 2015-2018

Como referido a Goldenergy contém mais de 450 mil clientes e contém muita informação sobre estes, sendo fornecido para este projeto mais de 9 milhões de registos. A quantidade de dados é tal que um humano não os consegue processar para extrair valor desta quantidade de informação e sendo uma base de dados real, a existência de incoerências e ruido é inevitável, sendo necessário um tratamento dos

dados profundo, uma vez que dados de baixa qualidade criam modelos de baixa qualidade.

Como existem dois tipos de clientes, os clientes novos e os clientes recorrentes. Os clientes novos são os clientes que realizam o primeiro pagamento e os clientes recorrentes são clientes com mais de um pagamento já realizado, sendo necessário a criação de um modelo para cada tipo de cliente.

A CPCIT4all tendo vários clientes industriais, pretende fornecer um serviço para lidar com estes problemas, sendo necessário uma ferramenta genérica que resolva os problemas de previsão e análise de dados específicos de cada empresa cliente da CPCIT4all.

2.2 Solução

As tecnologias de Inteligência Artificial, como aprendizagem automática e análise preditiva são mecanismos capazes de lidar com o tipo de problema que a CPCIT4all enfrenta, são capazes de detetar padrões e realizar previsões sobre grande quantidade de dados, fornecendo assim informação relevante sobre o negócio da empresa e informações sobre os seus clientes.

Através de um mecanismo automático de treino de modelos de previsão, é possível o fornecimento de modelos de previsão personalizados para a empresa cliente, atualizando os resultados de previsão de forma automática.

Os resultados da previsão destes modelos serão depois disponibilizados através de um relatório dinâmico de Power BI, onde é fornecido uma análise estatística sobre o conjunto de dados, ou seja, uma análise estatística sobre os clientes e é fornecido também uma análise estatística sobre os resultados de previsão.

Ajudando assim a empresa a ter uma melhor compreensão das faturas em atraso e da sua relação com os clientes, uma vez que para aqueles que realizam pagamentos em atraso muito frequentemente, a empresa atuará sobre estes de forma a resolver este problema.

2.3 Análise de Valor

Nesta secção é apresentado o modelo NCD, os conceitos de valor, onde são apresentados os benefícios e os custos da solução, os conceitos de valor para o cliente e valor percecionado. Esta secção termina com a apresentação do modelo CANVAS.

2.3.1 Modelo Fuzzy Front End

A primeira fase do processo de inovação é denominada de Fuzzy Front End. A Figura 3 apresenta as fazes do modelo FFE.

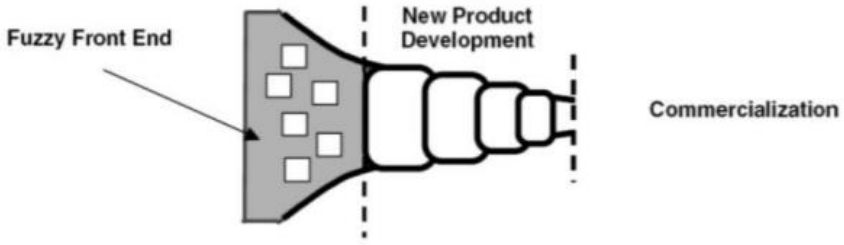

Figura 3 – Modelo Fuzzy Front End

- **FFE** - Fase experimental da inovação. Pode ser caótica uma vez que não apresenta grande organização. Baseada nos momentos "Eureka". Tem como objetivo identificar possíveis oportunidades de negócio;

- **New Product development (NPD)** - Método disciplinado e orientado a objetivos. Tem como objetivo aperfeiçoar o trabalho feito no ponto anterior;

- **Comercialização** - Promoção e venda do produto criado.

O modelo New Concept Development (NCD), foi desenvolvido para proporcionar uma linguagem e terminologia necessária para otimizar o processo de inovação FFE apresentado anteriormente.

2.3.2 Modelo New Concept Development – NCD

De forma sustentar a primeira fase do processo de inovação foi adotado o modelo teórico New Concept Development Model [3].

Caracteriza-se em três componentes, como ilustrado na Figura 4. Os cinco elementos chave **Identificação de Oportunidades** (Opportunity Identitifcation), **Análise de Oportunidades** (Opportunity Analysis), **Elaboração de ideias e Enriquecimento** (Idea Generation & Enrichment), **Seleção de ideias** (Idea Selection) e **Definição de Conceito** (Concept Definition), o motor (Engine) que faz uso dos cincos elementos chave, sob a liderança e cultura da organização e os fatores de influência.

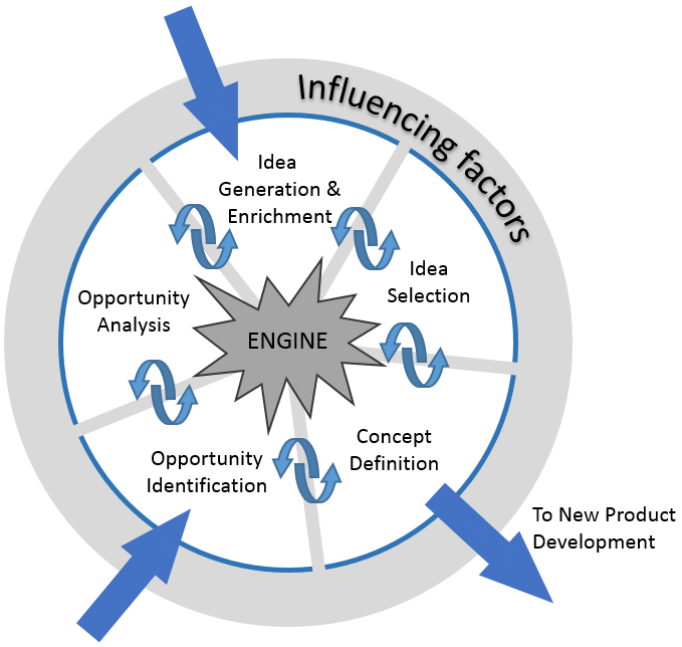

Figura 4 – Modelo NCD [3]

De seguida será apresentado os cinco elementos chave para nova conceção de um sistema avançado de previsão de pagamentos realizados em atraso.

2.3.2.1 Cinco elementos chave do modelo NCD

- **Identificação de Oportunidades** (Opportunity Identitifcation): Área de identificação de novas oportunidades. A identificação inclui a necessidade de atributos, tecnologias e modelo de negócio [3].

  Relativamente ao projeto:

  O pagamento em atraso das faturas é um problema muito comum no setor de serviços públicos, levando à criação de problemas de fluxo de caixa na empresa.

  Observando o valor apresentado pela solução modelo preditivo, os gestores das empresas deste setor apercebem-se da oportunidade de aumentar o processo de cobrança de faturas, aumento assim o fluxo de caixa da empresa.

- **Análise de Oportunidades** (Opportunity Analysis): Fase de análise das oportunidades identificadas no ponto anterior. A seleção da oportunidade a ser desenvolvida faz-se tendo em conta os atributos da empresa [3].

  Relativamente ao projeto:

Analisando o mercado, denota-se uma falta de oferta notória deste tipo de sistema sendo que os já existem não são fáceis de encontrar. É percetível, a necessidade de um sistema que possa diminuir ou até impedir o congestionamento do fluxo de caixa de uma empresa pertencente ao setor público.

- **Elaboração de ideias e Enriquecimento** (Idea Generation & Enrichment): Transformação das oportunidades em ideias de produtos. É um processo evolutivo em que existe combinação de pensamentos até se chegar a uma solução que atenda às necessidades dos clientes e à capacidade da empresa [3].

  Relativamente ao projeto:

  Identificada e avaliada a oportunidade, foi iniciado o processo de elaboração de ideias e do seu enriquecimento. Numa primeira fase, foi feito um estudo sobre que tipo de tecnologia se poderia utilizar para o efeito. Nesta, foi necessário identificar as capacidades que o sistema deveria possuir para que os clientes reconheçam o seu valor. Posto isto foram apresentadas as seguintes características:

  - Sistema automático de criação de modelos de previsão personalizados;
  - Apresentação dos resultados de previsão obtidos através de um relatório dinâmico, criado em Power BI [1], para uma visualização informativa e intuitiva por parte do cliente.

  No primeiro ponto é decidido o design do sistema a desenvolver de forma a que seja automático a criação de modelos de aprendizagem automática.

- **Seleção de ideias** (Idea Selection): Depois de elaborar as ideias é necessário selecionar uma ou mais ideias para desenvolvimento [3].

  Relativamente ao projeto:

  Analisando as ideias anteriormente descritas, foi feita a seleção que seria mais aconselhável. Posto isto, será necessário utilizar todas as ideias descritas, uma vez que fazem todas parte de um sistema completo;

- **Definição de Conceito** (Concept Definition): Um conceito apresenta uma forma escrita ou visual com características e benefícios únicos [3].

  Relativamente ao projeto:

  Com este sistema pretende-se fornecer um serviço de criação automática de modelos de previsão, e através destes modelos melhorar o processo de cobrança de faturas do

cliente. O seu uso diminuirá o número de clientes que realizarão pagamentos das suas faturas em atraso, o que levará a um aumento do fluxo de dinheiro da empresa.

2.3.3 Valor, Valor percecionado e Valor para o Cliente

A procura de vantagem das organizações face aos concorrentes é, cada vez mais, um objetivo crucial. A criação de valor é uma das formas de distinção de terceiros. Este é uma componente chave para a sustentabilidade e crescimento de uma organização, porém, é também um conceito de difícil alcance [4].

2.3.3.1 Valor

O valor pode ser caracterizado como a razão entre os benefícios e os custos/sacrifícios [4]. Dentro dos benefícios, incluem-se benefícios práticos e intangíveis. Já nos custos, associam-se custos monetários, de tempo, de energia e intangíveis. Valor resume-se, então:

$$Valor = \frac{Benefícios}{Custos} = \frac{\text{práticos} + \text{intangiveis}}{\text{monetários} + \text{tempo} + \text{energia} + \text{intangiveis}} \quad [4]$$

A Tabela 1 apresenta os benefícios e custos do projeto desta dissertação:

Tabela 1 – Comparação dos benefícios com custos para calcular o valor

	Serviço
Benefícios	✓ Previsão do pagamento em atraso das faturas dos clientes; ✓ Análise estatística; ✓ Aumento de fluxo de dinheiro da empresa;
Custos	✓ Investimento inicial do sistema e hardware; ✓ Manutenção do software e servidores;

2.3.3.2 Valor Percecionado e Valor para o Cliente

O conceito de **valor percecionado** ou valor atribuído, segundo Lindgreen e Wynstra [5], define-se como o valor que cada cliente assume que um determinado produto tem para si e para o seu negócio. Este valor pode ser diferente do valor percecionado pelo produtor. Uma vez que o produtor é mais sensível à qualidade do produto, o cliente assume uma posição mais cuidada no que toca ao preço do serviço oferecido. Os possíveis clientes deste produto poderão ter diversas opiniões do mesmo.

O **valor para o cliente** segundo Woodall [6], é a vantagem comercial que o cliente sente em comparação com a concorrência ao utilizar um determinado produto. Esta vantagem tanto pode ser apenas uma redução de esforço de produção, como, existência de benefícios, o resultado de uma operação Benefício/Custo com resultado positivo, entre outros.

O serviço fornecido pelo nosso produto, concede a possibilidade da criação e treino de modelos de previsão de forma automática e à empresa que o utilize, a possibilidade de aumentar a velocidade dos processos de cobrança de pagamentos de faturas, aumentado assim o fluxo de caixa da empresa. Contudo assumindo que o custo inicial é relativamente alto, o cliente deve projetar a diferença Benefício/Custo a longo prazo para avaliar o valor do produto para si.

2.3.4 Modelo Canvas

O modelo CANVAS é uma ferramenta de gestão estratégica que permite desenvolver modelos de negócio, quer estes sejam novos ou existentes [7]. É constituído por nove blocos individuais que podem ser divididos em duas áreas fundamentais: Lado FrontEnd e Lado BackEnd. Apesar da distinção, todos os blocos comunicam entre si, complementando-se. Estes blocos são:

- **Segmentos de Clientes**: Especifica quem vai ser o cliente alvo do negócio a criar. Define-se quais os clientes mais importantes;
- **Proposta de Valor**: Explica qual a proposta de valor que o negócio apresenta aos possíveis clientes. Define que tipos de problemas vão ser respondidos e quais as necessidades dos clientes que serão satisfeitas;
- **Canais de distribuição**: Define quais os canais que serão usados para chegar aos clientes;
- **Relações com clientes**: Neste bloco é esclarecido qual o tipo de relação que o cliente espera ter da empresa fornecedora do negócio;
- **Fontes de receita**: Especifica de que forma e pelo que que os clientes irão pagar;
- **Atributos Chave**: Define quais os atributos chave necessários para a proposta de valor apresentada;
- **Atividades Chave**: Distingue quais as atividades chave necessárias para manter os clientes interessados no negócio e para fornecer os serviços definidos na proposta de valor;

- **Parceiros Chave**: Enumera quais os parceiros mais importantes para fornecer o serviço/produto;

- **Estrutura de custos**: Especifica quais os custos mais importantes para o modelo de negócio.

  A Tabela 2 apresenta a adaptação do modelo CANVAS ao projeto apresentado. Relativamente aos parceiros chave, é importante denotar que durante a execução deste projeto ainda se estavam a contactar possíveis parceiros, não sendo ainda possível especificar os mesmos.

Tabela 2 – Modelo Canvas

<table>
<tr>
<td rowspan="2">Parceiros Chave
✓ Ainda por definir;</td>
<td>Atividades Chave
✓ Manutenção de software;
✓ Melhoramento das capacidades de previsão;</td>
<td rowspan="2">Proposta de Valor
✓ Previsão dos atrasos das faturas dos clientes;
✓ Fornecimento de uma interface intuitiva com análises sobre os clientes
✓ Redução do número de faturas em atraso;
✓ Aumento do fluxo de dinheiro da empresa;</td>
<td>Relações com os Clientes
✓ Suporte pós-venda;
✓ Constante melhoria dos algoritmos de previsão;
✓ Substituição de hardware defeituoso;</td>
<td rowspan="2">Segmentos do Clientes
✓ Indústrias da área de fornecimento de energia;</td>
</tr>
<tr>
<td>Atributos Chaves
✓ Servidores Cloud;
✓ Internet;</td>
<td>Canais de comunicação
✓ Demonstrações em clientes;</td>
</tr>
<tr>
<td colspan="3">Estrutura de custos
✓ Equipa de desenvolvimento e manutenção do software;
✓ Alocação Cloud;</td>
<td colspan="3">Fonte de Receitas
✓ Serviços de desenvolvimento à medida;
✓ Venda do sistema por trabalhador monitorizado;
✓ Licenciamento da solução;</td>
</tr>
</table>

3 Fundamentos e Estado de Arte

Neste capítulo serão descritos os fundamentos necessários para a compreensão da informação por parte do leitor e o estudo do estado de arte, onde serão apresentadas quatro soluções semelhantes existentes no mercado, a comparação do trabalho relacionado e as tecnologias utilizadas.

3.1 Fundamentos de Aprendizagem

Nesta secção vão ser apresentados os fundamentos necessários para a solução. A nossa solução utiliza o processo CRISP-DM (Cross-Industry Standard Process for Data Mining). Este processo é composto por seis etapas:

- **Compreensão do negócio** (Business Understanding): Compreensão dos objetivos do projeto e os seus requisitos;
- **Compreensão dos dados** (Data Understanding): Familiarização com o conjunto de dados iniciais, identificação da qualidade dos dados;
- **Preparação dos dados** (Data Preparation): Limpeza dos dados, seleção de atributos;
- **Mineração dos dados** (Modeling): Utilização de algoritmos de aprendizagem automática para criação de modelos;
- **Avaliação** (Evaluation): Avaliação dos modelos criados através do uso de métricas, identificação de problemas não identificados nas etapas anteriores e a verificação se os resultados atendem os objetivos do negócio;
- **Implantação** (Deployment): Colocação dos modelos resultantes em prática e configurar o sistema para realizar uma mineração contínua dos dados.

A Figura 5 ilustra as etapas do processo CRISP-DM.

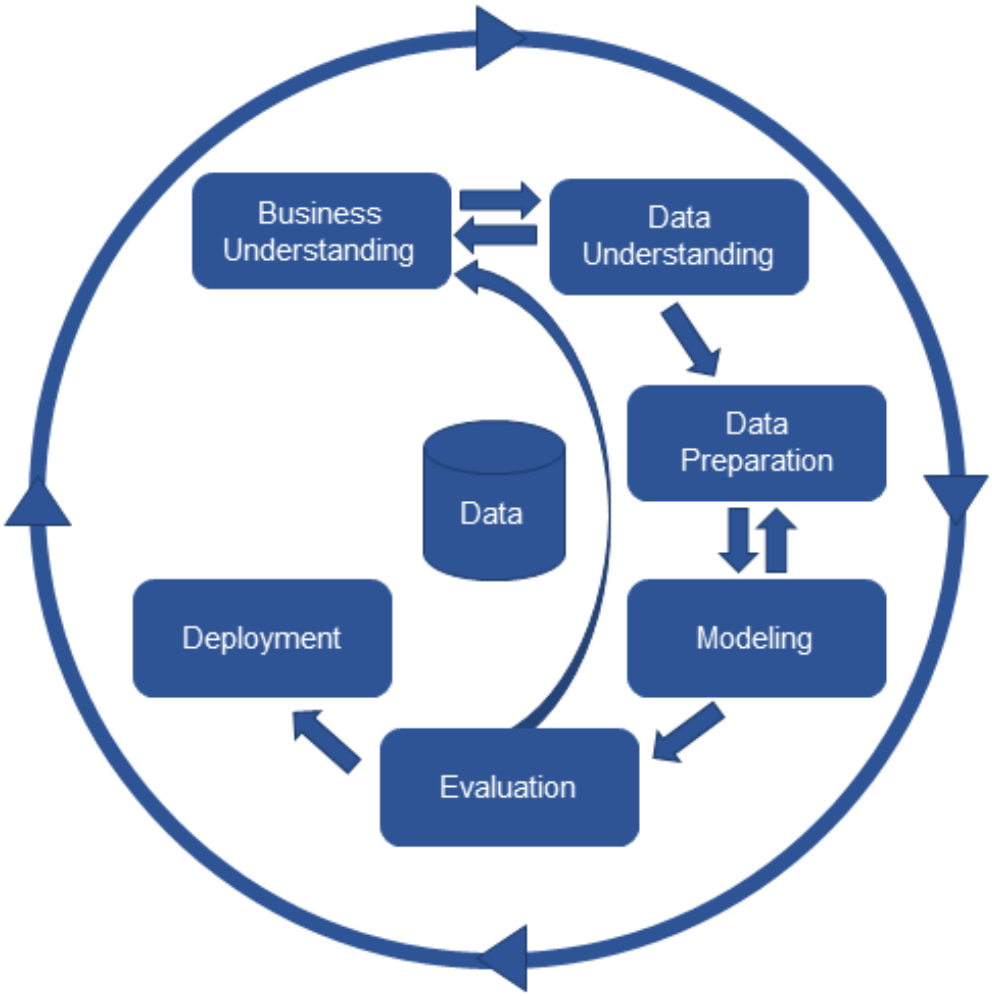

Figura 5 – Etapas do processo CRISP-DM [8]

De seguida será apresentado o processo de descoberta do conhecimento, processo este incluído no modelo CRISP-DM.

Serão também apresentados os algoritmos de Classificação utilizados nesta dissertação, assim como as técnicas de deteção de anomalias e as métricas de avaliação dos modelos de aprendizagem automática.

O capítulo termina com uma apresentação do *Cold-Start Problem*, problema relacionado com a extração de informação sobre os novos clientes.

3.1.1 Descoberta do Conhecimento

A quantidade de dados recolhidos e armazenados ao longo do tempo tem crescido de forma considerável em praticamente todas as áreas da sociedade. Com o aumento do número de pessoas e cidades ligadas através da internet, os conjuntos de dados são cada vez maiores. Segundo o estudo "The Digitization of the World" [9] é estimado que o tamanho total de dados digitais no mundo no ano de 2025 será de aproximadamente 175 Zeta bytes, conforme ilustrado na Figura 6.

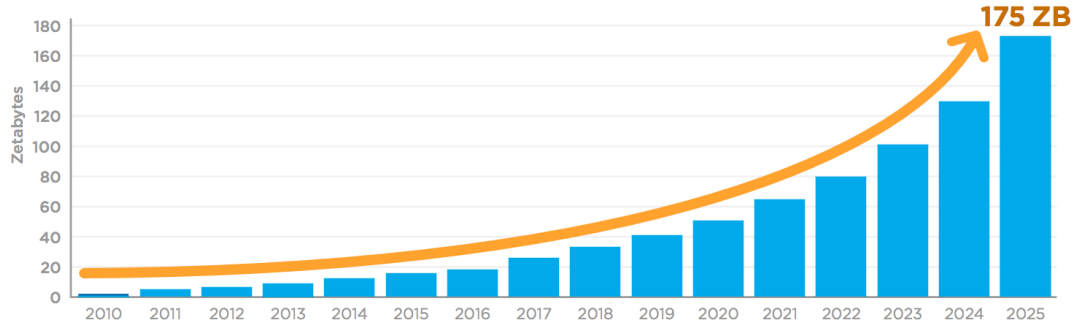

Figura 6 – Tamanho anual de dados digitais globais [9]

Um exemplo disso é o aumento de dados de clientes no setor público onde a GoldEnergy, cliente da CPCIT4all, se insere.

Nesta situação particular, tal como em muitas outras, o processamento dos dados é uma tarefa extremamente dispendiosa e em alguns casos até impossível. O crescente interesse na área de descoberta automática de conhecimento deve-se aos problemas referidos.

A Descoberta de Conhecimento em Base de Dados, **Knowledge Discovery in Databases** (**KDD**), trata-se de um processo de identificação de com valor para o universo em questão [10].

Neste processo, o conjunto de dados representam um conjunto de fatos. Os padrões entre eles representam o conhecimento extraído do universo em questão caso utilizador considere que estes padrões representem informação relevante [10].

Este processo é composto por três fases principais, ilustradas na Figura 7:

- Pré-processamento de dados;
- Mineração de dados;
- Pós-processamento de conhecimento;

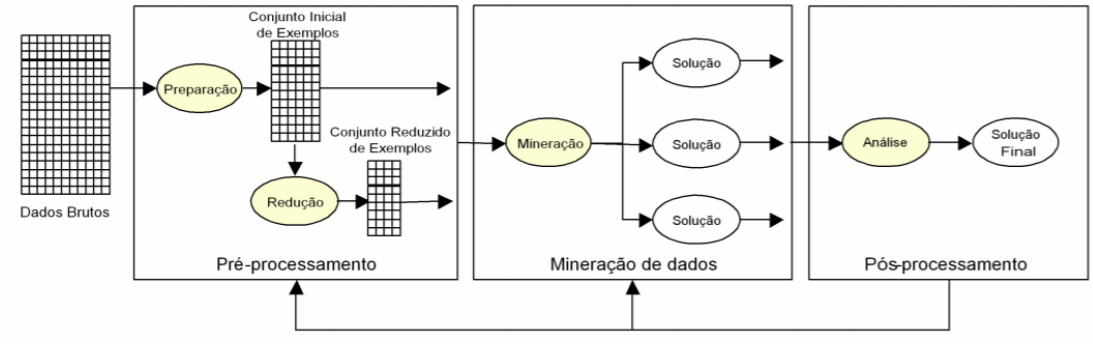

Figura 7 – Processo de Descoberta de Conhecimento [11]

3.1.1.1 Pré-Processamento de Dados

As bases de dados do mundo atual são altamente suscetíveis a dados com ruído, ausentes e inconsistentes devido ao seu tamanho normalmente enorme (geralmente vários gigabytes ou mais) e à sua provável origem de várias fontes heterogêneas. Dados de baixa qualidade levarão a resultados de mineração de baixa qualidade [12].

Esta fase visa essencialmente entender a qualidade dos dados de forma a prepará-los para a fase seguinte. Para este objetivo são realizadas as seguintes tarefas:

- Integração dos dados (Data Integration)

A integração dos dados consiste em agrupar dados de múltiplas fontes numa base de dados comum de forma coerente. Nesta tarefa existe um vasto número de questões a serem consideradas, como o esquema de integração, as possíveis redundâncias nos dados, assim como a deteção de eventuais conflitos [13].

- Limpeza dos dados (Data Cleaning)

A limpeza dos dados consiste na remoção de inconsistências e na correção dos erros nos dados. Pode igualmente ser necessário efetuar o preenchimento de valores em falta, assim como identificar ou remover os dados que não se enquadrem no universo em questão [13].

Se os utilizadores não considerarem os dados a utilizar fiáveis, então também não vão confiar nos resultados do processo de mineração, uma vez, os resultados poderão ser pouco precisos e com baixo valor.

- Seleção dos dados/Redução de dados (Data Reduction)

A tarefa de seleção dos dados consiste na aplicação de técnicas para reduzir o universo de dados em estudo, mantendo a integridade original dos dados. A mineração do grupo de dados reduzido deverá ser mais eficiente e produzir os mesmos resultados analíticos [13].

Existem várias estratégias utilizadas na seleção dos dados, nesta dissertação vamos utilizar as seguintes:

  - Filtro de baixa variâncias: os atributos que contêm poucos valores diferentes fornecem pouca informação [14].

  - Filtro de alta correlação: Atributos que contêm uma correlação alta com outros atributos fornecem informação muito semelhante à fornecida pelos outros atributos, então apenas um destes atributos será necessário para obter a informação [14].

  - Utilizar algoritmo Floresta Aleatória: Algoritmos como a Floresta Aleatória contêm funções para encontrar os atributos que fornecem mais informação. Estas funções constroem as árvores contra o atributo de destino e usa estatística de uso de cada atributo para encontrar o subconjunto de atributos que fornecem mais informação [14].

- Transformação dos dados (Data Transformation)

A tarefa de transformação dos dados consiste em transformar os dados em formatos apropriados para o processo de mineração através de operações de agregação, generalização, normalização ou discretização [13].

É nesta tarefa que a maioria dos erros são corrigidos, nomeadamente os erros com origem humana, por exemplo os erros de processamento de dados incorretos. Nos casos em que são encontradas discrepâncias, é necessário definir e aplicar uma série de transformações para os corrigir, tais como:

  - **Suavizar** (Smoothing): Remoção de ruído dos dados. Exemplos de técnicas para este problema são:

    - **Uso de uma constante global para preencher o valor ausente** [12], por exemplo quando existem valores ausentes, substituir por valores constantes como “other” ou “unknown”;

    - **Uso de uma medida de tendência central para o atributo** [12], por exemplo, quando se trata de dados numéricos, utilizar a média ou mediana para preencher o valor ausente.

- **Agregação**: Resumir um conjunto de valores num único, aplicando operações de agregação aos dados, através de operações aritméticas (média; máximo; soma; entre outros) [12].

- **Generalização**: Generalização dos dados através da aplicação de hierarquias de conceitos entre os dados [12], por exemplo, o valor "Maia" passa a ser generalizado para "Porto".

- **Normalização**: Dados são dimensionados de forma a serem inseridos em intervalos de referência, normalmente entre -1.0 e 1.0 ou 0.0 e 1.0 [12].

- **Construção de Atributos**: São criados atributos a partir de um conjunto de dados, com o objetivo de melhorem o processo de mineração [12]. Exemplo destas técnicas para extração de novos atributos são:

    - **Divisão de atributos**, tal como o nome indica, esta técnica divide um atributo em vários [12], por exemplo dividir uma data em dia, mês e ano.

    - **Binning**, é a técnica de agrupar valores dos atributos em grupos [12], por exemplo um atributo "país" com o valor "Portugal", podemos criar um atributo "Continente" como o valor "Europa".

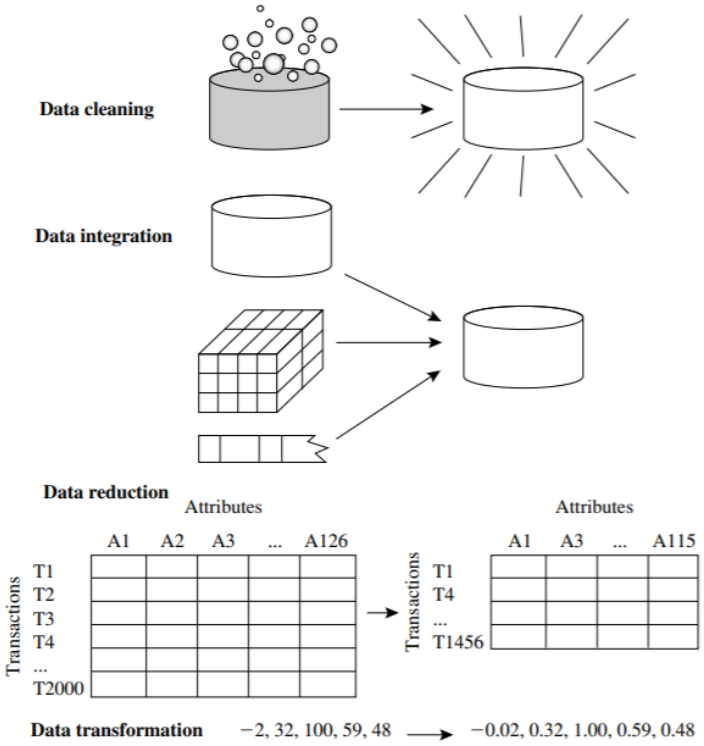

Figura 8 – Fases de Pré-Processamento de Dados [12]

Na fase de pré-processamento, ilustrada anteriormente na Figura 8, os métodos de visualização de dados, assim como a utilização de estatísticas descritivas (médias, desvios padrão), assumem um papel fundamental no conhecimento prévio dos dados, podendo mesmo auxiliar na seleção dos algoritmos mais adequados para a fase de mineração [13].

O pré-processamento é a atividade que requer mais esforço ao longo de todo o processo de descoberta de conhecimento. Segundo o autor Dorian Pyle estima-se que cerca de 80% do tempo despendido em todo o processo de modelagem preditiva seja utilizado no pré-processamento de dados [15].

3.1.1.2 Mineração de Dados

A mineração de dados abrange todo o processo de descoberta do conhecimento e da aplicação de algoritmos específicos para a extração de padrões do conjunto de dados.

As funcionalidades de mineração de dados incluem:

- **Caracterização e distinção**: Consistem em descrever classes e conceitos individuais em termos resumido e concisos. Estas descrições podem alcançadas através das técnicas, **caracterização dos dados** que consiste no sumário da classe

alvo dos dados em estudo, também podem ser alcançadas através da técnica **distinção de dados**, através da comparação da classe alvo com uma classe comparativa, ou com um conjunto destas [12].

- **Mineração de padrões, associações e correlações frequentes**: Consiste na mineração de padrões frequentes que existam nos dados em estudo. Os padrões encontrados levam à descoberta de associações e correlações interessantes nos dados.

- Classificação e Regressão:

A **classificação** é o processo de encontrar um modelo que descreva e distinga **classes/categorias** de dados, normalmente de valor **discreto**. O modelo é gerado através do uso de algoritmos de aprendizagem automática e com um conjunto de dados, designado por conjunto de treino, onde as classes dos atributos são previamente conhecidas, pertencendo assim ao tipo de aprendizagem supervisionada. Depois de encontrado esse modelo, é possível aplicá-lo de forma a prever a classe/categoria de um novo objeto [12].

A **regressão** prevê dados, normalmente de valor **contínuos** em vez de categóricos como a classificação. Ou seja, os modelos de regressão são usados para prever valores de dados numéricos ausentes ou indisponíveis [12].

Para a execução da tarefa de classificação é possível aplicar uma série de algoritmos de aprendizagem automática, nomeadamente: **árvores de decisão** [16], **regressão logística** [17], **máquina de vetor de suporte** [18], **redes neurais** [12], **GBM** [19] e algoritmos de conjuntos como as **florestas aleatórias** (Random Forest) [17], entre outros. Na Figura 9 são apresentados exemplos de diferentes algoritmos de classificação para um mesmo problema. Neste caso particular é relacionada a idade de um indivíduo X e o seu rendimento, inserindo-o numa determinada categoria (classe A, B ou C).

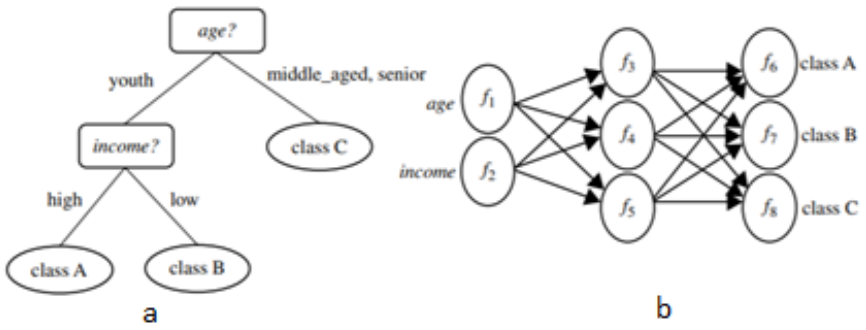

Figura 9 – Representação do modelo de classificação: (a) Árvores de decisão, (b) Redes Neurais [12]

- **Análise de agrupamentos**: Ao contrário da classificação e da regressão, que analisam conjuntos de dados com os atributos classificados, a análise de agrupamentos analisa conjuntos de dados sem se conhecer a classe dos atributos, sendo então um método de aprendizagem não supervisionada.

Normalmente, os algoritmos utilizados neste tipo de tarefa são aqueles que utilizam alguma medida de distância entre pontos, como o K vizinhos mais próximos (KNN) [20]. O objetivo desses algoritmos é maximizar a distância entre grupos e simultaneamente minimizar a distância entre indivíduos do mesmo grupo, tal como ilustrado na Figura 10.

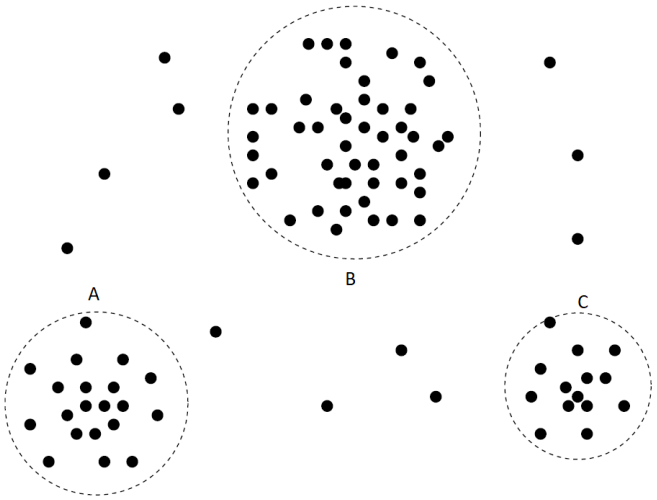

Figura 10 – Tarefa de agrupamento em que um conjunto de dados é dividido em três grupos [12]

- **Deteção de anomalias**: Um conjunto de dados pode conter objetos que não estão em conformidade com o comportamento geral ou modelo dos dados. Esses objetos de dados são anomalias. Muitos métodos de mineração de dados removem estes objetos porque os consideram como ruído ou exceções. No entanto, em alguns casos, por exemplo deteção de fraude e a deteção do pagamento em atraso de faturas, os eventos raros podem ser mais interessantes do que os que ocorrem com mais frequência. A esta tarefa é chamada mineração de anomalias ou mineração de casos raros.

O problema de deteção de anomalias será apresentado em mais detalhe na secção 3.1.3.

Tipo de Aprendizagem

As tarefas de mineração podem ser classificadas em duas categorias, a descrição e a previsão. As tarefas de mineração descritiva caracterizam propriedades dos dados num conjunto de dados alvo. As tarefas de mineração preditiva executam a indução nos dados atuais para fazer previsões [12].

Os **modelos descritivos** estão habitualmente associados à modelação de relações entre dados que não são previamente rotulados, ou seja, aprendizagem do tipo **não supervisionada**. Neste tipo de aprendizagem, a aprendizagem é efetuada descobrindo similaridades nos dados, ou seja, pretende-se encontrar agrupamentos de dados com características semelhantes. Os algoritmos de agrupamento (clustering) pertencem à aprendizagem não supervisionada [21].

Por exemplo, como está apresentado na Figura 11, um método de aprendizagem não supervisionada pode receber como entrada um conjunto de imagens de animais. Suponha que encontra 3 conjuntos de dados, esses conjuntos (clusters) podem corresponder a 3 raças de animais distintas. Contudo, como os dados de treino não estão classificados o modelo não nos consegue dizer qual a raça de cada conjunto encontrado.

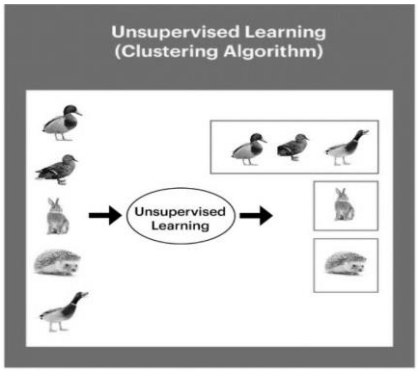

Figura 11 – Funcionamento de um algoritmo de aprendizagem não supervisionada [21]

Os **modelos de previsão** estão geralmente associados à modelação de relações entre dados que são previamente rotulados, ou seja, aprendizagem do tipo **supervisionada** [21].

Na Figura 12 é apresentado um exemplo de aprendizagem supervisionada de classificação, onde um conjunto de imagens de animais classificadas com a raça a que pertencem é usado como dados de treino. O modelo criado prevê a que raça uma imagem de um animal pertence.

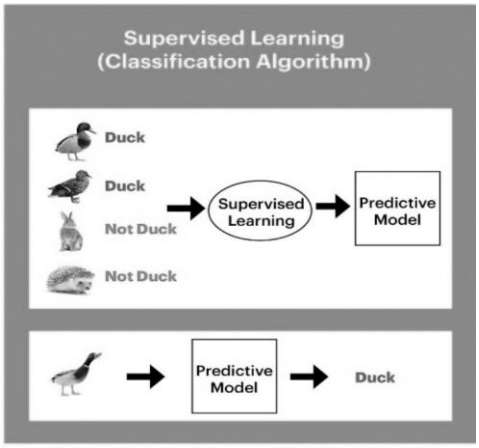

Figura 12 – Funcionamento de um algoritmo de aprendizagem supervisionada [21]

A importância, quer da previsão como da descrição para determinadas aplicações de mineração de dados, poderá variar consideravelmente dependendo da natureza dos dados e dos objetivos do utilizador.

Para o problema abordado nesta dissertação, o tipo de aprendizagem automática utilizado pertence à aprendizagem supervisionada.

3.1.1.3 Pós-processamento de conhecimento

O objetivo principal da fase de pós-processamento é avaliar, validar e consolidar o conhecimento extraído do processo de mineração de dados [11]. Recorrendo à tradução dos resultados em métricas pré-definidas, é possível efetuar uma avaliação do conhecimento extraído. Devem ser avaliados da mesma forma para garantir que os resultados são fiáveis e estatisticamente significativos.

Depois de avaliado e validado o conhecimento extraído é consolidado quando este conhecimento é associado a sistemas de apoio à decisão ou é disponibilizado ao utilizador através de documentação própria [13].

3.1.1.4 Sumário

Em suma, as três fases (pré-processamento de dados, mineração de dados e pós-processamento do conhecimento) são fundamentais para que o processo de descoberta de conhecimento seja bem-sucedido.

Este processo, como referido anteriormente, é um processo que se repete a si próprio até que o utilizador considere o modelo criado como sendo o melhor possível, isto é, não deve ser possível alterar os dados ou alterar o algoritmo utilizado, de forma a obter um modelo com melhores resultados comparando as métricas de avaliação.

3.1.2 Algoritmos de Classificação

A aprendizagem automática investiga como os computadores podem apreender (ou melhorar o seu desempenho) com base em dados. Uma área de pesquisa principal é que os programas de computador aprendam automaticamente a reconhecer padrões complexos e a tomar decisões inteligentes com base nos dados fornecidos [12].

Os algoritmos de aprendizagem automática são essenciais uma vez que conseguem detetar estes padrões e fornecer informação útil sobre um universo de dados extenso.

De seguida serão brevemente apresentados os diferentes algoritmos de aprendizagem automática de classificação utilizados nesta dissertação.

3.1.2.1 Árvores de Decisão

As árvores de decisão são algoritmos que utilizam a estratégia de divisão e conquista, ou seja, este algoritmo divide o conjunto de atributos em vários subconjuntos de menor dimensão (sem sobreposição) com valores de resposta semelhantes usando um conjunto de regras de divisão. No final, as soluções dos subconjuntos podem ser combinadas para gerar a solução do problema [22].

As árvores de decisão classificam instâncias ordenando-as desde a raiz até um determinado nó-folha, o qual designa a classificação da instância em causa. Cada nó na árvore especifica um determinado atributo da instância, enquanto cada ramo descendente corresponde a um dos possíveis valores para o atributo em questão. Uma instância é classificada começando pela raiz da árvore, testando o atributo definido pelo nó e posteriormente descendo o ramo correspondente ao valor do atributo dado. Todo este processo é depois repetido para a subárvore cuja raiz é um novo nó [13].

Exploramos agora o exemplo adaptado pelo autor [13] onde a Figura 13 ilustra uma árvore de decisão, classificando o dia como sendo possível ou não praticar ténis consoante os atributos meteorológicos.

Tabela 3 – Atributos e os seus possíveis valores

Atributos	Tempo	Humidade	Vento	Jogar
Valores	Sol	Alta	Forte	Não
	Nublado	Normal	Fraco	Sim
	Chuva			

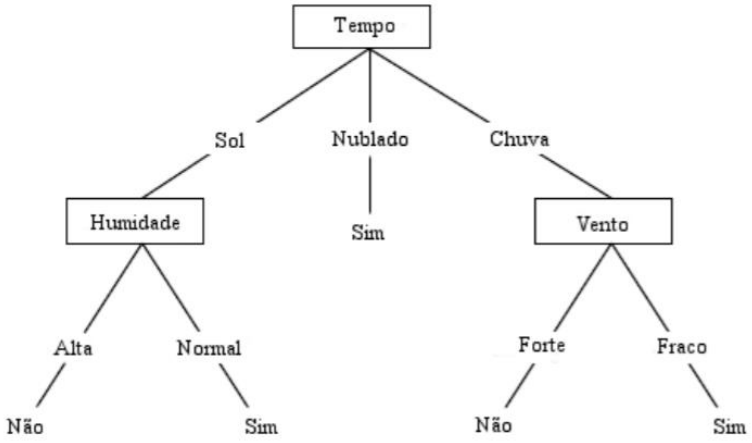

Figura 13 – Árvore de decisão que representa o conceito Jogar Ténis [13]

O resultado "Sim" da árvore presente na Figura 13 pode ser representada pela seguinte expressão:

$$Se(\,Tempo = Sol\ \cap Humidade = Normal)$$

$$\cup\,(Tempo = Nublado)$$

$$Se(\,Tempo = Chuva\ \cap Vento = Fraco)$$

Esta árvore pode também expressar as condições quando o resultado é "Não", nomeadamente quando está sol e a humidade está alta, ou então nos casos em que está a chover e o vento é forte.

3.1.2.2 Regressão Logística

A regressão logística é um modelo estatístico que, na forma básica, usa uma função logística para modelar uma variável dependente binária, embora existam muitas extensões mais complexas [23]. Matematicamente, um modelo logístico binário possui uma variável dependente com dois valores possíveis, como aceite/não aceite, representada por uma variável indicadora, na qual os dois valores são representados por "0" e "1" [12].

A regressão logística analisa dados distribuídos binominalmente da forma:

Regressão Linear: $Y = b0 + b1 * X1 + b2 * X2 \ldots . . + bk * Xk$

Função Sigmoide: $P = \frac{1}{1-e^{-Y}}$

Regressão Logística: $\ln\left(\frac{P}{1-P}\right) = Y$

Como podemos ver, a equação da regressão logística é muito semelhante à regressão linear. A diferença reside no fato de que a regressão linear fornece valores contínuos de Y para um dado X, a regressão logística também fornece valores contínuos entre 0 e 1 para um dado X, que é transformado posteriormente em Y = 0 ou Y = 1 com base no valor limite (0.5).

A Figura 14 ilustra um exemplo de uma função logística onde para um resultado com o valor 0.8 é atribuído à classe 1 e o resultado com valor 0.2 é atribuído à classe 0.

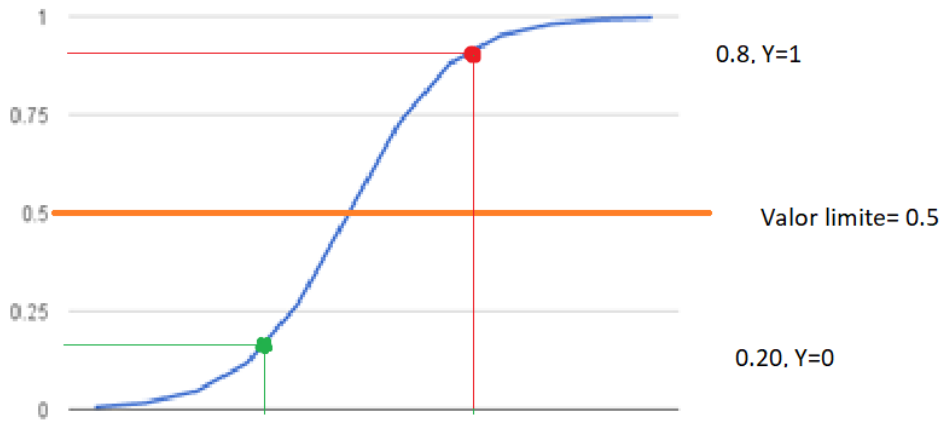

Figura 14 – Exemplo da representação gráfica de uma função Logística

3.1.2.3 One Class SVM

O algoritmo One-Class SVM, é um algoritmo que implementa a técnica de One-Class Classification, onde esta tenta identificar os objetos específicos de uma classe entre todos os objetos, através da aprendizagem de um conjunto de treino que contenha apenas objetos da classe que pretendemos identificar [24].

A Figura 15 é apresentado a representação gráfica de um algoritmo de One Class SVM onde o espaço de treino é limitado para a classe de treino e tudo o que fique fora desse espaço é considerado uma observação anómala.

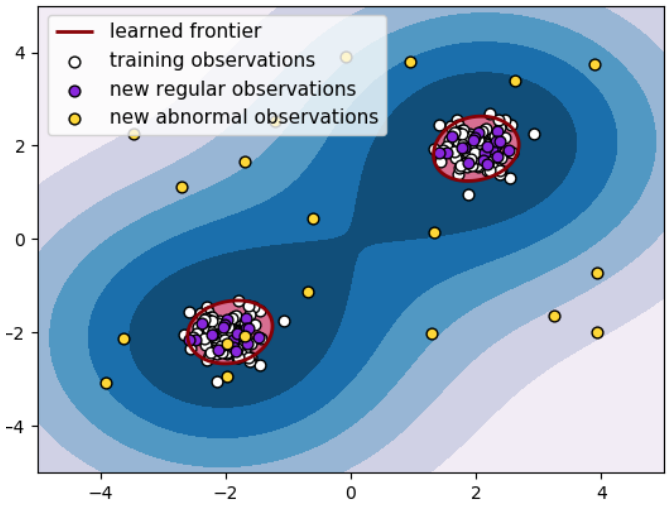

Figura 15 – Deteção de casos raros, do algoritmo One-Class SVM com um kernel não linear [25].

3.1.2.4 Floresta Aleatória

Neste algoritmo cada um dos seus classificadores atua como uma árvore de decisão, de modo que o conjunto de classificadores seja uma "floresta" dai o seu nome, Random Forest. As árvores de decisão individuais são geradas usando uma seleção aleatória de atributos em cada nó para determinar a divisão. Durante a classificação, cada árvore vota e a classe mais popular é retornada.

Exploramos agora um exemplo dado pelos autores Gustavo Machado, Mariana Recamonde Mendoza e Luís Gustavo Corbellini [17] sobre o estudo do uso do algoritmo floresta aleatória para a criação de um modelo de previsão de diarreia viral bovina.

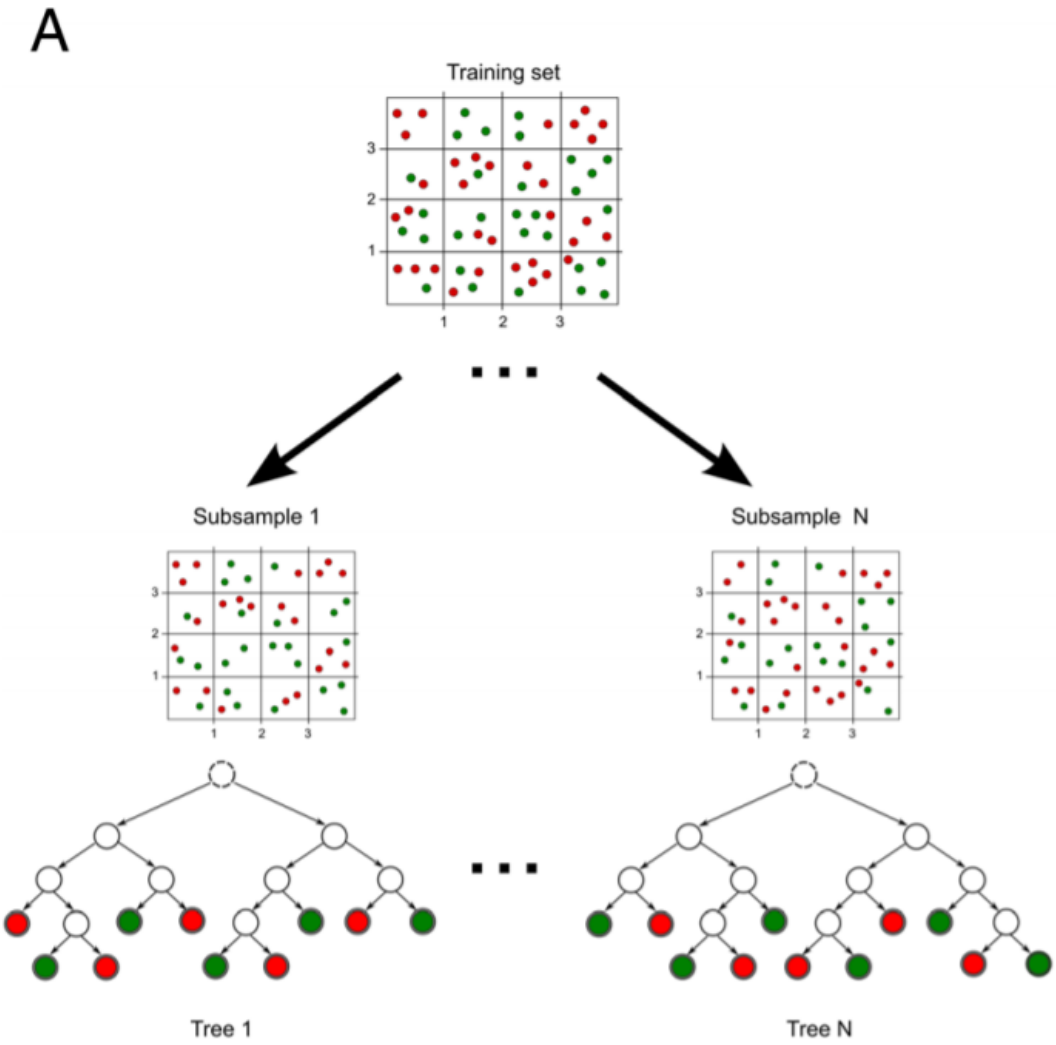

Figura 16 – Cada árvore de decisão no conjunto é construída sobre uma amostra aleatória dos dados originais, que contém exemplos positivos (rótulos verdes) e negativos (rótulos vermelhos) [17]

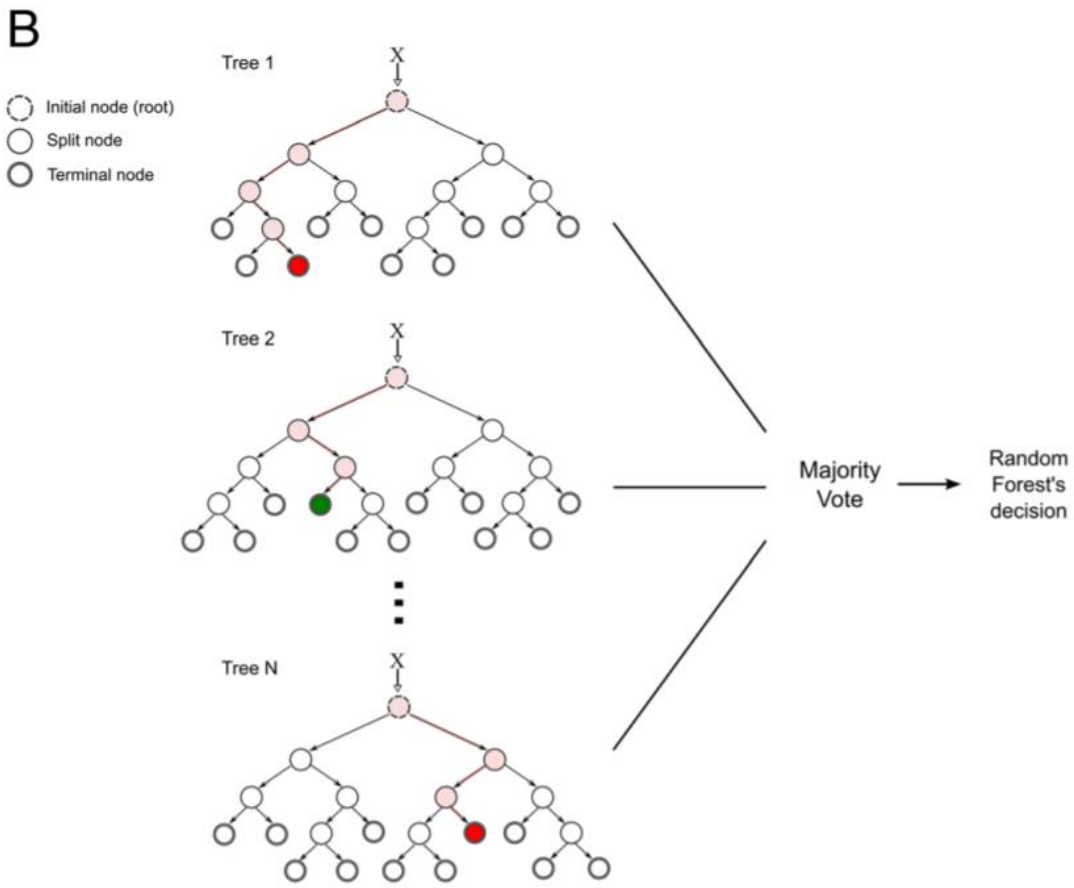

Figura 17 – Representação do funcionamento da lógica de previsão de um algoritmo floresta aleatória [17]

O modelo de previsão criado através do algoritmo floresta aleatória é baseado num procedimento de votação maioritária entre todas as árvores individuais. ´

Na Figura 16 é apresentado a forma de divisão dos dados de treino pelas arvores que pertencem a floresta, onde cada árvore de decisão no conjunto é construída sobre uma amostra aleatória dos dados originais, que contém exemplos positivos (rótulos verdes) e negativos (rótulos vermelhos).

Na Figura 17 é ilustrado o procedimento realizado para cada árvore. Onde para cada novo ponto de dados (ou seja, X), o algoritmo inicia no nó raiz de uma árvore de decisão e percorre a árvore (ramificações destacadas) testando os valores das variáveis em cada um dos nós divididos visitados (nós rosa), de acordo com cada um, seleciona o próximo ramo a seguir. Esse processo é repetido até que um nó folha seja atingido, o que atribui uma classe a esta instância: nós verdes representam a classe positiva, nós vermelhos representam a classe negativa. No final do processo, cada árvore emite um voto para o rótulo de classe preferido e o modo das saídas é escolhido como a previsão final [17].

3.1.2.5 Redes Neurais

As redes neurais artificiais são um método para solucionar problemas através da simulação do cérebro humano, inclusive em seu comportamento, ou seja, aprendendo, errando e fazendo descobertas [26]. São técnicas computacionais que apresentam um modelo inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência, logo quantos mais dados tivermos mais inteligente o modelo será.

As redes neurais possuem nós ou unidades de processamento. Cada unidade possui ligações para outras unidades, nas quais recebem e enviam sinais. Cada unidade pode possuir memória local e estas unidades são a simulação dos neurônios, recebendo e retransmitindo informações.

Os resultados são classificados com valores entre 0 e 1. No nosso caso é necessário classificar como pertencendo à classe 0 ou à classe 1, onde é utilizado, tal como na regressão logística, a função sigmoide [27] que para os valores contínuos obtidos são classificados como 0 ou 1. Depois de obtidos todas as votações, é escolhido como resultado a mais popular.

Uma rede neural pode possuir uma ou múltiplas camadas. A rede neural escolhida para o exemplo é constituída por uma camada de entrada (com seis variáveis), uma camada intermediária (com vinte neurônios) e uma camada de saída (com um neurônio) conforme a Figura 18. O neurônio de saída contém a classificação binaria (0 ou 1) para o conjunto de atributos.

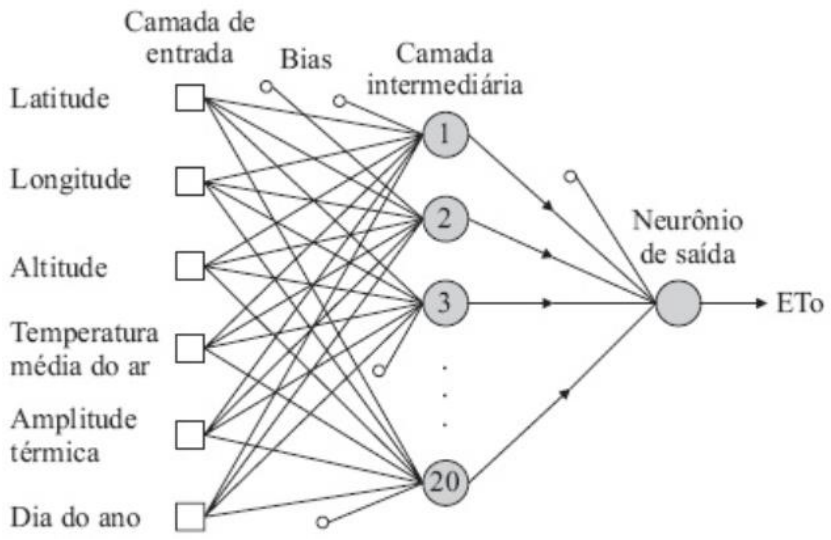

Figura 18 – Rede Neural [27]

3.1.2.6 Gradient Boosting Machine (GBM) e Light Gradient Boosting Machine (LightGBM)

As máquinas de aumento de gradiente (GBMs) são algoritmos de aprendizagem de automática extremamente populares que provaram ser bem-sucedido em muitos domínios e são um dos métodos mais utilizados nas soluções vencedoras as competições do Kaggle [28]. Enquanto florestas aleatórias constroem um conjunto de árvores independentes profundas, os GBMs constroem um conjunto de árvores rasas em sequência com cada árvore aprendendo e melhorando a anterior [19].

Embora as árvores rasas sejam, por si só, modelos preditivos bastante fracos, elas podem ser "impulsionadas (boosting)" para produzir um conjunto poderoso que, quando ajustado adequadamente, geralmente é difícil de superar com outros algoritmos. Este capítulo abordará os fundamentos para entender e implementar algumas implementações populares de GBMs.

Na Figura 19 é apresentado a abordagem do um algoritmo GBM, explicado anteriormente.

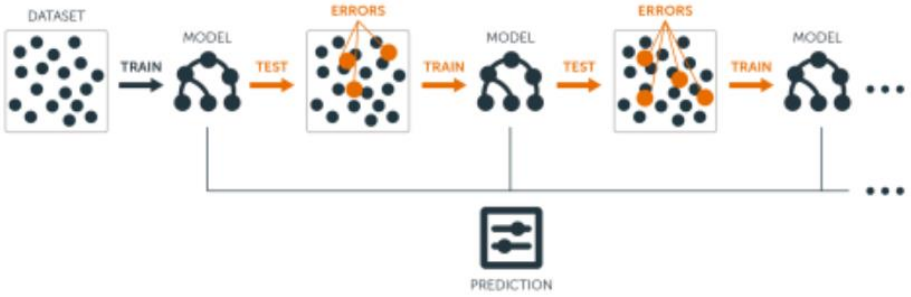

Figura 19 – Abordagem sequencial de conjuntos [22]

A Figura 20 ilustra a evolução do custo do erro através das adições das árvores sucessivamente de 0 a 1024 árvores. Como podemos ver, o custo do erro representado pela linha vermelha vai diminuindo à medida que o número de árvores aumenta. Obtendo, assim, a solução ótima na árvore número 1024, onde o custo do erro é praticamente nulo, sobrepondo-se à linha azul (valores verdadeiros).

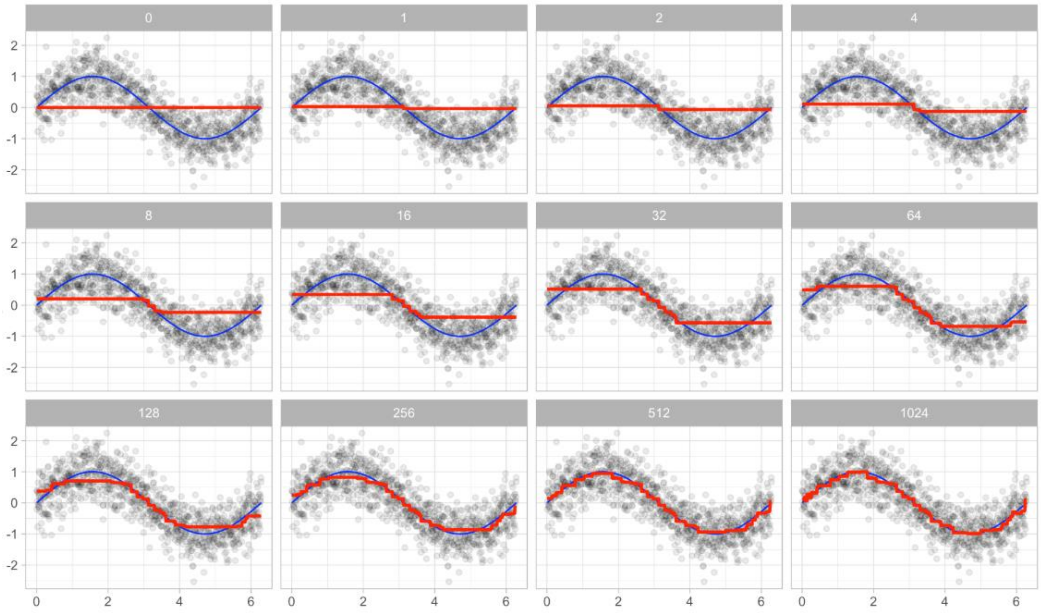

Figura 20 – Evolução do custo do erro (linha vermelha) através das adições das árvores sucessivamente de 0-1024 árvores [22]

O LightGBM é uma versão do GBM tradicional, onde a grande diferença está na forma da construção do conjunto de árvores. O LightGBM foca-se no "*leaf-wise tree growth*" enquanto o GBM tradicional utiliza "*traditional level-wise tree growth*". Isto quer dizer que à medida que a árvore cresce em profundidade, esta foca-se em crescer apenas um ramo em vez de crescer múltiplos ramos. Esta diferença torna o LightGBM num algoritmo com um desempenho muito superior ao GBM, especialmente quando se trata de trabalhar com grandes números de dados [22].

3.1.3 Deteção de anomalias

Uma anomalia é uma observação que difere tanto das outras que levanta suspeitas sobre o mecanismo que a gerou ser diferente [29]. Como podemos observar na Figura 21, as anomalias, representadas pelo ponto A2 e pelo grupo A1 aparecem isoladas dos restantes objetos, cujo comportamento segue o esperado, pertencendo aos grupos classe 1 e classe 2.

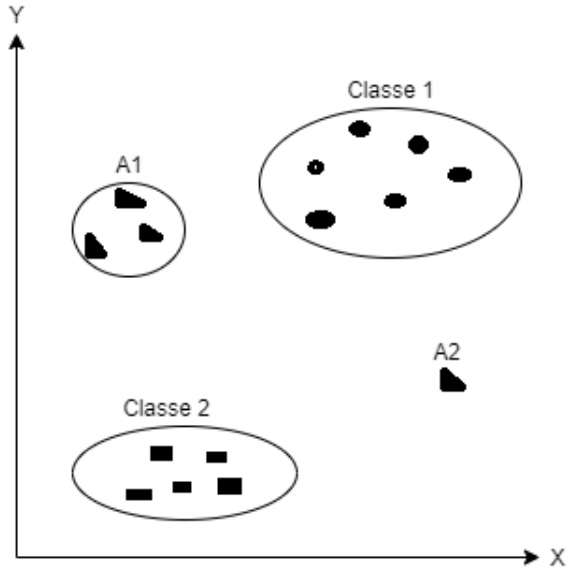

Figura 21 – Exemplos de anomalias, onde Classe 1 e Classe 2 são grupos gerados por observações normais e A1 e A2 são anomalias (adaptado de [29]).

A deteção de anomalias está relacionada com a remoção de ruídos e com a deteção de casos raros [29] [30]. O objetivo da redução de ruídos consiste em remover dados que representam fenômenos não interessantes ao usuário. Já na deteção de casos raros [29] [31], tem como objetivo reconhecer objetos que representam um novo padrão nos dados que se encontram normalmente escondido devido à discrepância do número de ocorrências comparado com as ocorrências normais. Como por exemplo, um cliente realizar pagamentos de faturas em atraso, onde maioritariamente os clientes pagam atempadamente ou a deteção de fraudes financeiras, onde maioritariamente se realizam transações autorizadas.

Devido a este não balanceamento de amostras, muitos métodos comuns de classificação não funcionam bem para a deteção de anomalias devido às suas limitações em enfrentar este problema, através do foco no problema especifico, a maioria das técnicas utilizam várias informações, como a natureza dos dados, o tipo das anomalias a serem detetadas e a disponibilidade de dados previamente rotulados [29].

De seguida é brevemente apresentado técnicas para lidar com o problema do não balanceamento de dados, a técnica de pesos (weighting) e a técnica de amostragem (sampling).

3.1.3.1 Técnica de Pesos

A aprendizagem sensível a custos é um tipo de aprendizagem que leva em consideração os custos/pesos da classificação incorreta (e possivelmente outros tipos

de custos). O objetivo deste tipo de aprendizagem é minimizar o custo total do modelo criado. Ou seja, na aprendizagem sensível ao custo, o custo de classificar um exemplo positivo como negativo pode ser diferente de classificar um exemplo negativo como positivo. Ao contrário da aprendizagem insensível ao custo que não leva em consideração os custos da classificação incorreta [32].

A Figura 22, apresenta para modelo criado utilizando o algoritmo SVM a diferença da sua previsão para uma abordagem não sensível a custos (linha continua) e para uma abordagem sensível a custos (linha a tracejado). Com podes verificar, através da abordagem sensível a custo, são identificados mais registos pertencendo às anomalias.

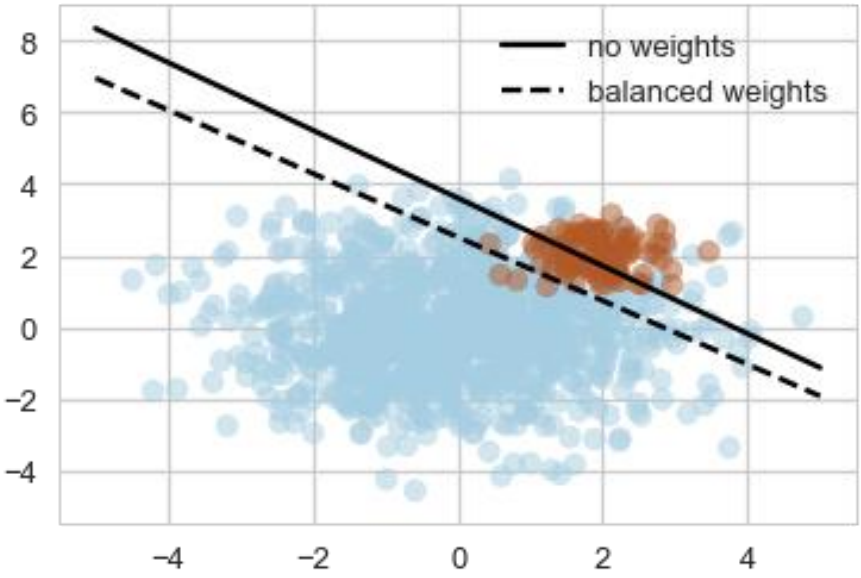

Figura 22 – Representação da diferença de classificação de uma abordagem sem pesos (linha continua) e uma abordagem com pesos (linha tracejado) [33]

3.1.3.2 Técnica de Amostragem

A técnica de amostragem é uma técnica muito comum para lidar com conjuntos de dados não balanceados, pois vai transformar o conjunto não balanceado num com as proporções das classes balanceadas. Existem duas principais abordagens, a técnica de Subamostragem (Undersampling) e a técnica de Sobre amostragem (Oversampling), sendo também muitas vezes utilizado uma combinação das duas [34].

- Subamostragem (UnderSampling)

Esta técnica consiste em extrair um menor conjunto de dados classificados com a classe maioritária, de forma a balancear a quantidade de dados pertencentes a cada classe.

Contudo esta técnica pode provocar a perda de dados importantes ao reduzir o tamanho do conjunto de dados da classe maioritária, chamado de Underfitting. Esta perda de informação pode levar a um modelo não conseguir modelar os dados de treino, de forma a não conseguir a generalização para os novos dados e consequentemente obter resultado pouco precisos [34].

- Sobre amostragem (OverSampling)

Esta técnica, ao contrário da subamostragem, aumenta o número de registos do conjunto de dados da classe minoritária. Isto resulta num conjunto de dados maior, o que terá um impacto no desempenho computacional. A utilização desta técnica pode levar a Overfitting, que ocorre quando um modelo aprende os detalhes de ambas as classes nos dados de treino na medida em que afeta negativamente o desempenho do modelo em novos dados.

Devido a este problema de Overfitting, foi desenvolvida a técnica de Synthetic Minority Oversampling Technique (SMOTE) [35]. Este método gera dados sintéticos com base nas semelhanças do espaço dos atributos entre instâncias minoritárias existentes. Para criar os dados sintéticos, encontra o os pontos mais próximos de cada instância minoritária, seleciona um deles ao acaso, e calcula as interpolações lineares para produzir uma nova instância minoritária [34].

3.1.4 Métricas de avaliação de Modelos

Existem diversas medidas que podem ser utilizadas para avaliar os resultados dos modelos criados pelos algoritmos de aprendizagem automática. Dentre estas, as mais utilizadas fazem uso dos dados presentes na matriz de confusão. A matriz de confusão é composta pelo mesmo número de linhas e colunas, ambos iguais ao número de classes a serem detetadas pelo algoritmo. Cada coluna da matriz representa os casos distribuídos nas classes pelo algoritmo e cada linha representa os casos nas classes das quais eles realmente pertencem [36].

No caso de previsão binaria, tem-se somente duas classes: 0 e 1. Quando se utiliza o algoritmo obtêm-se quatro resultados:

- **Verdadeiros Positivos** (VP), demostra quantos casos foram identificados como a classe que queremos prever, “1”, corretamente.

- **Falsos Positivos** (FP), demonstra quantos casos foram identificados como “1”, mas na verdade são “0”.

- **Verdadeiros Negativos** (VN), demonstra o número de casos “0” que também foram classificados como tal.

- **Falsos Negativos** (FN), demonstra o número de casos "1" que foram classificadas erradamente, com "0".

	Classe Predita	
Classe Verdadeira	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 23 – Matriz de Confusão [36]

Através dos valores presentes na matriz de confusão é possível calcular várias medidas de desempenho utilizadas para a avaliação dos algoritmos de classificação. As métricas utilizadas nesta dissertação são:

- **Precisão geral ou Accuracy** é a taxa que considera apenas os exemplos verdadeiramente positivos para mostrar qual a percentagem de acerto do algoritmo;

$$Accuracy = \frac{VP + VN}{Total}$$

- **Revocação ou Recall** demonstra o número de exemplos classificados como pertencentes a uma classe, que realmente são daquela classe, dividido pela quantidade total de exemplos que pertencem a esta classe, mesmo que sejam classificados como outra. No caso binário, verdadeiros positivos divididos pelo total de positivos [36].

$$Recall = \frac{VP}{VP + FN}$$

3.1.5 *Cold-Start Problem*

O *Cold-start problem*, representa o problema de encontrar padrões em clientes novos, sobre os quais não se contêm informação histórica.

Segundo o estudo "*Cold Start Solutions For Recommendation Systems*" [37], os autores estudam o problema, para novos utilizadores e para novos produtos,

avaliando e comparando as técnicas: Aprendizagem Ativa (Active Learning), Atributos Semânticos (*Semantic Attributes*), Traços de Personalidade (*Personality Traits*) e Cruzamento de Domínios (*Cross-Domain*).

Apresentamos agora uma breve descrição das técnicas aprendizagem ativa, traços de personalidade e cruzamento de domínios.

- Aprendizagem Ativa

Aprendizagem ativa, geralmente faz parte de um tópico de pesquisa mais amplo da aprendizagem automática, uma área de pesquisa bem conhecida que se concentra no design e desenvolvimento de novos algoritmos para realizarem tarefas de regressão e classificação [37]. Estes algoritmos, normalmente, necessitam de um conjunto de dados grande para conseguirem aprender os padrões escondidos nestes, e contruir modelos para prever dados sem precedentes.

- Atributos Semânticos

As soluções tradicionais para o *cold-start problem* são baseadas na abordagem popular Filtragem Baseada em Conteúdo, *Content Based Filtering* (CBF) [37]. Estas abordagens criam perfis do utilizador associando as suas preferências aos atributos semânticos do conteúdo do produto. Através da exploração do conteúdo dos produtos é possível solucionar o problema do novo produto.

- Traços de Personalidade

Uma das soluções naturais para enfrentar o cold-start problem é usar um conjunto adicional de atributos, com o objetivo de se contruir o perfil do novo utilizador [37]. Uma das formas mais representativas de atributos, são os relacionados com a personalidade dos utilizadores. Estes traços de personalidade são baseados em características previsíveis e estáveis dos utilizadores e descrevem “o padrão de comportamento consistente e processos interpessoais originados nos indivíduos” [37].

- Cruzamento de Domínios

Cruzamento de domínios é outra solução para este problema e é baseada na exploração de domínios auxiliares, com o objetivo de gerar recomendações para o domínio de destino.

Nesta dissertação é estudado a técnica de aprendizagem ativa para a criação de um modelo de previsão do pagamento em atraso de faturas para clientes novos, os quais apenas contemos um conjunto de características que os descrevem e os dados da sua primeira fatura.

3.2 Estado de Arte

Para resolver o problema descrito, foi necessário fazer uma pesquisa de soluções existentes que partilham os mesmos problemas e os mesmos objetivos que o nosso. Este estudo permite que seja feita uma análise dos vários designs das soluções encontradas, influenciado a escolha do design da solução deste projeto.

Na Figura 24 está apresentado o processo de transformação do pedido de um cliente em dinheiro para uma empresa. O problema, como descrito anteriormente, visa identificar os pagamentos que serão pagos em atraso pelos respetivos clientes. As etapas do processo que serão afetadas pelo sistema a criar estão realçadas na figura.

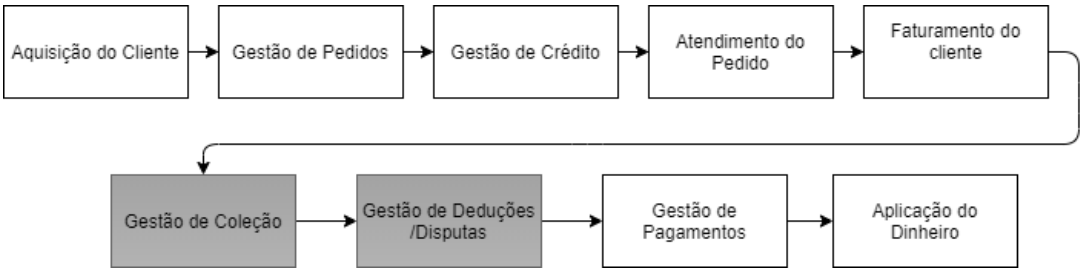

Figura 24 – Processo de Pedido-para-Dinheiro (*Order-to-Cash* (O2C)) [38]

3.2.1 Trabalho Relacionado

Estudo 1

Problema:

A dissertação de Hu Peiguang [38] combina a análise financeira com a análise preditiva. Nesta dissertação, o autor pretende não só prever se o pagamento será realizado em atraso, mas também o intervalo de dias do atraso.

Contudo, para o nosso problema, pretendemos apenas estudar o caso binário de previsão, isto é, se o pagamento da fatura será realizado em atraso.

A análise começa pela análise do balanceamento dos dados. Tendo 72464 faturas na base de dados, que contêm 4291 clientes diferentes e onde 73% das faturas são pagas depois do limite de pagamento, conclui-se que os dados contêm uma distribuição aceitável uma vez que a classe maioria é a classe que é pretendido prever.

Solução:

Após a remoção de erros e ruido dos dados, o próximo passo realizado é a criação de novos atributos. Os atributos históricos para cada cliente, apresentados na Tabela 5, são atributos extraídos sobre os dados das faturas originalmente fornecidas apresentados na Tabela 4.

Tabela 4 – Atributos representantes de um Fatura [38]

Nome	Significado
Número do Cliente	-
Nome do Cliente	-
Número do Documento	-
Referência	Número de referência na base de dados
Centro de lucro	-
Data do Documento	Data da criação da fatura
Moeda do Documento	-
Data limite pagamento fatura	-
Valor fatura	-
Data Pagamento fatura	-
Termo Pagamento	Identificador se o pagamento foi pago em atraso ou não

Tabela 5 – Atributos Históricos para cada cliente [38]

Nº	Atributos	Descrição
1	Número do Cliente	-
2	Número de faturas pagas	Número total de faturas pagas pelo cliente
3	Número de faturas pagas em atraso	Número total de faturas pagas em atraso pelo cliente
4	Rácio do número de faturas pagas em atraso	Rácio entre os pontos 3 e 2
5	Soma do valor base das faturas pagas	Soma dos valores base de todas as faturas pagas do cliente
6	Soma do valor base das faturas pagas em atraso	Soma dos valores base de todas as faturas pagas em atraso de um cliente
7	Rácio da soma do valor	Rácio entre os pontos 6 e 5
8	Média de dias de atraso das faturas pagas em atraso	Média de dias em atraso de todas as faturas pagas em atraso de um cliente
9	Número de faturas pendentes	Número de faturas pendentes de um cliente
10	Número de fatura pendentes em atraso	Número de faturas pendentes em atraso de um cliente.

Nº	Atributos	Descrição
11	Rácio de faturas pendentes em atraso	Rácio entre os pontos 10 e 9
12	Soma do valor base das faturas pagas	Soma dos valores base de todas as faturas pendentes do cliente
13	Soma do valor base das faturas pendentes em atraso	Soma dos valores base de todas as faturas pendentes em atraso de um cliente
14	Rácio da soma do valor	Rácio entre os pontos 13 e 12
15	Média de dias de atraso das faturas pendentes em atraso	Média de dias em atraso de todas as faturas pendentes em atraso de um cliente

Após a extração dos novos atributos, a maioria da análise consiste em testar vários modelos de classificação, de aprendizagem supervisionada. Os algoritmos que são comparados são o classificador de árvore de decisão [12], o classificador Floresta Aleatória [12], Adaptive Boosting [13], Regressão Logística [12], e máquina vetor suporte (SVM) [18]. Obtendo como melhores resultados o classificador Floresta Aleatória, como apresenta a Tabela 6.

Tabela 6 – Previsão de resultados dos algoritmos binários [38]

Modelo	Precisão de previsão fora da amostra
Árvore de decisão (Decision Tree)	0.861
Floresta Aleatória (Random Forest)	**0.892**
AdaBoost	0.863
Regressão Logística (Logistic Regression)	0.864
Máquina de vetores de suporte (SVM)	0.869

Estudo 2

Problema:

O artigo [39] realiza uma análise do processo de cobrança de faturas usando registos de faturas de quatro empresas, acumulando um número total de mais de 170.000 faturas. A principal diferença com o primeiro estudo é a origem dos dados de treino, no qual os dados vieram de apenas um negócio. Ao contrário do presente, onde foram utilizadas quatro fontes de informação diferentes, tornando o problema mais interessante devido ao novo ângulo a ser estudado: a questão de saber se esses conjuntos de dados podem ser interligados para criar um modelo com melhor desempenho.

Solução:

Tal como no primeiro estudo, foi concluído que através da construção de novos atributos representantes do histórico de pagamentos dos clientes, os resultados de previsão melhoram significativamente. Pode-se então concluir, que estes atributos são mais valiosos que os atributos representantes do cliente e da própria fatura, como podemos ver na Tabela 7, onde estão apresentados os resultados de previsão para a empresa "A", utilizando conjunto de atributos diferentes.

Como podemos observar, existe um grande ganho ao utilizar os dados históricos para o caso dos clientes recorrentes e existe também um ganho significativo ao utilizar os dados representativos dos clientes para o caso dos clientes novos.

Tabela 7 – Resultados da previsão para a empresa "A"

Atributos	Todas as Faturas (accuracy)	Faturas dos clientes recorrentes (accuracy)	Faturas dos clientes novos (accuracy)
Dados Faturas	68.38	70.77	70.64
Dados Faturas + Cliente	73.33	74.66	79.48
Dados Faturas + Histórico	85.08	85.08	N/A
Dados Faturas + Cliente + Histórico	86.24	87.56	N/A

É também concluído que através da utilização da aprendizagem sensível a custo foi possível melhor a accuracy das faturas pagas em atraso de risco, isto é, faturas pagas com mais de 90 dias de atraso que estavam sub-representadas nos dados de estudo, lidando assim com o problema de não balanceamento dos dados.

Por fim o autor compara os resultados obtidos pelo modelo de treino com apenas os dados da empresa respetiva e com o conjunto de dados de treino de todas as empresas. É concluído que existe mais a ganhar ao treinar sobre o conjunto de dados criado com os dados de todas as empresas, conforme ilustrado na Tabela 8.

Tabela 8 – Precisão geral da previsão do modelo criado a partir do conjunto de dados de todas as empresas e o modelo criado especificamente para cada empresa [39]

Data para Teste	Accuracy do modelo utilizando dos dados de todas as empresas	Accuracy do modelo utilizando apenas os dados das respetivas empresas
Empresa A	92.65	86.24
Empresa B	95.81	93.09
Empresa C	77.26	66.21
Empresa D	84.10	71.34

Estudo 3

Problema:

O artigo [40] pretende, tal como nos anteriores, prever se uma fatura será paga fora do limite. Enquanto os estudos anteriores estudavam a abordagem multiclasse, este pretende um modelo de classificação binário.

O estudo é baseado em mais de 500 mil faturas e tal como nos outros estudos, os dados encontram-se não balanceados. Em 236124 Faturas presentes na base de dados, cerca de 211812 (90%) são faturas pagas dentro do limite.

Solução:

O autor avalia e compara vários algoritmos de aprendizagem automática, e tal como nos estudos anteriores, é necessário extrair novos atributos representantes do histórico do cliente, e utilizar técnicas para lidar com o não balanceamento de dados. Neste estudo foi utilizado a abordagem sensível a custo.

Na Tabela 9 é apresentado os algoritmos de aprendizagem automática utilizados e os respetivos resultados obtidos segundo as métricas escolhidas pelo autor.

Tabela 9 – Previsão do pagamento atrasado de faturas dos classificadores avaliados após a adição dos novos atributos [40].

Método	F1-score	Recall	Accuracy
SVM	97.95	77.80	89.31
Regressão Logística	97.56	68.80	91.50
GBM	96.74	51.96	96.18
Floresta Aleatória	96.73	53.46	93.34
Rede Neural	96.15	60.23	78.20

Método	F1-score	Recall	Accuracy
Árvore de decisão Balanceada (ADB)	96.24	59.19	77.55
Adaboost Balanceada (AB)	96.40	70.95	72.93
Regressão Logística Balanceada (RLB)	**97.09**	**84.83**	**74.15**
GBM Balanceada (GBMB)	**98.01**	**81.86**	**86.77**

3.2.2 Análise comparativa do trabalho relacionado

Na secção anterior, foi apresentado três exemplos reais de estudos sobre a criação de modelos de previsão sobre o pagamento atrasado de faturas de clientes.

Através da análise destes estudos foi possível identificar que todos dividiram a solução em 4 fases principais:

- Limpeza e pré-processamento de dados;
- Análise estatística e seleção de atributos;
- Construção de modelos de aprendizagem supervisionada;
- Teste e avaliação do desempenho dos modelos de classificação.

Foi também concluído que é essencial extrair atributos históricos sobre o comportamento dos clientes em relação ao pagamento de faturas e extrair atributos que descrevam o perfil do cliente, para que seja possível criar um modelo de sucesso para a previsão do pagamento atrasado de faturas. Os atributos históricos extraídos que foram comuns para todos os estudos, estão apresentados na Tabela 10.

Tabela 10 – Atributos Históricos sobre cada Cliente

Nº	Atributos	Descrição
1	Número de faturas pagas	Número total de faturas pagas pelo cliente
2	Número de faturas pagas em atraso	Número total de faturas pagas em atraso pelo cliente
3	Rácio do número de faturas pagas em atraso	Rácio entre os pontos 2 e 1
4	Soma do valor base das faturas pagas	Soma dos valores base de todas as faturas pagas do cliente

Nº	Atributos	Descrição
5	Soma do valor base das faturas pagas em atraso	Soma dos valores base de todas as faturas pagas em atraso de um cliente
6	Rácio da soma do valor	Rácio entre os pontos 5 e 4
7	Média de dias de atraso das faturas pagas em atraso	Média de dias em atraso de todas as faturas pagas em atraso de um cliente
8	Número de faturas pendentes	Número de faturas pendentes de um cliente
9	Número de fatura pendentes em atraso	Número de faturas pendentes em atraso de um cliente.
10	Rácio de faturas pendentes em atraso	Rácio entre os pontos 9 e 8
11	Soma do valor base das faturas pagas	Soma dos valores base de todas as faturas pendentes do cliente
12	Soma do valor base das faturas pendentes em atraso	Soma dos valores base de todas as faturas pendentes em atraso de um cliente
13	Rácio da soma do valor	Rácio entre os pontos 12 e 11
14	Média de dias de atraso das faturas pendentes em atraso	Média de dias em atraso de todas as faturas pendentes em atraso de um cliente

Também foi notado um problema comum de não balanceamento de dados, uma vez que se trata de pagamentos de faturas e a maior parte dos clientes realizam o pagamento antes da data limite. Devido a este problema será necessário utilizar técnicas de balanceamento como as técnicas de **sobre amostragem**, **subamostragem**, **SMOTE** e utilizar algoritmos capazes de integrar **pesos** e algoritmos de **aprendizagem sensível a custos**. O algoritmo mais utilizado, que suporte pesos, nestes estudos foi o classificador Floresta Aleatória. A Tabela 11 apresenta os algoritmos comuns utilizados nas soluções dos estudos acima apresentados.

Tabela 11 – Classificadores de aprendizagem Supervisionada

Classificador
Máquina de vetores de suporte (SVM)
Regressão Logística
GBM
Floresta Aleatória
Floresta Aleatória (pesos)
Floresta Aleatória (sensível a custos)
Rede Neural
Árvore de decisão

3.2.3 Tecnologias Utilizadas

A presente secção apresenta as tecnologias utilizadas na construção da solução desta dissertação, nomeadamente a linguagem de programação Python [41], a ferramenta

de controlo de versões Git [42] e a ferramenta Power BI [1] usada com o único propósito para visualização dos resultados de previsão.

3.2.3.1 Pyhton

A linguagem de programação Python foi desenvolvida por Guido Van Rossum. O Python é uma linguagem modular, multiplataforma e de fácil aprendizagem [41].

O Python permite a implementação de programas e a execução de protótipos, tornando o processo de desenvolvimento muito mais rápido.

Na Figura 25 é apresentado a previsão do número de perguntas por ano no website Stack Overflow das linguagens de programação [41].

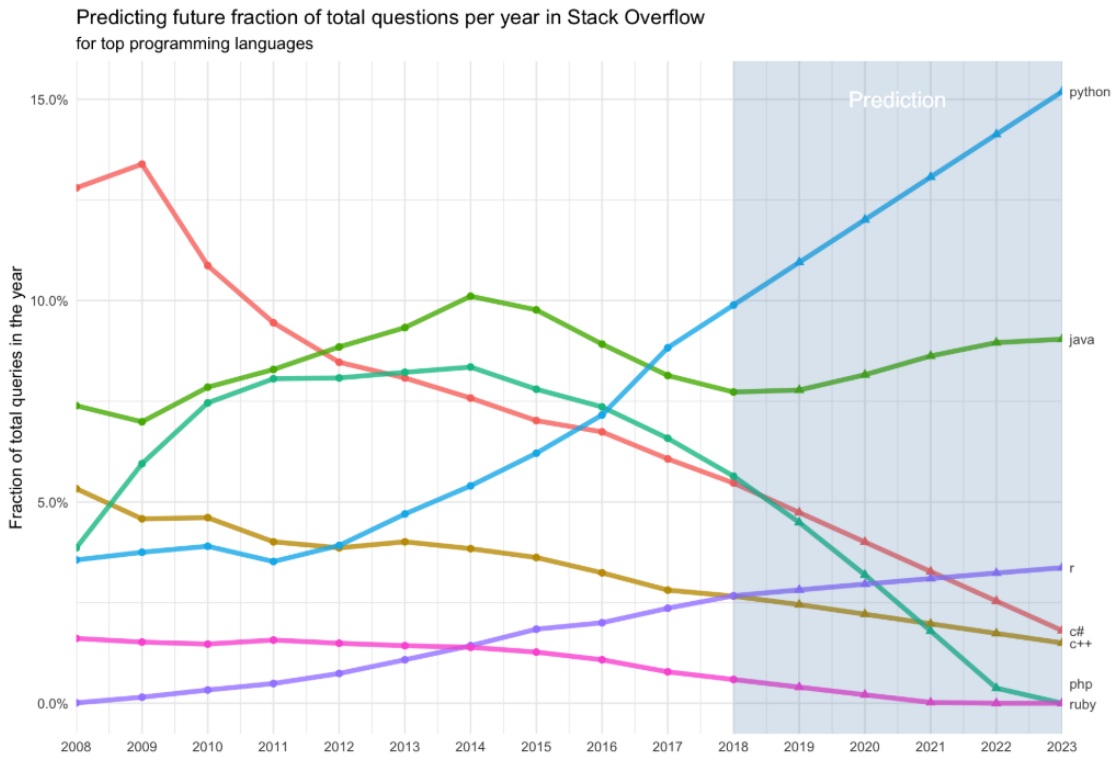

Figura 25 – Previsão de Popularidade de Linguagens de Programação [41]

3.2.3.2 Git

O Git é uma ferramenta de controlo de versão open-source que permite que várias pessoas possam trabalhar no mesmo projeto ao mesmo tempo com segurança. São guardadas todas as alterações feitas nos arquivos ao longo do projeto, permitindo revertê-los a qualquer momento para um estado anterior [42].

3.2.3.3 Power BI

O Power BI é uma solução de análise de negócios que permite a visualização, análise e a partilha de dados e informações de uma organização ou incorporação na sua aplicação ou site, através do fornecimento de interfaces intuitivas como dashboards e relatórios dinâmicos [1].

4 Descrição Técnica

O presente capítulo expõe e contextualiza, inicialmente, uma análise ao caso de estudo. Em seguida, é apresentada a análise arquitetural e do design arquitetural, onde são expostos o design do sistema em completo e a apresentação do design do treino do modelo de previsão. Por fim é apresentado o design da componente de PowerBI.

4.1 Engenharia de Requisitos

Na engenharia de requisitos, é apresentada a análise de requisitos funcionais e não funcionais que surgiram no projeto. Os requisitos são classificados de acordo com cada uma das diferentes categorias do modelo FURPS+ [43].

4.1.1 Requisitos Funcionais

Os requisitos funcionais definem as principais características do sistema. Os nossos requisitos funcionais são:

- **Tratamento de dados**: O sistema deve ser capaz de tratar os dados recebidos, de forma prepará-los para modelação.

- **Criação e treino dos modelos de previsão**: O sistema deve ser capaz de treinar e avaliar modelos de previsão, fornecendo assim o melhor modelo possível.

- **Análise de dados em PowerBI**: O sistema deve fornecer uma análise aos dados fornecidos e aos resultados de previsão, através de um relatório dinâmico automaticamente atualizado sempre que novos dados são inseridos.

4.1.2 Requisitos Não Funcionais

Na classificação FURPS+, existem requisitos não funcionais, incluindo usabilidade, confiabilidade, desempenho.

- **Usabilidade**: Interface simples, intuitiva e eficiente;
- **Fiabilidade/Segurança**: O sistema deve estar acessível vinte e quatro horas por dia, sete dias por semana;
- **Desempenho**: O sistema deverá suportar múltiplos acessos concorrentes sem que coloque em causa a produtividade dos utilizadores.

4.2 Caso de Estudo

Os dados com que lidamos nesta dissertação, são dados de faturas. O fornecedor desta informação é a empresa GoldEnergy, cliente da CPCIT4all. Contudo é uma empresa com uma grande quantidade de dados, fornecendo mais de 9 milhões de registos pertencendo a mais de 425 mil clientes diferentes. Cada registo de faturas é caracterizado pelos atributos apresentados na Tabela 12.

Tabela 12 – Atributos dos dados fornecidos pela GoldEnergy

Atributo	Significado
Contrato	Id de contrato associado a um cliente.
Nome Cliente	Nome de cliente
Distrito	Distrito associado do local consumo
Tipo Documento	Tipo documento na conta corrente do cliente sempre “Fatura” ou “NotaCrédito”
ID_Documento	Nº Documento da Fatura ou Nota de crédito
Valor Documento	Valor com IVA incluído da fatura ou nota Crédito
Data Registo	Data em que o documento é emitido no sistema
Data Vencimento	Data limite para pagamento por parte do cliente, apenas é relevante para o tipo documento Fatura.
Método Pagamento	Forma pela qual o cliente vai pagar o documento. É sempre “Normal” ou “DébitoDireto”
Pago Reembolsado	Se o documento foi liquidado na totalidade.
TipoCliente	Se é cliente “Industrial” ou “Doméstico”
TarifaSocial	O cliente com tarifa social tem um desconto adicional no documento emitido.
Campanha Em Vigor	Nome da Campanha em vigor no período de emissão do documento
EnergiaDocumento	3 tipos possíveis: “Gás”, “Eletricidade” ou “Gás+Eletricidade”

Atributo	Significado
Rescindiu	A partir da emissão do documento o cliente tinha o Contrato Rescindido.
Data Pagamento Ou Reembolso	Data em que o cliente liquidou a fatura ou que foi reembolsado se tratou de uma NotaCrédito
Data Emissão Corte	Data em que foi emitido um aviso de corte. Um aviso de corte representa um documento que o cliente recebe com uma notificação de pagamento urgente, caso os documentos incluídos no corte não sejam totalmente liquidados. O corte é efetuado para as energias ativas.
ID_AvisoCorte	Nº de identificação do aviso de corte. Um aviso de corte pode conter várias faturas que o cliente tem por pagar.
Data Vencimento Corte	Data limite de pagamento do aviso de corte.
Pago Pelo Aviso Corte	"Sim" caso o cliente tenha pago pela referência multibanco associada ao documento do corte. O cliente pode ter pago os documentos incluídos no corte por outros meios, nesse caso seria exibido "Não" nesta coluna.
Data Pagamento Corte	Se o cliente pagou porque recebeu o aviso corte é exibida a data na qual o fez.
Tipo Adesão Gás	Se a fatura do cliente contém serviços de gás e se é a primeira vez que os requisita. Contêm os valores de "Primeira Instalação", "Mudança Comercializador" e "Não".
Tipo Adesão Energia	Se a fatura do cliente contém serviços de energia elétrica e se é a primeira vez que os requisita. Contêm os valores de "Primeira Instalação", "Mudança Comercializador" e "Não".
Escalão Gás	Escalão do preço do gás, contem valores de 0 a 5, sendo 0 não pagar por gás e sendo 5 o escalão mais caro.
Potencia Energia	Potência de energia elétrica que é fornecida. Contem 13 valores diferentes.

Após a análise da distribuição de faturas pagas em atraso, apresentado na Figura 26, concluímos que os dados se encontram não balanceados, contendo 84% dos registos com faturas pagas dentro do limite do tempo contra apenas 16% dos registos com faturas pagas depois do limite. Sendo o nosso objetivo a previsão da classe minoritária, será necessário utilizar técnicas de balanceamento de dados e de deteção de anomalias.

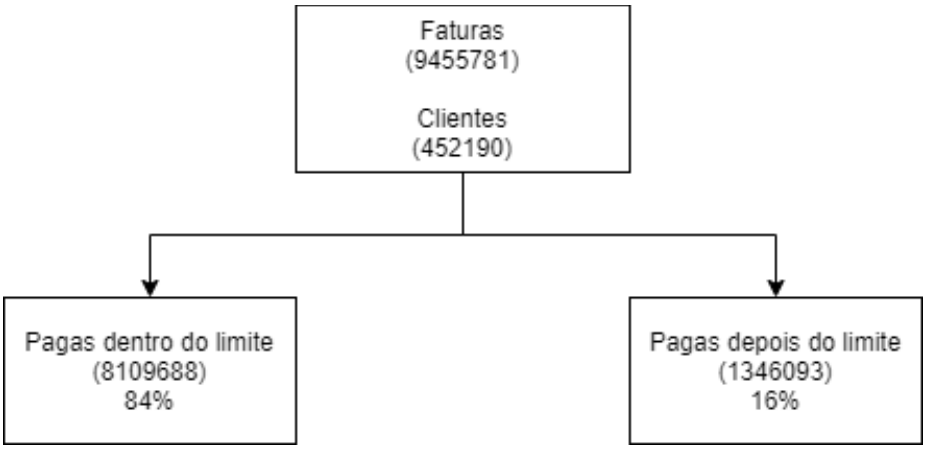

Figura 26 – Distribuição dos dados fornecidos para o caso de estudo

4.3 Análise e Design Arquitetural

Na criação do sistema pretendido, foi necessário realizar um design arquitetural que permitisse a automação do processo de modelação dos dados provenientes dos clientes. Fornecendo um modelo personalizado para cada cliente, obtendo assim melhores resultados de previsão, uma vez que cada modelo de aprendizagem automática será específico para cada cliente.

Foi identificado a existência de 3 componentes principais:

- Componente de base de dados, onde estão guardados os dados a serem analisados.
- A componente do modelo de previsão, a componente mais importante do sistema, onde é criado o melhor modelo previsão possível para os dados fornecidos.
- E por fim a componente de PowerBI, onde são apresentados os resultados de previsão e uma análise estatística ao cliente.

De seguida é apresentado o design arquitetural do sistema, sendo apresentado o diagrama de componentes, o diagrama de fluxo do sistema, o diagrama de implantação, o modelo de dados e o diagrama de domínio. É também apresentado o funcionamento e design da componente de criação dos modelos de previsão, componente principal do sistema.

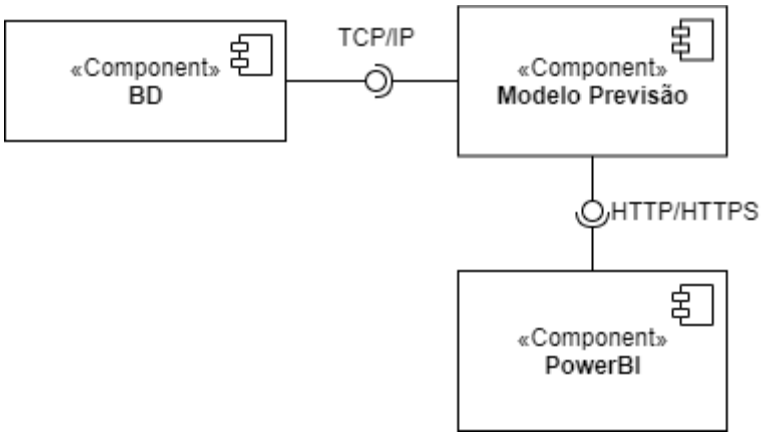

Figura 27 – Diagrama de Componentes

Na Figura 27 é apresentado o diagrama de componentes, onde a componente "Modelo Previsão" consome os dados provenientes da componente "BD" através de uma ligação tcp/ip. O resultado da componente "Modelo Previsão" é depois consumido pela componente "PowerBI" através de pedidos HTTPS.

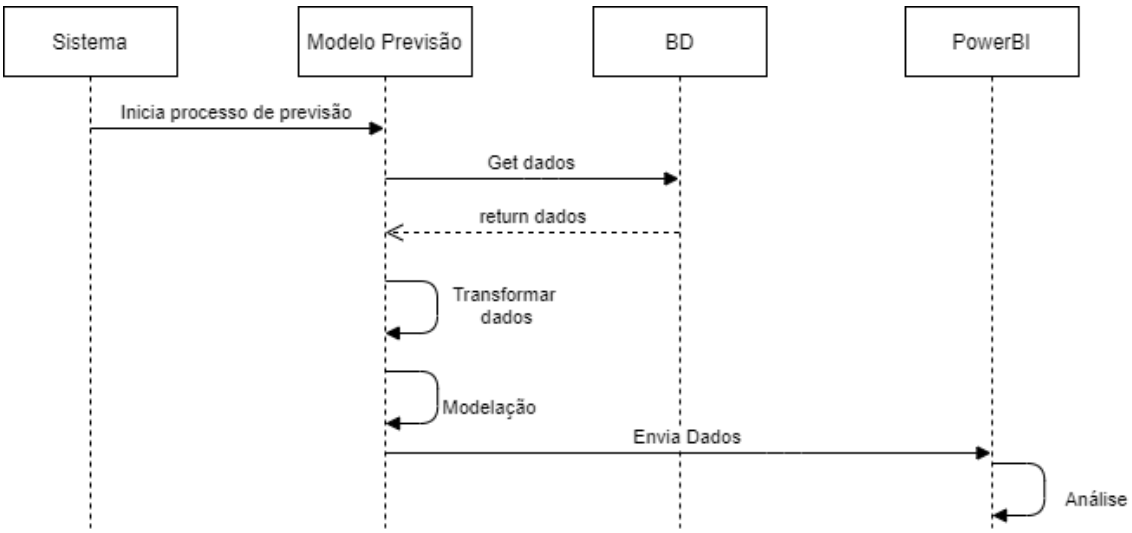

Figura 28 – Diagrama de Fluxo do Sistema

Na Figura 28 é apresentado o diagrama de fluxo do sistema. O sistema inicia o processo de previsão, onde a componente "Modelo Previsão" faz o pedido de GET dos dados à Base de dados, que os devolve à componente inicial.

Aqui os dados sofrem transformações para ficarem prontos para o processo de modelação, onde os resultados obtidos pelo modelo são enviados para o PowerBI, atualizando os relatórios dinâmicos e consequentemente as análises estatísticas fornecidas ao utilizador.

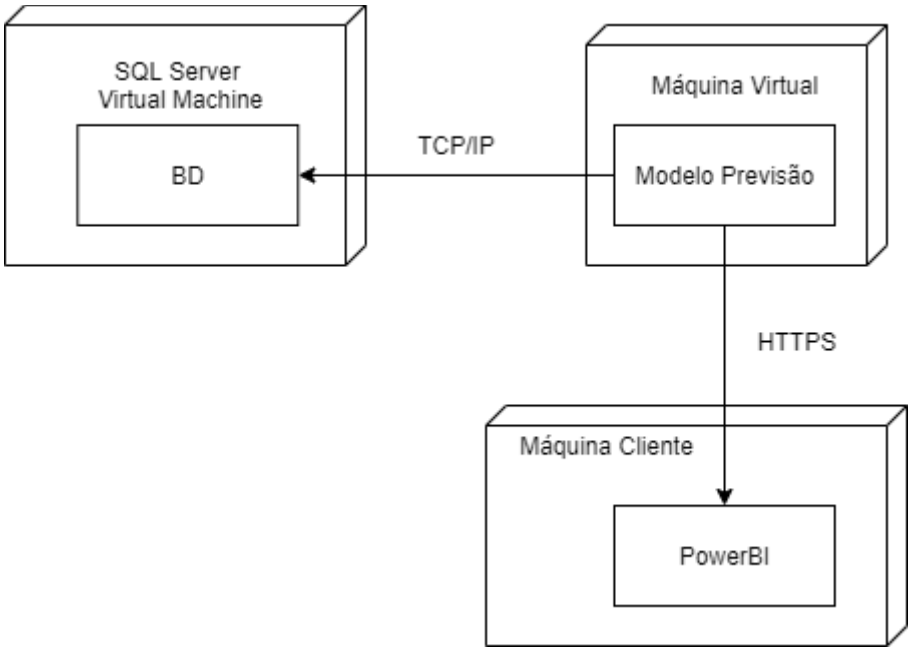

Figura 29 – Vista de Implantação

Na Figura 29 é apresentado o diagrama de Implantação do sistema, onde a componente “BD” está incorporada numa máquina virtual e pode ser acedida através de SQL Server. A componente de previsão, “Modelo Previsão”, está incorporada numa outra máquina virtual, sempre disponível para realizar uma previsão e por fim a componente de “PowerBI” está disponível na máquina do cliente.

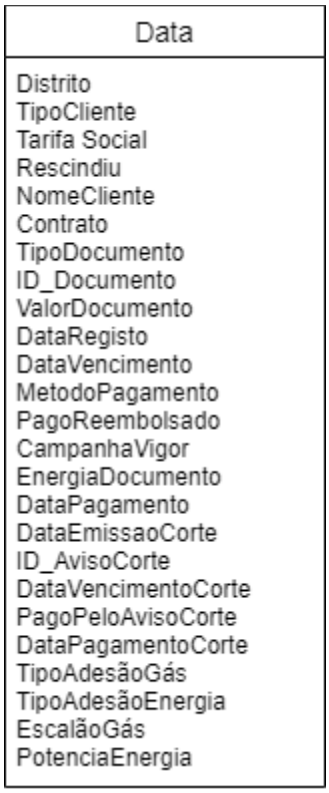

Figura 30 – Modelo Dados

Na Figura 30 está apresentado o modelo de dados do sistema, para o caso de estudo, onde a base de dados é composta por apenas uma tabela representativa dos dados dos clientes, contratos e faturas.

Tabela 13 – Atributos e o seu tipo

Nº	Atributo	Tipo
1	Distrito	Discreto
2	TipoCliente	Discreto
3	TarifaSocial	Discreto
4	Rescindiu	Discreto
5	NomeCliente	Discreto
6	Contrato	Discreto
7	TipoDocumento	Discreto
8	ID_Documento	Discreto
9	ValorDocumento	Continuo
10	DataRegisto	Date
11	DataVencimento	Date
12	MetodoPagamento	Discreto
13	PagoReembolsado	Discreto
14	CampanhaVigor	Discreto
15	EnergiaDocumento	Discreto
16	DataPagamento	Date
17	DataEmissaoCorte	Date

18	ID_AvisoCorte	Discreto
19	DataVencimentoCorte	Date
20	PagoPeloAvisoCorte	Discreto
21	DataPagamentoCorte	Date
22	TipoAdesãoGás	Discreto
23	TipoAdesãoEnergia	Discreto
24	EscalãoGás	Discreto
25	PotenciaEnergia	Discreto

Na Tabela 13 são apresentados os atributos e o seu tipo, sendo todos os atributos do tipo discreto, à parte do atributo "ValorDocumento" que pertence ao tipo contínuo.

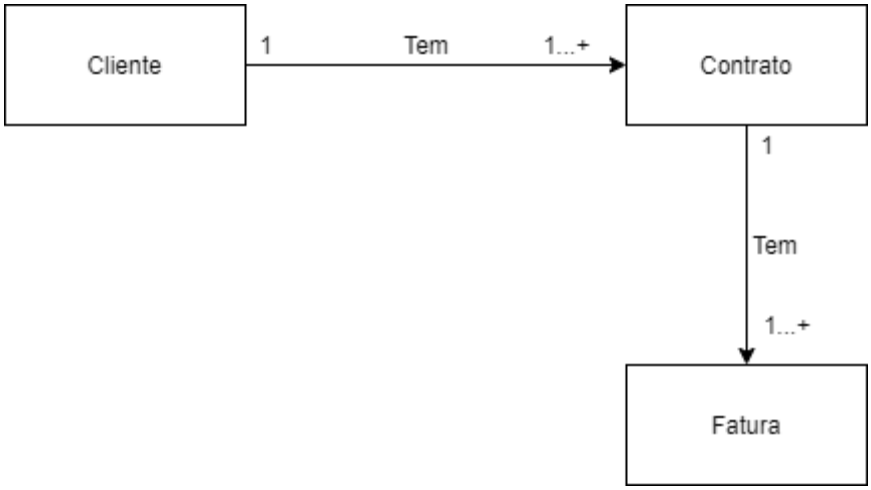

Figura 31 – Modelo de Domínio

Na Figura 31 é apresentado o modelo de domínio do sistema, sendo composto por dados representativos do cliente dos contratos e das faturas. Cada cliente tem um ou mais contratos, e cada contrato tem uma ou mais faturas.

As seguintes tabelas de atributos são especificas para o nosso caso de estudo. Na Tabela 14 é apresentado os atributos característicos de um cliente.

Tabela 14 – Atributos característicos do Cliente

Atributos Característicos do Cliente				
Distrito	TipoCliente	Tarifa Social	Rescindiu	Nome Cliente

Na Tabela 15 é apresentado os atributos característicos de uma fatura, onde esta pertence a um contrato.

Tabela 15 – Atributos característicos de uma fatura

Atributos Característicos da Fatura				
Contrato	Tipo Documento	ID_Documento	Valor Documento	Data Registo
Data Vencimento	Método Pagamento	Pago Reembolsado	Campanha Em Vigor	EnergiaDocumento
Data Pagamento Ou Reembolso	Data Emissão Corte	ID_AvisoCorte	Data Vencimento Corte	Pago Pelo Aviso Corte
Data Pagamento Corte	Tipo Adesão Gás	Tipo Adesão Energia	Escalão Gás	Potencia Energia

4.3.1 Design Treino Modelo Preditivo

O sistema cria um modelo específico para cada cliente, mas o seu processo de treino é idêntico para todos. Na presente secção é apresentado o processo de treino de um modelo de previsão.

Como referido no contexto vamos utilizar o processo CRISP-DM, onde a solução é composta por 4 processos, ilustrados de seguida na Figura 32:

- O processo de limpeza e pré-processamento dos dados é representado pela etapa azul da figura. Este processo representa mais de 80% do trabalho realizado, uma vez que é necessário analisar e corrigir quais quer incoerências e erros que possam existir nos dados.

- O processo de construção e seleção de atributos é representada pela etapa laranja da figura. Este processo é responsável pela extração de novos conjuntos de dados representativos de nova informação sobre o perfil dos clientes, como os atributos históricos, apresentados na Tabela 17 e o conjunto de atributos Temporais, apresentados na Tabela 18.

- O processo de modelação é representado pelas etapas verde e vermelha na figura. Este processo é responsável pela criação de modelos de previsão, através da combinação de algoritmos de machine learning e as técnicas de lidar com o não balanceamento dos dados. Os algoritmos utilizados serão os apresentados na Tabela 16;

- O processo de teste e avaliação do desempenho dos modelos de classificação criados, representado pela etapa a roxo na figura. Neste, os modelos criados no processo anterior são avaliados e comparados através das métricas recall e accuracy, de forma a obter o melhor possível. Este processo de experiência e teste é repetido para cada conjunto de atributos, técnica de lidar com o problema de casos raros e algoritmos de aprendizagem automática, até que seja considerado que não seja possível melhorar os resultados.

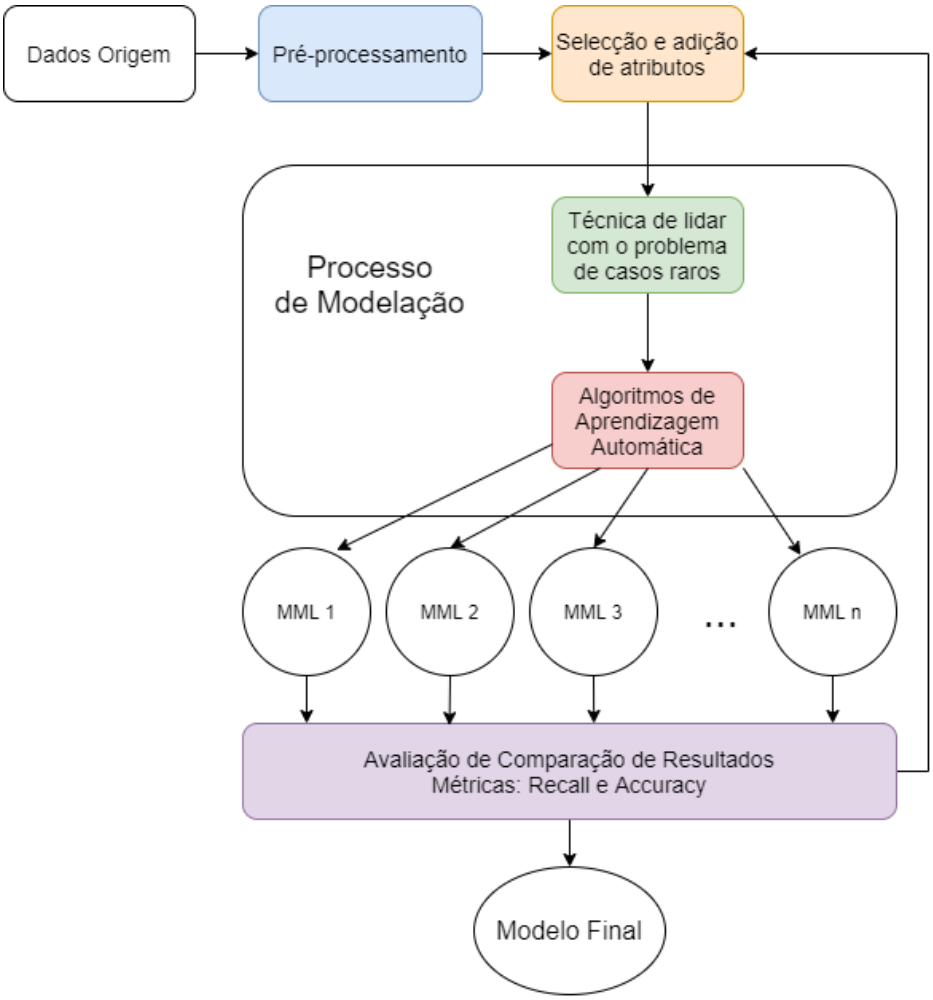

Figura 32 – Design Arquitetural

De seguida são apresentados os algoritmos de aprendizagem automática utilizados e o processo de extração de dados históricos e temporais.

Tabela 16 – Alternativas de design

Algoritmos Aprendizagem Automática
Árvore de Decisão
Floresta Aleatória
Regressão Logística
One-Class SVM

Algoritmos Aprendizagem Automática
Rede Neural
LighGBM

No processo de seleção e construção de atributos, relativamente ao nosso caso de estudo, foram extraídos dados estatísticos representado o histórico de um cliente e foram também extraídos dados temporais.

Da tabela 15 apresentada acima, o valor do documento é generalizado, classificado em 6 categorias diferentes consoante a montante.

Tabela 17 – Atributos Históricos extraídos para cada Cliente

Nº	Atributos	Descrição
1	Número de faturas pagas	Número total de faturas pagas pelo cliente
2	Número de faturas pagas em atraso	Número total de faturas pagas em atraso pelo cliente
3	Rácio do número de faturas pagas em atraso	Rácio entre os pontos 2 e 1
4	Soma do valor base das faturas pagas	Soma dos valores base de todas as faturas pagas do cliente
5	Soma do valor base das faturas pagas em atraso	Soma dos valores base de todas as faturas pagas em atraso de um cliente
6	Rácio da soma do valor	Rácio entre os pontos 5 e 4
7	Média de dias de atraso das faturas pagas em atraso	Média de dias em atraso de todas as faturas pagas em atraso de um cliente
8	Número de faturas pendentes	Número de faturas pendentes de um cliente
9	Número de fatura pendentes em atraso	Número de faturas pendentes em atraso de um cliente.
10	Rácio de faturas pendentes em atraso	Rácio entre os pontos 9 e 8
11	Soma do valor base das faturas pagas	Soma dos valores base de todas as faturas pendentes do cliente
12	Soma do valor base das faturas pendentes em atraso	Soma dos valores base de todas as faturas pendentes em atraso de um cliente
13	Rácio da soma do valor	Rácio entre os pontos 12 e 11
14	Média de dias de atraso das faturas pendentes em atraso	Média de dias em atraso de todas as faturas pendentes em atraso de um cliente

Na tabela 17 está apresentado os atributos extraídos sobre o histórico de um cliente, estes atributos são dados estatísticos sobre as faturas de um cliente.

Tabela 18 – Atributos Temporais

<table>
<tr><th colspan="3">Temporais</th></tr>
<tr><td colspan="2">Para cada Cliente</td><td>Para cada faturas de cada cliente</td></tr>
<tr><td>últimas 5 faturas</td><td>últimas 3 faturas</td><td>Pagou a anterior</td></tr>
<tr><td colspan="2">Número de faturas não pagas dentro do limite</td><td>Número de faturas não pagas seguidas até a atual</td></tr>
<tr><td colspan="2">Valor médio das faturas</td><td></td></tr>
<tr><td colspan="2">Rácio de faturas não pagas dentro do limite</td><td></td></tr>
</table>

Na Tabela 18 está apresentado os atributos temporais extraídos sobre o cada cliente, sobre as suas últimas 3 e 5 faturas e para cada fatura deste cliente, se a anterior foi paga.

4.3.2 Design PowerBI

Os resultados de previsão das novas faturas serão depois fornecidos ao gestor de cobranças de pagamentos (o utilizador), através de um relatório dinâmico, conforme ilustrado na Figura 33.

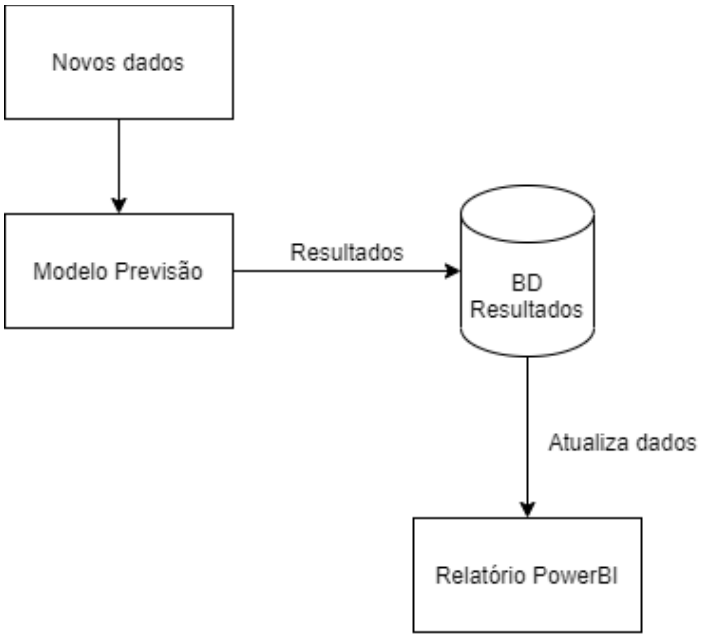

Figura 33 – Processo atualização do Relatório PowerBI

O relatório apresenta uma análise dos dados fornecidos para o estudo, ou seja, fornece uma análise estatística sobre as previsões, mas também sobre os dados como um todo.

Sendo o relatório do PowerBI um relatório dinâmico, todas as métricas estatísticas são filtradas pelos filtros apresentados na Tabela 19, fornecendo assim uma análise com o detalhe que o cliente desejar.

Tabela 19 – Filtros do relatório PowerBI

Filtro	Variação
Nome Cliente	Contem todos os nomes dos clientes
Contrato	Contem todos os números de contrato
Anos/Meses	Contem todos os anos e os seus respetivos meses

De seguida é apresentado as métricas de análise estatística fornecidas através do relatório dinâmico do PowerBI.

Tabela 20 – Métricas de análise estatística

Métrica
Valor total faturas
Valor total fátuas em atraso
Rácio do valor total das faturas em atraso
Número de faturas
Número de faturas em atraso
Percentagem de faturas em atraso

A Figura 34 apresenta uma análise mais geral dos dados completos fornecidos no nosso caso de estudo, fornecendo as métricas de análise estatística apresentadas, anteriormente, na Tabela 20.

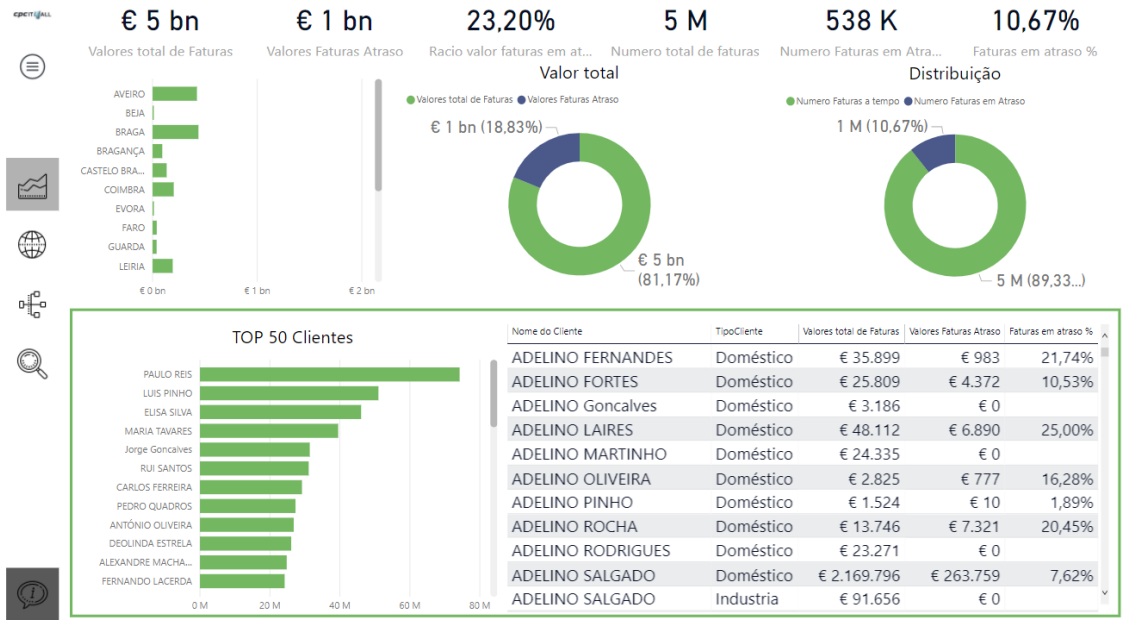

Figura 34 – Página Home do relatório PowerBI

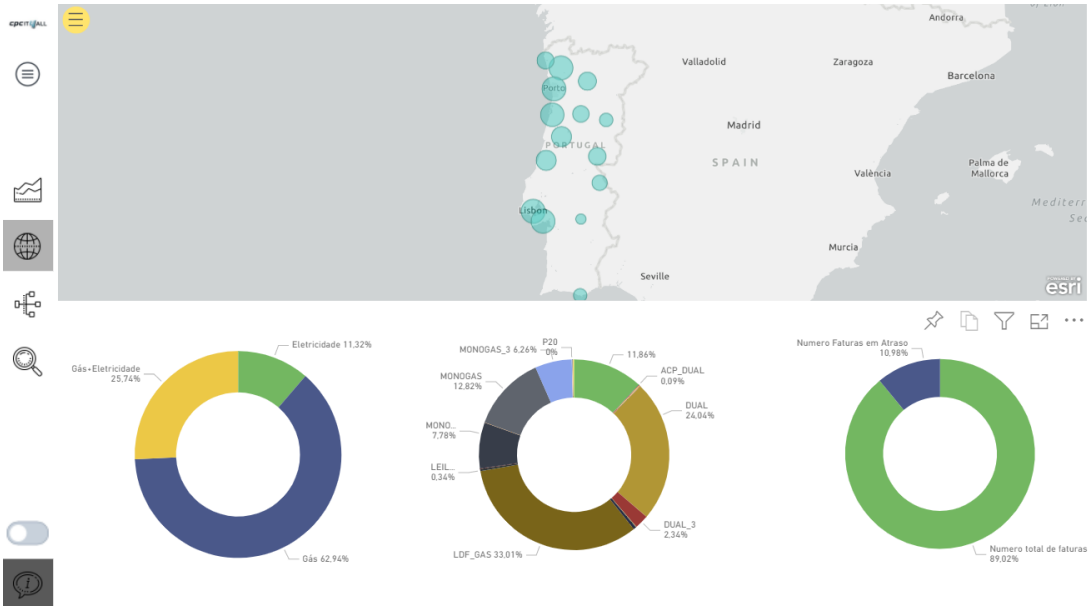

Figura 35 – Página Visualização Geográfica

A Figura 35 apresenta a página de visualização geográfica, apresentando um mapa com os distritos e o seu respetivo volume do negócio, neste caso, valor total das faturas.

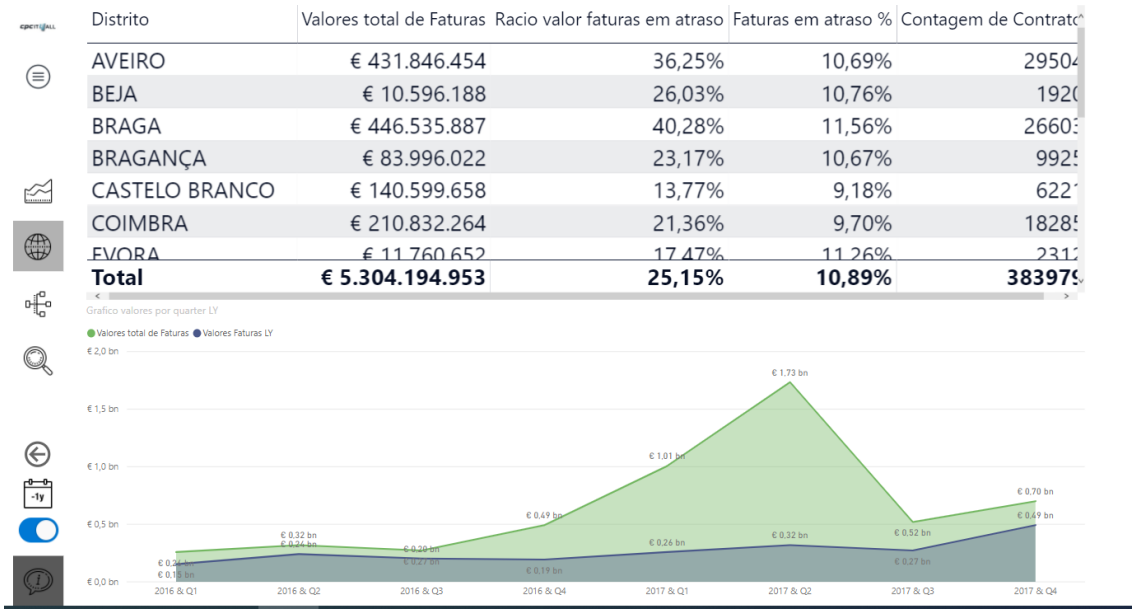

Figura 36 – Visualização Geográfica - comparação anos

A Figura 36 apresenta outra vertente da página de visualização geográfica onde é possível comparar os dados de cobranças ao com o respetivo intervalo de tempo do ano anterior.

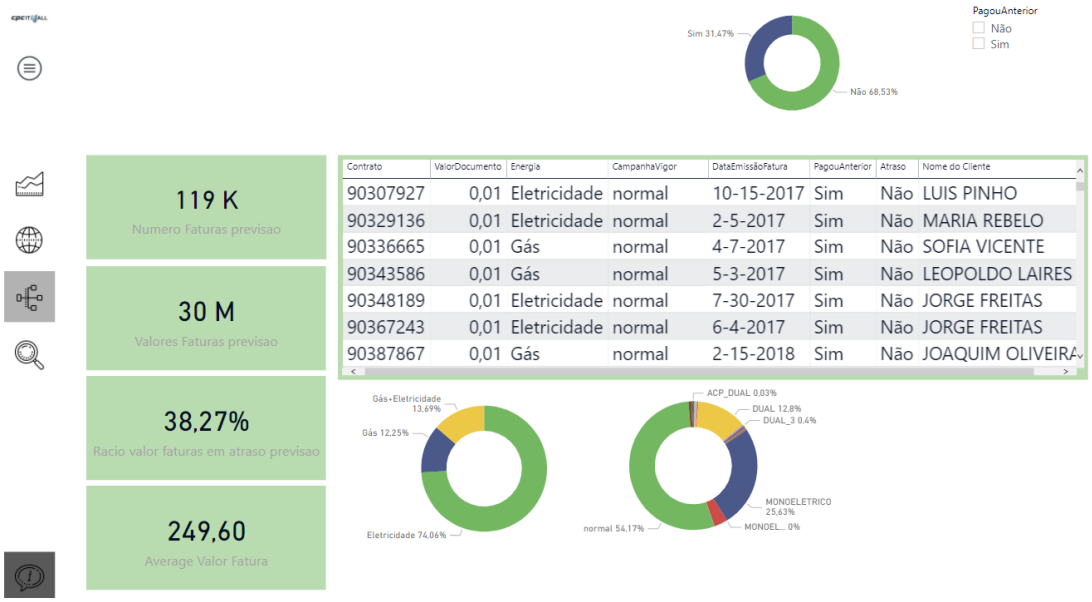

Figura 37 – Página Resultados Previsão

A Figura 37 apresenta os resultados de previsão assim como uma análise estatística sobre estes dados. Apresentando o número de faturas analisadas, o valor total destas faturas, o rácio do valor das faturas classificadas como potencial atraso e o valor médio das faturas.

Figura 38 – Visualização detalhe cliente

A Figura 38 apresenta uma vista em detalhe do cliente, onde são apresentados ao longo do tempo estipulado nos filtros, o valor das faturas pagas dentro e fora do limite, assim como o valor total de cobranças e o valor total de cobranças atrasadas.

5 Experiências

Para a criação do mecanismo de previsão, foi necessário estudar o comportamento de combinações de várias técnicas e algoritmos de machine learning, de forma a obter a combinação que obtenha os melhores resultados, gerando assim o melhor modelo de previsão possível.

De seguida é apresentado a configuração, resultados e análise comparativa das seis experiências mais relevantes para o nosso caso de estudo.

5.1 Configuração

Durante todas as experiências a ser descritas, foi utilizado mesmo conjunto de dados de treino e o mesmo conjunto de dados te teste. Sendo o conjunto de dados de teste composto pela última fatura de cada contrato, é criando um conjunto de teste com mais de 400 mil faturas.

O treino do modelo de aprendizagem automática para clientes com registos anteriores, é utilizado um conjunto de dados composto pelos atributos históricos, apresentados anteriormente na tabela 17, pelos atributos característicos dos clientes, apresentados na tabela 14, pelos dados característicos de uma fatura, apresentados na tabela 15 e pelos atributos temporais, apresentados na tabela 18.

As experiências apresentadas sobre este modelo são as 1,2,3 e 5.

Em relação ao treino do modelo de aprendizagem automática para clientes novos, é utilizado um conjunto de dados composto pelos atributos característicos dos clientes, apresentados na tabela 14 e pelos dados característicos de uma fatura, apresentados na tabela 15.

As experiências apresentadas sobre este modelo são a 4 e 6.

Nas experiências 1 e 4 é usado a técnica de amostragem UnderSampling para lidar com o problema de casos raros.

Na experiência 2 é usado a técnica de pesos para lidar com o problema de casos raros, onde são avaliados 3 conjunto de pesos diferentes usando o algoritmo Random Forest.

Na experiência 3 é usado a técnica de "One Class Classification" para lidar com o problema de casos raros.

Nas experiências 5 e 6 é usado a técnica de amostragem SMOTE para lidar com o problema de casos raros.

Tabela 21 – Resumo configuração experiências

Experiência	Modelo	Dados	Técnica Casos raros	Distribuição dados
1	Clientes com Histórico	Históricos Temporais + Clientes + Fatura	UnderSampling: 1 para 1	0: 1315206 1: 1315206
2	Clientes com Histórico	Históricos Temporais + Clientes + Fatura	Pesos: - 0:9, 1:1 - 0:1, 1:9	0: 7688385 1: 1315206
3	Clientes com Histórico	Históricos Temporais + Clientes + Fatura	One Class Classification	0:0 1: 1315206
4	Clientes Novos	Clientes + Fatura	UnderSampling: 1 para 1	0: 1315206 1: 1315206
5	Clientes com Histórico	Históricos Temporais + Clientes + Fatura	SMOTE: 1 para: 60%, 75% e 85%	0: 7688385 1: 60%, 75% e 85%
6	Clientes Novos	Clientes + Fatura	SMOTE: 1 para: 60%, 75% e 85%	0: 7688385 1: 60%, 75% e 85%

A Tabela 21 apresenta um resumo da configuração das experiências descritas nesta secção, apresentando o modelo a que pertencem, os dados que utilizam, a técnica de lidar com o problema dos casos raros utilizada e a distribuição dos dados das classes 0 – Faturas pagas dentro do tempo limite e 1 – Faturas pagas depois do tempo limite.

5.2 Resultados

De seguida são apresentados os resultados das experiências acima descritas:

Classe 0 – Pago dentro do limite

Classe 1 – Pago em atraso

Experiência 1:

Tabela 22 – Resultados experiência 1

Algoritmo	Resultados	
	Accuracy (%)	Recall (%)
LGBM	80.01	88.56
Random Forest	82.22	85.02
Decision Tree	78.52	81.85
Rede Neural	69.48	90.67
Logistic Regression	77.88	78.12

Experiência 2:

O algoritmo de aprendizagem automática utilizado para a abordagem de pesos é o Random Forest.

Tabela 23 – Resultados experiência 2

Distribuição: Pesos utilizados	Resultados	
	Accuracy (%)	Recall (%)
1 para 1	87.63	56.37
0:9, 1:1	84.34	62.23
0:1, 1:9	87.57	55.83

Experiência 3:

Tabela 24 – Resultados experiência 3

Algoritmo	Resultados	
	Accuracy (%)	Recall (%)
One-Class SVM	59.24	92.56

Experiência 4:

Tabela 25 – Resultados experiência 4

Algoritmo	Resultados	
	Accuracy (%)	Recall (%)
LGBM	58.73	76.34
Random Forest	69.23	59.27
Decision Tree	64.68	59.82
Rede Neural	27.86	96.37
Logistic Regression	31.82	84.27

Experiência 5:

Distribuição: Registos da classe 0 - 7688385 e foram testadas para a classe 1, 60%, 75% e 85% do número de registos da classe 0.

Foram obtidos os melhores resultados para o rácio de 75% do número de resultados da classe 1.

Tabela 26 – Resultados experiência 5

Algoritmo	Resultados	
	Accuracy (%)	Recall (%)
LGBM	89.91	63.53
Random Forest	89.23	61.49
Decision Tree	85.71	59.28
Rede Neural	85.55	71.21
Logistic Regression	82.97	70.85

Experiência 6:

Distribuição: Registos da classe 0 - 7688385 e foram testadas para a classe 1, 60%, 75% e 85% do número de registos da classe 0.

Foram obtidos os melhores resultados para o rácio de 75% do número de resultados da classe 1.

Tabela 27 – Resultados experiência 6

Algoritmo	Resultados	
	Accuracy (%)	Recall (%)
LGBM	86.71	11.60
Random Forest	86.92	20.87
Decision Tree	81.50	28.74
Rede Neural	56.51	56.88
Logistic Regression	40.30	62.38

5.3 Análise

De seguida é apresentado a análise dos resultados as experiências apresentadas anteriormente.

Modelo para clientes recorrentes:

Relativamente o modelo de aprendizagem automática para clientes com histórico, a experiência que obteve melhores resultados foi a experiências 1, onde foi utilizado a técnica de Random Undersampling para lidar com problema do não balanceamento dos dados.

Nesta experiência, os algoritmos que obtiveram a melhor relação entre recall e accuracy, dando prioridade à métrica de recall, foram o LightGBM e Random Forest.

Modelo para clientes novos:

Relativamente o modelo de aprendizagem automática para clientes sem histórico, a experiência que obteve melhores resultados foi a experiências 4, onde foi utilizado a técnica de Random Undersampling para lidar com problema do não balanceamento dos dados.

Nesta experiência, os algoritmos que obtiveram a melhor relação entre recall e accuracy, dando prioridade à métrica de recall, foram o LightGBM e Decision Tree.

Conclusão das Experiências:

Através destas experiências foi possível concluir que para ambos os modelos necessários, o algoritmo LightGBM em conjunto com a técnica de Random Undersampling para lidar com o problema do não balanceamento dos dados, obtiveram os melhores resultados.

O modelo para clientes recorrentes, é classificado corretamente 88.56% dos pagamentos que serão pagos fora do limite e tendo 80% de precisão na classificação geral da fatura, tendo um erro geral de 20%.

O modelo para clientes novos, é classificado corretamente 76.34% dos pagamentos que serão pagos fora do limite e tendo 58.73% de certeza na avaliação geral da fatura, tendo um erro geral de 41.27%.

6 Conclusão

A componente mais importante do sistema desenvolvido é a componente de aprendizagem automática. Embora existam outras alternativas da metodologia a utilizar para este problema, como é o exemplo do processo SEMMA, o CRISP-DM é o processo mais utilizado devido a sua complexidade. Por este motivo apenas o processo CRISP-DM é usado e descrito no presente relatório.

Como referido, o processo escolhido é composto por 4 processos. O primeiro é constituído por um pré-processamento e limpeza de dados, onde são aplicadas técnicas de pré-processamento para tornar os dados coerentes e com uma qualidade desejada.

O segundo é constituído pela análise e seleção de atributos, onde são aplicadas técnicas de extração de novos atributos de forma a criar padrões possíveis. De forma idêntica, são removidos os atributos que não possuem qualquer valor para os algoritmos de aprendizagem automática, como por exemplo as datas, os nomes e os identificadores.

O terceiro é constituído pela construção de modelos de aprendizagem supervisionada utilizando os dados selecionados no processo anterior.

O último processo é constituído pela avaliação e comparação dos modelos de aprendizagem automática criados no processo anterior.

Estes processos repetem-se até que seja considerado que não seja possível obter melhores resultados, adquirindo assim o melhor modelo possível.

Os resultados de classificação do modelo são depois disponibilizados para o utilizador através de um relatório dinâmico criado em Power BI, que fornece uma análise estatística sobre os dados, sendo atualizado automaticamente sempre que novas faturas sejam avaliadas pelo modelo de previsão.

Trabalho futuro:

O modelo desenvolvido é um modelo de aprendizagem supervisionada de classifica binaria, do tipo estático. Estas características fazem com que sempre que se deseje treinar um novo modelo, será necessário treinar com o total do conjunto de dados e não só com os dados novos. Tornado assim o processo de treino mais demorado e com requisitos computacionais maiores.

No futuro, iremos transformar o processo de treino do modelo estático para um modelo do tipo dinâmico, o que otimizará o processo de treino. Isto ocorre uma vez que o modelo será atualizado, ou seja, não será criado um novo modelo com os dados todos, serão apenas necessários os novos dados e o modelo antigo. Diminuindo assim o período de treino e os requisitos computacionais necessários.

7 Referências

[1] "O que é o Power BI?," Microsoft, [Online]. Available: https://powerbi.microsoft.com/pt-pt/what-is-power-bi/. [Acedido em 7 6 2020].

[2] C. Ruci e J. Parks, "Predictive Analytics: Powering The Utility Sector," 31 Maio 2019.

[3] F. Gama, N. Amboni, G. D. Alpersted e M. C. B. Moraes, "Processo de captação de novas oportunidades no desenvolvimento de novos produtos em uma empresa industrial de motores elétricos," Setembro 2016.

[4] A. Sakamoto e L. C. D. Serio, "Cadeia de valor de empresas de software, a partir de suas inovações: um estudo empírico," Janeiro 2009.

[5] A. Lindgreena e F. Wynstra, "Value in business markets: What do we know? Where are we going?," 2005.

[6] W. T, "Conceptualising 'Value for the Customer: An Atttributal, Structural and Dispositional Analysis," 2003.

[7] J. R. Andrade, "CANVAS Modelo estrutura para criação de modelos de negócio".

[8] S. Kumar, "Meet CRISP-DM: An approach to answer all your questions," 4 Maio 2019.

[9] D. Reinsel, J. Gantz e J. Rydning, "The Digitization of the World From Edge to Core," Novembro 2018.

[10] U. M. Fayyad, G. Piatetsky-Shapiro e P. Smyth, "From data mining to knowledge discovery: An overview," *Advances in knowledge discovery and data mining,* 1996.

[11] H. D. Lee, "Seleção de atributos importantes para a extração de conhecimento de bases de dados," 2005.

[12] J. Han, M. Kamber e J. Pei, "Data Mining Concepts and Techniques," *Data Mining,* 2011.

[13] P. Ferreira, "Aplicação de Algoritmos de Aprendizagem Automática para a Previsão de Cancro de Mama," Outubro 2010.

[14] F. Clésio, "7 TÉCNICAS PARA REDUÇÃO DA DIMENSIONALIDADE," 13 Junho 2015.

[15] D. Pyle, "Data Preparation for Data Mining," *Data Preparation for Data Mining,* 1999.

[16] J. M. P. Gama, "Combining Classification Algorithms," 1999.

[17] G. Machado, M. R. Mendoza e L. G. Corbellini, "What variables are important in predicting bovine viral diarrhea virus? A random forest approach," 2015.

[18] J. Suykens, T. V. Gestel, J. D. Brabanter, B. D. Moor e J. Vandewalle, "Least Squares Support Vector".

[19] V. Morde, "XGBoost Algorithm: Long May She Reign!".

[20] S. Jiang, G. Pang, M. Wu e L. Kuang., "An Improved K-Nearest-Neighbor Algorithm for Text Categorization," 2012.

[21] L. Zhou, "Simplify Machine Learning Pipeline Analysis with Object Storage," *Data In, Intelligence Out,* 3 Maio 2018.

[22] B. Boehmke e B. Greenwell, "Hands-On Machine Learning with R," 6 Dezembro 2019.

[23] C. Sammut e G. I. Webb, "Encyclopedia of Machine Learning," 2010.

[24] B. A, C. L, C. C, R. M, S. F e A. S. G, “A One-Class SVM Based Tool for MachineLearning Novelty Detection in HVACChiller Systems,” Agosto 2014.

[25] S. learn, “One-class SVM with non-linear kernel (RBF),” [Online]. Available: https://scikit-learn.org/stable/auto_examples/svm/plot_oneclass.html. [Acedido em 6 2020].

[26] M. Sheinman, “Hyperparameters for Neural Networks,” 2018.

[27] S. Z. Sidney, F. S. Elias, F. d. C. Daniel e B. Salassier, “Reference evapotranspiration estimate in Rio de Janeiro state using artificial neural networks”.

[28] https://www.kaggle.com/, “Kaggle,” 2019.

[29] G. Costa, “Deteção de anomalias utilizando métodos paramétricos e múltiplos classificadores,” Outubro 2014.

[30] D. Hawkings, *Identification of Outliers,* 1980.

[31] M. Markou e S. Singh, “Novelty detection,” 2003.

[32] C. X. Ling e V. S. Sheng, “Cost-Sensitive Learning,” 2010.

[33] Z. Wang, “Practical tips for class imbalance in binary classification”.

[34] J. Brownlee, “Imbalanced Classification,” *Tour of Data Sampling Methods for Imbalanced Classification,* 1 2020.

[35] N. V. Chawla, K. W. Bowyer, L. O. Hall e W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” 2002.

[36] M. Filho, “As Métricas Mais Populares para Avaliar Modelos de Machine Learning,” 25 Maio 2015.

[37] M. Elahi e F. B. Moghaddam, “Cold Start Solutions For Recommendation Systems,” 2019.

[38] H. Peiguang, “Predicting and Improving Invoice-to-Cash Collection,” 2015.

[39] S. Zeng, P. Melville, C. A. Lang, I. Boier-Martin e C. Murphy, “Using Predictive Analysis to Improve Invoice-to-Cash Collection”.

[40] S. K. K. Mani, S. Dechu, C. K. Maurya e T. Tater, “Prediction of Invoice Payment Status in Account Payable Business Process,” Novembro 2018.

[41] M. Antoniou, “Predicting the future popularity of programming languages,” 15 Setembro 2019.

[42] E. Cavalcante, “Sistemas de Controle de Versão,” 2012.

[43] Wikipédia, “FURPS,” [Online]. Available: https://pt.wikipedia.org/wiki/FURPS.