UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE INFORMÁTICA

IDENTIFICAÇÃO DE PERFIS DE CLIENTES BANCÁRIOS: UMA PERSPETIVA DE CIÊNCIA DE DADOS

Ema Alexandra Guilherme Mendes

Mestrado em Ciência de Dados

Dissertação orientada por:
Profª. Doutora Soraia Vanessa Meneses Alarcão Castelo
Eng. Dino Coutinho

2023

Agradecimentos

A conclusão desta dissertação marca a conclusão de uma fase importante na minha vida e o começo de uma nova, é com grande satisfação e entusiamo que expresso aqui o mais profundo agradecimento a todos aqueles que contribuíram para sua concretização. Gostaria de agradecer em primeiro lugar à minha orientadora, a Professora Soraia Meneses Alarcão, e ao meu supervisor, Dino Coutinho, pelo constante apoio prestado durante a realização deste trabalho e mentoria. Quero também deixar um agradecimento à Innovation Makers pela confiança depositada em mim e pela oportunidade que proporcionaram. Finalmente, um agradecimento à minha família e amigos que acreditaram em mim e me apoiaram, e um agradecimento muito especial ao André Torcato, a minha grande fonte de apoio ao longo deste ano.

Dedicatória.

Dedico este trabalho a todo caminho que fiz até aqui e ao meu eu passado. Desde a licenciatura foram incontáveis as vezes que achei que não conseguiria, mas hoje estou aqui. Poderei dizer que sou mestre e não poderia me sentir mais orgulhosa por isso. O caminho nem sempre é fácil, mas ele é necessário para dar valor quando chegamos ao fim. Dedico a tese a todos os momentos de dificuldade superados nesta jornada. E à Ema do futuro, acredita em ti e agora é sempre a subir!

Resumo

Com o avanço tecnológico, para os bancos permanecerem relevantes foi necessário inovarem os seus serviços. Essa tendência surgiu do aumento da competição, envolvendo não só os bancos como também novas entidades (*neobanks*, etc). A análise de dados dos clientes é fundamental para o crescimento e sustentabilidade do banco, pois permite conhecer os clientes com detalhe, aumentando a sua satisfação e fidelização.

A área que desenvolve esta análise denomina-se *Customer Intelligence* (CI). Para um dado banco, os problemas de CI centram-se nas dificuldades em determinar o "valor" de cada cliente ou grupo de clientes, quais os que representam maior risco financeiro e como se deve personalizar as ofertas proporcionadas. Academicamente, os estudos realizados em CI são reduzidos, tanto na sua disponibilidade como na quantidade de métodos aplicados e na apresentação clara das conclusões, para administradores bancários, responsáveis pela tomada de decisão.

O objetivo deste trabalho foi desenvolver uma metodologia robusta e completa, de forma exploratória, utilizando os dados demográficos, transacionais e de empréstimos de clientes bancários. O culminar desta metodologia foi a criação de um *dashboard* que evidencia de forma objetiva o resultado das conclusões essenciais para conhecer um dado cliente, contendo um resumo das análises desenvolvidas, a atribuição de um grupo ao cliente e a caracterização deste.

A metodologia desenvolvida centra-se em quatro fases: 1) resumo do comportamento transacional de um cliente utilizando a análise *Recency Frequency Monetary*, de onde derivaram seis segmentos de clientes; 2) agrupamento dos clientes, utilizando *k-means*, tendo-se determinado a existência de três grupos; 3) identificação dos perfis de cada grupo utilizando técnicas de prospeção de dados; 4) classificação de um cliente, utilizando algoritmos de aprendizagem automática, como o *XGBoost*, que obteve a melhor performance.

Palavras-chave: *customer intelligence*, banca, algoritmos de agrupamento, análise de padrões, reportagem com *dashboard*

Abstract

The banking sector is essential for the functioning of society. With technological advancements, banks had to innovate their services to remain relevant. This trend arose due to the increasing competition, involving not only established banks but also new entities (neobanks, etc). Customer data analysis is crucial for the growth and sustainability of any bank, as it allows for a detailed understanding of customers, thereby enhancing their satisfaction and loyalty.

The field that develops this analysis is called Customer Intelligence (CI). For a given bank, CI issues revolve around difficulties in determining the "value" of each customer or customer group, identifying those posing higher financial risk, and personalizing the provided offers. Academically, studies conducted in CI are limited, both in their availability and the quantity of applied methods, as well as in the clear presentation of conclusions for bank administrators responsible for decision-making.

The objective of this work was to develop a robust and comprehensive methodology in an exploratory manner, using demographic, transactional, and loan data from bank customers. The culmination of this methodology was the creation of a dashboard that objectively highlights the results of essential conclusions to understand a given customer. This dashboard includes a summary of the developed analyses, customer grouping, and characterization of customers.

The developed methodology focuses on four phases: 1) summarizing the transactional behavior of a customer based on RFM analysis, which resulted in six customer segments; 2) customer clustering using the k-means algorithm, revealing the existence of three groups; 3) identification of profiles for each group using data mining techniques; 4) customer classification using machine learning algorithms such as XGBoost, which achieved the best performance.

Keywords: customer intelligence, banking industry, clustering algorithms, pattern analysis, dashboard reporting

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Com o evoluir da tecnologia surgem novas necessidades e expectativas por parte da população ([10]). Os bancos, tendo uma forte presença no dia-a-dia das pessoas, necessitam de estar a par das novas tendências na transição para o digital e da integração com Aprendizagem Automática (AA) nos seus variados serviços e produtos. Assim, ao inovarem com o objetivo de criar melhores condições e atender às necessidades dos seus clientes, mantêm-se nas suas preferências, apesar da existência de cada vez mais concorrência. Ao focarem-se na experiência do cliente (ou *user experience*), os bancos são capazes de promover a satisfação e fidelização destes.

Neste trabalho, desenvolvemos de forma exploratória uma metodologia baseada em métodos de Ciência de Dados (CD), com uma estrutura semelhante à de um projeto em CD ([11]). O objetivo foi estudar o comportamento de clientes através das suas transações e desta forma compreender melhor as suas preferências e necessidades, identificando e caracterizando grupos de clientes similares. Este trabalho está a ser realizado em parceria com a *Innovation Makers* (INM). A INM é uma empresa de serviços e consultoria de tecnologias da informação que atua no setor da banca, com mais de 13 anos de atividade em 7 países, num total superior de 70 projetos desenvolvidos em soluções financeiras inovadoras, com maior presença em Angola e Moçambique. Os projetos focam-se na experiência do utilizador e acompanham o crescimento de negócio digital dos seus clientes incluindo a investigação, desenvolvimento e implementação das soluções a nível de *software* e de *hardware*, em todo o seu ciclo. Entre as soluções destacam-se as de *eWallet*, *Banking*, pagamentos e gestão financeira ([12]).

Em seguida, apresentamos a motivação para o nosso trabalho, os objetivos de investigação e a estrutura do presente documento.

1.1 Motivação

A indústria bancária enfrenta o desafio de compreender e atender de forma eficaz às necessidades específicas dos seus clientes. Por exemplo, um cliente pode valorizar mais as funcionalidades relacionadas com gestão financeira, planeamento de pensões, e gestão de investimentos ou empréstimos imobiliários. Outro cliente pode preferir serviços digitais e transações online, enquanto outro pode valorizar o atendimento presencial num sucursal. Além disso, os compor-

tamentos relacionados com transações também variam: alguns clientes realizam mais frequentemente transferências internacionais, enquanto outros preferem poupanças regulares. Cada cliente possui características únicas e expectativas diferentes, o que torna essencial para os bancos compreenderem essas necessidades e adaptarem os seus produtos e serviços para atendê-las de forma eficiente. As abordagens tradicionais, como produtos padronizados, segmentação demográfica, atendimento presencial e campanhas de marketing massivas, muitas vezes falham em capturar as nuances e complexidades dos comportamentos destes clientes. Assim, surge a área de *Customer Intelligence* (CI), englobada em *Customer Relationship Management*, que consiste na coleta e análise de informações dos clientes para entender as suas necessidades e comportamentos, fornecendo *insights* valiosos para tomada de decisão. Desta forma, as empresas podem personalizar as suas ofertas e melhorar a experiência geral do cliente, impulsionando o seu crescimento desta bem como a fidelização do mesmo.

Abordar este problema é de extrema importância no atual cenário bancário, cada vez mais competitivo. A concorrência surge devido à entrada de novas entidades bancárias no mercado, a globalização, os avanços tecnológicos, a crescente procura por serviços financeiros e a busca de melhores taxas e condições por parte dos clientes. Isto leva os bancos a investir continuamente na inovação, otimização e melhoria dos seus produtos e serviços para atrair e reter clientes.

Neste trabalho, construiu-se uma nova metodologia para extração de conhecimento através de dados dos clientes, de forma exploratória.

A adoção de estratégias de segmentação de clientes orientadas a dados é fundamental no apoio à tomada de decisão. Ao segmentar os clientes com base nos seus perfis de consumo, os bancos podem identificar quais os produtos ou serviços são mais relevantes para cada segmento, permitindo assim uma alocação de recursos mais eficiente e uma oferta mais personalizada. Ao permitir campanhas de marketing direcionadas, a segmentação de clientes aumenta a relevância das comunicações e a probabilidade de conversão, fomentando o *cross-selling* (venda de vários produtos) e *upper-selling* (venda de produtos mais caros) dos seus produtos e serviços. Por último, a segmentação de clientes pode ser utilizada no desenvolvimento de produtos e serviços inovadores que atendam às necessidades específicas de cada grupo. Por exemplo, com base nas preferências de um segmento voltado para a tecnologia, podem ser lançadas aplicações bancárias mais avançadas como soluções de pagamentos digitais ou um Gestor de Finanças Pessoais (GFP). Uma aplicação GFP oferece recursos de gestão financeira abrangentes, permitindo que os clientes monitorizem as suas despesas, criem orçamentos, recebam alertas de gastos excessivos e obtenham *insights* sobre a sua saúde financeira geral. Estas tecnologias impactam positivamente o desempenho dos bancos, a fidelidade dos clientes e impulsionam o crescimento geral do setor bancário (13; 14; 15).

Em suma, conhecer os clientes é diferenciador em qualquer setor, pois permite que as empresas atendam às necessidades e desejos específicos do seu público-alvo. Um exemplo de sucesso que ilustra esta importância é o da Corporação Bancária de Hong Kong e Xangai (HSBC) (16). A HSBC utiliza Inteligência Artificial (IA) para oferecer aos clientes que utilizam cartão de crédito nos EUA uma experiência de compra personalizada. A HSBC está a trabalhar num programa de recompensas que processa dados de clientes para prever como estes irão reclamar os seus pontos, para que possam oferecer melhores ofertas de mercado, como viagens, mercadorias, cartões-presente e dinheiro. A tecnologia recomenda uma categoria de resgate para cada titular de cartão de crédito. A HSBC enviou comunicações por email com base nessas recomendações, ao mesmo tempo que enviou uma categoria aleatória para um grupo de controle. No primeiro lote, cerca de 70% aproveitaram as recompensas e o número de e-mails abertos aumentou 40%, potenciando assim um melhor relacionamento e comunicação com estes.

Este trabalho contribui para o avanço do conhecimento no setor financeiro ao explorar a aplicação de CD, principalmente no desenvolvimento de algoritmos de agrupamento, na segmentação de clientes no setor bancário e análise de perfis com Prospeção de Dados (PD). Os resultados deste trabalho têm implicações práticas para os bancos, bem como outras indústrias que lidem com clientes, como as das telecomunicações e retalho.

1.2 Objetivos

O principal objetivo desta dissertação é identificar perfis de clientes bancários com métodos de análise de dados e implementação de modelos de AA. Mais especificamente, modelos não supervisionados e supervisionados com a finalidade de segmentar clientes e classificá-los, respetivamente. Através de técnicas de prospeção de dados será possível traçar os padrões de cada grupo produzido, para criar os respetivos perfis financeiros. Será elaborado, por último, um *dashboard* que permite resumir os perfis encontrados e *insights*, usando a ferramenta do Power BI. Com isto, podemos formular as seguintes questões de investigação:

1. Quais são os principais atributos e características dos diferentes segmentos de clientes?
2. De que forma as técnicas aplicadas contribuem para a segmentação eficaz dos clientes e a criação de perfis financeiros?
3. Com a metodologia desenvolvida quais recomendações podem ser efetuadas para cada grupo identificado?

1.3 Estrutura do documento

Este documento está organizado da seguinte forma:

- **Capítulo 1 (Introdução)** – introduzimos o problema e a sua motivação. Em seguida, especificamos os principais objetivos do trabalho e a estrututa deste documento;

- **Capítulo 2 (Trabalho Relacionado)** – é efetuada a contextualização geral do problema. Descrevemos o sistema financeiro atual e alguns dos principais produtos e serviços fornecidos pelas instituições bancárias. Em seguida, dada a nossa fonte de dados, resumimos sucintamente o contexto da República Checa nos anos 90. Além disso, apresentamos na mesma secção o trabalho relacionado; para tal descrevemos algumas das aplicações de IA no setor financeiro e o seu impacto. Por fim, abordamos alguns estudos existentes relacionados com a metodologia proposta neste trabalho, dentro e fora da área de estudo, bem como uma revisão dos modelos mais utilizados e respectivas limitações.

- **Capítulo 3 (Dados)** – descrevemos o conjunto de dados usado para desenvolver o trabalho, seguido pela análise exploratória dos dados e limpeza/pré-processamento;

- **Capítulo 4 (Métodos e resultados da metodolgia)** – apresentamos a metodologia desenvolvida, alguns fundamentos teóricos, e os respetivos resultados aplicados ao conjunto de dados da República Checa;

- **Capítulo 5 (Suporte à decisão baseada em dados)** – é apresentada a implementação do modelo através de uma *Application Programming Interface* (API) e a disponibilização dos resultados principais da metodologia num *dashboard*;

- **Capítulo 6 (Conclusões e trabalho futuro)** - apresentamos as conclusões do documento, descrição das limitações e trabalho futuro.

Capítulo 2

Contexto e trabalho relacionado

Dado o contexto específico em que surge este trabalho começamos por descrever os aspetos relevantes do sistema financeiro e os seus produtos e serviços, bem como a perspetiva da banca para o futuro. Sendo o nosso conjunto de dados referente à República Checa, entre os anos 1993 e 1999, resumimos o contexto do país e da indústria bancária nesse período. Em seguida, no trabalho relacionado, discutimos algumas das aplicações atuais de IA no setor financeiro. Por fim, apresentamos o estado de arte quanto aos tipos de técnicas de AA aplicadas a dados bancários, ou análises semelhantes noutras indústrias.

2.1 O sistema financeiro

Tradicionalmente, um banco é uma instituição financeira que aceita depósitos da população e cria um depósito à vista, simultaneamente concedendo empréstimos (17). Alguns dos serviços que prestam incluem depósitos, contas poupança, financiamento de casas e automóveis, empréstimos, transações internacionais, realização de pagamentos de serviços, consulta de extratos, entre outros. Durante muitas décadas, a sua função não saiu desta esfera de permitir e facilitar transações, sendo um lugar seguro para uma pessoa ou empresa ter e gerir o seu dinheiro.

Com a crescente tendência de digitalização que se verifica no século XXI, os bancos expandiram a sua atividade para fora dos balcões presenciais nas instituições, permitindo agora ao cliente aceder à sua conta bancária e efetuar praticamente todas as operações das caixas de multibanco e dos balcões, através da internet ou pelo telemóvel, no que se denomina *internet banking* e *mobile banking*, respetivamente. Outro serviço importante de destacar é o *open-banking*, que permite que o cliente forneça os seus dados financeiros a sistemas externos, possibilitando a outras instituições, como, por exemplo, seguradoras, conseguirem saber o perfil do cliente sem qualquer interação prévia. Isto facilita a adoção de serviços ou produtos por parte dos clientes, dado que as entidades já os conhecem (e.g. o seu nível de risco ou comportamentos), e receber ofertas personalizadas.

Após a pandemia de COVID-19, devido às medidas colocadas em prática pelos governos de vários países, tornou-se clara a necessidade e preferência do uso de canais não presenciais, mas simultaneamente, destacaram-se as suas limitações. Segundo um estudo efetuado por *McKinsey & Company*, uma consultora de gestão que atua a nível global, verificou-se um aumento de 81%

para 95% do nível de adoção digital na Europa só em 2020 (18). Com esta tendência, têm surgido novas possibilidades para efetuar soluções mais personalizadas, visto que é possível coletar dados de toda a interação de um cliente com um dado serviço. Assim, existe uma tendência de, em vez de apostar em estratégias de *marketing* generalizadas para atrair mais público, focar antes em melhorar o envolvimento e relação com o cliente. Isto significa que é necessário oferecer personalização numa escala cada vez maior, que, por sua vez, implica perceber as necessidades e expectativas individuais dos consumidores digitais através da sua pegada digital.

Apesar desta tendência, verifica-se que as instituições financeiras têm dificuldades na inovação digital devido a vários fatores. Por exemplo, embora exista investimento em IA, só 8% dos bancos são capazes de aplicar *insights* preditivos dos seus modelos de AA nas campanhas de marketing. Outro exemplo, só 16% contêm protocolos padrão no desenvolvimento de algoritmos (19) que permitem acionar sobre os *insights* dos respetivos modelos. Além disso, num estudo efetuado pela *PwC Global FinTech Survey 2016* (20), apenas metade dos inquiridos, administradores no setor bancário, acredita que são centrados no consumidor. Em comparação, mais de 80% dos participantes de um semelhante inquérito a *FinTechs*, responderam que acreditam ser centrados no consumidor.

Esta dificuldade de inovação deriva de um extenso conjunto de regulamentações para garantir a estabilidade e a segurança financeira. Essas regulamentações podem limitar a capacidade dos bancos de desenvolver novas ideias e tecnologias, pois precisam de cumprir requisitos específicos. Muitos bancos têm sistemas e infraestrutura antigos que foram desenvolvidos ao longo de décadas. Esses sistemas podem ser complexos e difíceis de modificar, o que dificulta a implementação de soluções inovadoras; além disso a modernização dessas infraestruturas pode ser um processo demorado e caro. Os bancos precisam também de ser cautelosos ao introduzir novas tecnologias e serviços para garantir a conformidade com as leis e regulamentos aplicáveis, precisando de avaliar cuidadosamente os riscos associados à inovação, como por exemplo a cibersegurança, proteção de dados e prevenção de fraudes (21; 22).

Ainda assim, os bancos tradicionais enfrentam o desafio adicional de se manterem competitivos e na escolha dos seus consumidores diante das novas tecnologias e concorrentes modernos, reforçando um cenário de constante mudança e necessidade de inovar.

Destacam-se, em seguida, algumas das entidades a surgir no setor bancário (23):

- *FinTechs*: empresas de inovação tecnológica financeira que oferecem serviços financeiros inovadores, como pagamentos digitais, empréstimos *peer-to-peer*, gestão de investimentos e *cashback*. Alguns exemplos conhecidos são o *PayPal*, *Square* e *TransferWise*.

- *Big Techs*: Gigantes da tecnologia, como a *Google*, *Amazon*, *Apple* e *Facebook*, que estão a explorar oportunidades no setor financeiro, oferecendo serviços como pagamentos móveis, carteiras digitais e assistentes virtuais para questões financeiras.

- *Neobanks*: Instituições financeiras digitais que operam exclusivamente online, sem agências físicas, e fornecem serviços bancários simplificados e personalizados, como a *N26*, *Revolut*

e *Chime*. Normalmente estes são capazes de oferecer melhores condições, por exemplo em relação à isenção de comissões e taxas.

- Telecomunicações: Empresas de telecomunicações que expandiram as suas ofertas para incluir serviços financeiros, como contas bancárias móveis e transferências de dinheiro. Um exemplo é o *Orange Bank*, lançado pela empresa de telecomunicações *Orange*.

Não obstante, os bancos tradicionais possuem igualmente algumas vantagens exclusivas. Primeiro, dado existirem há mais tempo e estarem bem estabelecidos, possuem um nível de experiência bem como de confiança com os clientes que lhes é única. As pessoas sentem-se mais seguras em utilizar serviços proporcionados pelos bancos tradicionais (24). Em segundo lugar, estes têm uma grande base de dados histórica, que permite ter um excelente ponto de partida para investir, com um certo grau de confiança, em projetos baseados em dados (25).

Ao focar em construir e nutrir a relação com o cliente, os bancos conseguem aumentar o grau de satisfação e nível de retenção dos mesmos. Este foi o caso do *Royal Bank of Scotland*. Este banco no Reino Unido combinou a análise de grandes quantidades de dados, inlcusive interações com os utilizadores e dados de texto, para ter um panorama completo das reclamações dos clientes e poderem identificar e agir de forma mais rápida e eficaz (26). Além disso, o *Royal Bank of Scotland* conseguiu perceber onde os seus clientes estavam a perder dinheiro, como por exemplo a pagar duplamente por serviços em pacotes de contas, e alertar para estas situações (27).

Resumindo, o desafio dos bancos atualmente é em apostar na confiança e suporte aos clientes: apesar de cada vez menos existir uma interação cara-a-cara, é possível conhecê-los a nível pessoal e oferecer ativamente serviços/soluções valiosas para as suas necessidades financeiras e pessoais.

Sendo os dados do nosso trabalho referente ao uso dos canais bancários pode ser importante perceber que tipo de operações um cliente tem ao seu dispor visto que podem ser fundamentais para a segmentação de clientes. Os bancos comerciais vendem e oferecem aos seus clientes um leque de produtos e serviços semelhantes, entre instituições bancárias (28). Destacamos os seguintes:

- Depósito à ordem: o tipo de depósito associado à abertura de conta e que permite a movimentação dos fundos depositados a qualquer momento;

- Depósito a prazo: depósito feito a termo, onde é estipulado um termo para o seu reembolso, só podendo ser restituído terminado o prazo estipulado;

- Produto de crédito: A concessão de crédito é o ato pelo qual um banco coloca dinheiro à disposição da pessoa ou empresa, ficando estas responsáveis por fazer a devolução na data acordada no contrato, com um valor extra no que toca ao cálculo da taxa de juro sobre o valor a prescindir. Os bancos disponibilizam diversas modalidades de crédito com diferentes custos e comissões como é o caso de crédito à habitação, crédito pessoal, crédito automóvel, microcrédito e cartões de crédito;

- Transferências bancárias: trata-se de uma ordem dada por um consumidor (ordenante) ao seu banco, para movimentar os seus fundos a débito de uma conta de depósito de que ele é titular e a crédito de outra conta (beneficiário), no mesmo banco ou noutro banco;

- Cheques bancários: corresponde a uma ordem de pagamento dada pelo titular da conta ao banco, para que este pague o beneficiário, o titular ou terceiro, determinada quantia. Este pode também ser atribuído a título de crédito, ou seja, um documento que representa um crédito que alguém tem sobre outra pessoa;

- *Internet Banking* e *mobile banking*: são serviços que permitem aos clientes bancários acesso, através da Internet, às suas contas e um conjunto de operações bancárias que são frequentemente efetuadas numa agência ou numa caixa de pagamento automático, sem ser necessário a deslocação dos clientes aos respetivos pontos;

- Cartões de débito e crédito: os cartões de débito são instrumentos emitidos pelos bancos membros do sistema bancário, utilizados eletronicamente em caixas automáticas e Terminal de Pagamento Automático (TPA), para movimentação e consulta de contas de depósito à ordem, ao passo que os cartões de crédito são instrumentos de pagamento, com a forma de cartões ou outro dispositivo ou código, que são fornecidos pela instituição financeira emissora para possibilitar aos seus utilizadores a realização de transações financeiras, como por exemplo, pagamentos e/ou levantamentos de numerário, nos terminais que os aceitem.

Além destes produtos, podemos afirmar que o futuro será caracterizado por uma transformação significativa impulsionada pela tecnologia. Os bancos compreendem e fazem um esforço no sentido de se adaptarem às necessidades dos seus clientes e à evolução das mesmas, oferecendo serviços digitais cada vez mais sofisticados. Espera-se uma maior enfâse nas experiências de utilizador. Além disso, os bancos estão a explorar tecnologias como IA e AA para oferecer recomendações personalizadas de produtos financeiros (29), serviços de atendimento ao cliente automatizados e análises avançadas para gestão de riscos. A expansão da tecnologia *blockchain* também pode levar a uma maior eficiência das transações, reduzindo a necessidade de intermediários e acelerando processos como no caso de pagamentos internacionais.

O futuro do sistema financeiro atual será caracterizado por uma digitalização e uma maior conexão entre instituições financeiras e consumidores. Um estudo desenvolvido pela *McKinsey & Company* descobriu que as empresas que se destacam recorrendo à personalização geram 40% mais receita dessas atividades do que os concorrentes (10). Espera-se uma maior inclusão financeira, com o acesso a serviços financeiros extendido a populações mais desfavorecidas, impulsionado por soluções inovadoras como *Fintechs* e criptomoedas. No entanto, com a crescente dependência da tecnologia, surgem desafios relacionados com a cibersegurança e a proteção de dados pessoais, pelo que será necessário um equilíbrio cuidadoso entre inovação e proteção dos consumidores.

2.2 Indústria bancária na República Checa dos anos 90

Durante os anos 90, a República Checa passou por um período de transição significativo após a dissolução da Checoslováquia, em 1993. Essa década foi marcada por mudanças políticas, económicas e sociais.

No início dos anos 90, a República Checa adotou uma abordagem de transição rápida para uma economia de mercado. Isso envolveu a privatização de empresas estatais, incluindo os bancos, que, anteriormente, eram controlados pelo Estado. A privatização foi acompanhada por reformas para reestruturar e modernizar o setor bancário (30). Durante esse período, muitos bancos estrangeiros estabeleceram presença na República Checa, trazendo consigo *expertise*, capital e tecnologia avançada que ajudou a fortalecer o sistema bancário do país e a introduzir práticas bancárias modernas. Os bancos estrangeiros também competiram com os bancos domésticos, impulsionando a eficiência e a qualidade dos serviços oferecidos (31).

A liberalização financeira e a abertura dos mercados permitiram que os bancos checos expandissem os seus serviços, diversificando os seus produtos e alcançando um público mais amplo. O setor bancário viu um aumento na oferta de empréstimos para empresas e consumidores, facilitando o crescimento económico e o desenvolvimento de pequenas e médias empresas (32).

No entanto, esse período de transição também apresentou alguns desafios. A rápida liberalização e a falta de regulamentação adequada levaram a um aumento da especulação financeira. Isso culminou numa crise financeira no final dos anos 90, quando alguns bancos enfrentaram problemas de liquidez e incumprimento. Para lidar com esses problemas, o governo checo implementou medidas regulatórias mais rigorosas e adotou uma abordagem mais cautelosa em relação à liberalização financeira. O setor bancário passou por um processo de consolidação, com fusões e aquisições que fortaleceram as instituições financeiras e melhoraram a estabilidade do sistema.

Em suma, os anos 90 foram um período de transformação e desafios para a indústria bancária na República Checa. Apesar das dificuldades previamente mencionadas, o setor bancário emergiu mais forte e resiliente. As reformas realizadas naquela época ajudaram a estabelecer as bases para um sistema bancário sólido e bem-regulado, que continua a desempenhar um papel vital no apoio ao crescimento económico e ao desenvolvimento do país.

2.3 Inteligência artificial no setor financeiro

A IA foi fundada há mais de 70 anos, mas só ultimamente tem sofrido um crescimento exponencial nas suas aplicações, em grande parte devido aos avanços na área, melhoria do *hardware* em termos de processamento e armazenamento, e o surgimento de armazenamento e computação em nuvem (33).

As tecnologias de IA estão cada vez mais a ser integradas nos setores financeiros a nível global. O desenvolvimento e popularização de *internet banking* e *mobile banking* foi um dos grandes impulsionadores do crescimento da informação bancária em tempo real. Aliado à disponibilidade cada vez maior de grandes volumes de dados de outras fontes torna-se cada vez mais impor-

tante conseguir implementar ferramentas analíticas nos serviços que consigam tirar proveito dessa informação em tempo útil, i.e., que seja possível tomar ações durante a janela de oportunidade identificada (34). Algumas das principais aplicações consistem na melhoria de satisfação do cliente, *marketing* e otimização de gestão estratégica, segurança e deteção de fraudes, gestão de risco e investimentos e finalmente, estratégia de expansão da rede de balcões presenciais e avaliação de performance interna.

Mais especificamente, no que toca à gestão de relação com o cliente, o cerne deste trabalho, o uso de técnicas de PD, ou *data mining*, tem sido bastante utilizado. As principais técnicas aplicadas incluem uma combinação de análise de grupos, regras de associação e técnicas de classificação e previsão (35). Especialmente em economias emergentes, i.e. de países em desenvolvimento, a aplicação destas tecnologias tem tido grande impacto (36), não só na automatização dos seus processos de negócio como também na inclusão de novas fontes de dados para ultrapassar determinados obstáculos. Um exemplo é a atribuição de microcrédito a clientes com baixo rendimento ou sem histórico bancário. Fazendo uso de outras fontes de dados provenientes de telemóveis e satélites, conseguem determinar a identidade e credibilidade de indivíduos necessária para o acesso aos serviços financeiros. Assim, AA e IA têm tido um grande papel na luta pela inclusão financeira (37).

Outra importante aplicação de IA, consiste na automatização de processos relacionados com a gestão de relacionamento com o cliente e serviço ao cliente. Dado a baixa acessibilidade nos países em desenvolvimento a sucursais e serviços bancários, o custo de chegar aos mesmos é alto em comparação com as suas transações e a receita que contribuem. Desta forma, a IA pode ajudar a reduzir os custos associados, ao permitir o processamento de um grande volume de transações de baixo valor e portanto com u custo mais reduzido, que por sua vez, converte-os em clientes com potencial para o banco (37).

De facto, IA já está a deixar a sua marca no setor bancário. Um estudo feito pelo *World Economic Forum*, em 2020, a *FinTechs* e bancos tradicionais concluiu que 85% dos participantes (de um total de 151 entidades), estão a usar algum tipo de aplicação de IA nos seus serviços (38), e no relatório da *Insider Intelligence* acerca do uso de IA no setor bancário, cerca de 80% dos bancos reconhece as vantagens oferecidas por estas tecnologias (39). Enquanto muitos bancos estão a iniciar a sua jornada, esta tecnologia já provou ser um grande avanço para conectar com os clientes e elevar sua experiência bancária (40). As suas aplicações desempenharão um papel importante o futuro do setor bancário (41).

2.4 Segmentação de clientes

Um dos focos desta dissertação é a segmentação de clientes e a construção de perfis financeiros recorrendo a algoritmos de AA. Esta é uma subárea da IA que se foca no desenvolvimento de algoritmos e modelos capazes de aprender a partir de um conjunto de dados. A aprendizagem automática pode ser dividida em supervisionada e não supervisionada, sendo ambas vertentes utilizadas no trabalho. Sendo assim, é importante distingui-las. A AA não supervisionada utiliza

dados não classificados, sendo o objetivo destes modelos, através da exposição às características dos dados sem as saídas esperadas, encontrar padrões, estruturas ou relações íntrinsecas nos dados e classificá-los de acordo com o encontrado. Este é o objetivo da segmentação de clientes, onde, somente com acesso às suas características, pretende-se descobrir qual a estrutura íntrinseca existente nos dados de clientes bancários. Por outro lado, a aprendizagem automática supervisionada constrói modelos que aprendem a classificar os rótulos já existentes nos dados. Por exemplo, os modelos supervisionados podem tentar prever se uma transação é fraudulenta ou legítima com base em exemplos rotulados, enquanto modelos não supervisionados tentam identificar grupos de transações sem rótulos, como segmentação de clientes com base em padrões de gastos.

Além disso, é necessário discutir a diferença entre segmentação de clientes e criação de perfis, duas análises relacionadas mas com objetivos diferentes. A segmentação é o ato de dividir uma base de dados em secções distintas ou segmentos. Existem duas abordagens para segmentação: orientada a mercado e orientada a dados. A primeira permite que os gerentes usem características que determinam ser importantes impulsionadoras para os seus negócios, enquanto que a orientada a dados usa técnicas de agrupamento para encontrar grupos homogéneos, com base em métricas de distância. A criação de perfis é o ato de utilizar dados para descrever ou traçar o perfil de um grupo de clientes definido (42). Existem vários tipos de segmentação (43) que se encontram resumidos na tabela 2.1.

Tabela 2.1: Tipos de segmentação e exemplos de variáveis.

Tipo de segmentação	**Descrição**	**Exemplo de variáveis**
Geográfica	Baseada em região, cidade ou densidade.	Região: Norte, Sul, Centro
Demográfica	Com base em variáveis demográficas.	Idade, Género, Salário
Psicográfica	Baseada em classe social, estilo de vida e personalidade.	Classe social: Alta, Média, Baixa
Comportamental	Tem em conta padrões comportamentais, por exemplo pela interação do cliente com determinado produto ou serviço, como a conta à ordem, o cartão de crédito ou o *Internet Banking*	Frequência de uso do cartão de crédito, Frequência de login *Internet Banking*

As áreas de estudo em que são aplicadas técnicas de segmentação de clientes são inúmeras, para este trabalho de pesquisa destacaram-se as áreas da banca, de telecomunicações e saúde pelos métodos e análises semelhantes ao que pretendemos implementar.

Foi efetuada uma revisão sistemática dos vários algoritmos de segmentação de clientes (ver tabela 2.2) onde encontrámos um extenso leque de modelos que podem ser utilizados para a análise de dados bancários. Resumindo, a tabela 2.2 apresenta diversos modelos de segmentação de dados baseados em diferentes métodos. Os métodos baseados em centróides, como *k-means* e *k-medoids*, são fáceis de implementar, porém o primeiro exige a escolha manual do número de grupos, enquanto o último destaca-se por conseguir lidar com *outliers*. Os métodos hierárquicos, como *Hi-*

erarchical Clustering Aglomerative (HCA), não exigem um número de grupos pré-definido, mas podem ser sensíveis à métrica escolhida e não lidam bem com tamanhos de grupos variados. Para colmatar esta limitação, surgem os métodos de densidade, como *Density-Based Spatial Clustering of Applications with Noise* (DBSCAN), que são eficazes em identificar grupos sobrepostos e *outliers*, mas falham ao lidar com densidades diferentes. Abordagens probabilísticas, como o *Gaussian Mixture Model*, utilizam distribuições para lidar com grupos sobrepostos, mas são mais complexas devido a funções de custo e pressupostos de distribuições normais. Além disso, outras alternativas incluem a *Fuzzy Theory* que permite um ponto pertencer a mais que um grupo, o *BIRCH* eficiente em memória e grandes quantidades de dados, e o 2 *Step Clustering*, uma melhoria do *k-means*, pois é capaz de automaticamente determinar o número ideal de grupos.

No que toca à segmentação de clientes o algoritmo mais utilizado na literatura continua a ser o *k-means* principalmente devido à fácil implementação e interpretabilidade, apesar de menos sofisticado.

Tabela 2.2: Modelos utilizados para segmentar dados e as suas respectivas vantagens e desvantagens.

Modelo	Características	Fonte
k-means	**Vantagens**: • Simples de implementar e de escalar; • Converge sempre e adapta-se facilmente a novos exemplos. **Desvantagens**: • Escolha manual do número de grupos; • Depende dos valores iniciais; • Não consegue formar grupos de diferentes tamanhos e densidades; • Sensível a *outliers*.	(44), (45) (46), (47) (48),(49) (50), (51) (52)
Fuzzy Theory (fuzzy c-means)	**Vantagens**: • Funciona melhor que *k-means* quando existem pontos que se sobrepõem; • Os pontos não estão restringidos a pertencer só a um grupo. **Desvantagens**: • Algoritmo mais lento; • Sensível à inicialização da matriz dos pesos.	(53), (49) (54)
k-medoids	**Vantagens**: • Fácil de implementar e é escalável; • Converge rapidamente e é robusto; • Menos sensível a *outliers* que *k-means*. **Desvantagens**: • Não é adequado quando os grupos têm formas arbi-	(55), (47)

Continua na próxima página

	trárias; • Depende dos valores iniciais.	
Hierarchical Clustering	**Vantagens**: • Não é necessário especificar o número de grupos; • Constrói uma estrutura hierárquica de grupos. **Desvantagens**: • Não consegue lidar com grandes quantidades de dados; • Depende da métrica escolhida, podendo ser sensível a ruído e *outliers*; • Não consegue lidar com grupos de tamanhos diferentes; • A ordem dos dados tem impacto no resultado final.	(56)
DBSCAN	**Vantagens**: • O modelo baseia-se em variância e distribuições probabilísticas para segmentar os dados e atribuir as classes; • Consegue lidar bem com grupos que se sobreponham; • Bons para detecção de anomalias. **Desvantagens**: • Não consegue identificar grupos que variam em densidade; • Não funciona bem com dados multidimensionais; • Lento a executar.	(57), (58)
Gaussian Mixture Model	**Vantagens**: • O modelo baseia-se em variância e distribuições probabilísticas para segmentar os dados e atribuir as classes; • Consegue lidar bem com grupos que se sobreponham; • Bons para detecção de anomalias. **Desvantagens**: • Os algoritmos de funções de custo não são triviais; • Assume distribuições normais nas variáveis; • Mais lento que o *K-means*; • Dificuldade em incorporar no modelo variáveis categóricas.	(59)
BIRCH (balanced iterative reducing and clustering using hierarchies)	**Vantagens**: • Funciona muito bem com conjuntos de dados grandes; • Não faz um varrimento de todos os dados e grupos existentes, sendo eficiente em memória; • Método dinâmico e incremental. **Desvantagens**: • Não consegue lidar com variáveis categóricas; • Tem que se escolher manualmente o número de grupos.	(60)

Continua na próxima página

2 Step Clustering	**Vantagens**: • Uma melhoria em relação ao *k-means* clássico na medida em que automaticamente define o melhor número de grupos; • Usado para grandes quantidades de dados. **Desvantagens**: • Influenciado pela ordem das instâncias nos dados.	(44), (51) (61)

De seguida, apresentamos em maior detalhe os trabalhos que se destacam pelas metodologias utilizadas. Em (44) foram comparados os resultados dos métodos *k-means*, com variáveis normalizadas e não normalizadas, e *Two Step Clustering* para segmentar clientes em função do tipo do programa de fidelização em que se enquadraram, usando variáveis como a quantia do empréstimo, o tempo de relação com o banco e o número de pagamentos em falta nos últimos 12 meses. Desta análise resultaram 3 *tiers* de clientes.

O método *k-medoids*, parte da família dos métodos *k-Center*, foi escolhido em (55) para segmentar investidores e delinear os perfis. Foi também monitorizado como os padrões de comportamento mudam com o tempo devido a flutuações de mercado, tendo sido analisado mudanças individuais de classe, e também a distância de cada investidor ao centro do seu grupo, permitindo criar séries temporais que se traduzem em *insights* que antes não existiam. Foram encontrados 4 perfis de investidores.

Para segmentar clientes em função do seu estilo de vida com base em dados transacionais de retalho, no estudo (56), foi usado um método de *Divisive Hierarchical Clustering*, especificamente, o VARCLUS que utiliza a correlação entre variáveis como uma métrica de distância entre pontos. Os dados incluiam características demográficas como a data de nascimento, género, número de pessoas do agregado e nome, bem como informações das transações realizadas, a data e hora da transação, a loja, os produtos e respetivos preços. Foram encontrados e caracterizados 6 tipos diferentes de grupos através da visualização das distribuições das categorias dos produtos comprados e seus preços e características demográficas.

Para colmatar as desvantagens de usar o algoritmo *K-Means*, em (48) foi implementada uma rede neuronal *Self Organizing Map* para determinar o número ideal de grupos de um método *K-means*, para providenciar um modelo dinâmico de PD, na segmentação de clientes bancários. Foram identificados 3 grupos diferentes, com características bastante distintas no que toca à quantia de empréstimos, contribuição para o banco e proporção de clientes.

Em (59), os autores aplicaram uma abordagem à segmentação baseada em modelos, mais especificamente, em estatística Bayesiana recorrendo a *Gaussian Mixture Models*, diferindo da usual abordagem determinística, de forma a agrupar instituições bancárias de acordo com determinadas métricas e indicadores financeiros. As variáveis selecionadas incluem o património total, o total de empréstimos e créditos, lucro e despesa líquida, número de empregados, capital social e capital próprio.

2.5 Classificação de segmentos de clientes

Após efetuar a segmentação de clientes, é possível utilizar um método de classificação para identificar, com base nas classes atribuídas, qual a classe a que um novo cliente pertence sem ser necessário retreinar o modelo de agrupamento, processo que pode levar bastante tempo e ser computacionalmente pesado bem como modificar os grupos anteriormente analisados, dependendo da quantidade de clientes e dados existentes.

Em relação ao problema de classificação, apresentamos na tabela 2.3 o estado de arte quanto aos modelos mais utilizados bem como as respetivas vantagens e desvantagens. Os métodos de conjunto (ou *ensemble*), como *Random Forests*, oferecem elevada *accuracy*, em geral, e escalabilidade para grandes volumes de dados, embora exijam mais complexidade no ajuste de parâmetros e requeiram um poder computacional considerável. Modelos baseados em probabilidades, como *Naive Bayes*, são simples e rápidos, embora assumam independência condicional. Redes neurais, como *Neural Networks*, são indicadas para aprendizagem contínua, mas necessitam de dados volumosos e ajustes de parâmetros complexos. Regressão Logística é simples e ágil, mas limitada a decisões lineares. Máquinas de Vetores de Suporte são eficientes em dados multidimensionais e podem lidar com complexidade, embora possam ter desempenho inferior em grandes conjuntos de dados, com muitas instâncias. Por fim, o () é rápido e robusto, todavia pode enfrentar dificuldades com dados não estruturados e apresentar tempos de treino mais longos.

Tabela 2.3: Modelos utilizados para classificar dados e as suas respectivas vantagens e desvantagens.

Modelo	Características	Fonte
Random Forests	**Vantagens**: • Modelo geralmente com elevada *accuracy*; • Eficiente com grandes quantidades de dados; • Conseguimos perceber a contribuição de cada variável na atribuição da classe; • Modelo robusto. **Desvantagens**: • Existe um elevado número de parâmetros; • Requer um elevado poder computacional (mas pode ser paralelizado); • Elevado tempo de treino; • Difícil de interpretar.	(46), (51)
Naive Bayes	**Vantagens**: • Simples de implementar; • Muito rápido e escalável;	(60)
Continua na próxima página		

	• Resiliência a ruído e pouco suscetível a *overfitting*. **Desvantagens**: • Assume independência condicional, o que na maioria dos casos não é verdade; • Carece do problema da frequência zero, sendo necessário fazer um *smoothing* netes casos; • Pode ser demasiado simples para casos com grande complexidade.	
Neural Networks	**Vantagens**: • Bom para casos de aprendizagem contínua dado que não precisa de ser treinado com os dados iniciais; • Tolerância a falhas; • Pode aproximar qualquer função contínua e diferenciável. **Desvantagens**: • Falta de transparência nas decisões; • Necessita de grandes quantidades de dados; • Difícil de construir a arquitetura e otimizar; • Para redes muito complexas, o tempo de treino é muito grande e requer muitos recursos computacionais.	(62), (63)
Logistic Regression	**Vantagens**: • Fácil de implementar e eficiente; • Pode facilmente ser estendido a múltiplas classes (*multinomial logistic regression*); • Rápido a classificar; • Menos suscetível a *overfitting*. **Desvantagens**: • Constrói apenas superfícies de decisão lineares; • Baixa *performance* em problemas complexos.	(63)
Support Vector Machines	**Vantagens**: • Eficiente em dados multidimensionais (principalmente quando o número das dimensões é maior que o número das amostras); • É eficiente no uso da memória; • É possível introduzir não linearidade com o *kernel trick*; • Modelo robusto e com regularização. **Desvantagens**: • Tem menor *performance* em conjuntos de dados com muitos dados; • Sensível a ruído e sobreposição de classes; • Dificuldade em parametrizar o modelo e interpretar.	(62)

Continua na próxima página

XGB (Extreme Gradient Boosting)	**Vantagens**: • Tem implementado paralelização de processamento sendo bastante rápido e eficiente; • Pode utilizar-se regularização ; • Modelo robusto e resiliente. **Desvantagens**: • Não funciona bem com dados esparsos ou não estruturados; • Sensível a *outliers*; • Falta de transparência nas decisões; • O tempo de treino pode ser elevado.	(64)

Podemos assumir que, dada a revisão bibliográfica e natureza do problema, serão produzidos mais que dois grupos. Por isso, os algoritmos que se destacam mais na tarefa de classificação são as *Random Forests* e o XGB. Apesar de um elevado número de parâmetros é possível regular a complexidade do modelo, ambos são robustos e escaláveis em grandes quantidades de dados e interpretáveis (através da importância das variáveis preditoras). Estas características são cruciais para o tipo de modelo explicativo que pretendemos desenvolver.

Em (46), os autores criaram uma metodologia para encontrar os padrões de clientes de uma empresa de telecomunicações, por forma a conseguir prever o seu comportamento no futuro. Recorream ao algoritmo *k-means* para segmentar os clientes e, posteriormente, uma rede neuronal e árvore de decisão para classificar e prever o nível de "atratividade" destes, e de *churn* primária e secundária. Esta análise é feita com base em características sociodemográficas como a região, estado civil, salário, nível de escolaridade, anos de empregabilidae e se está reformado, e comportamentais, como as horas de uso de cada serviço e que tipo de serviços aderem. O nível de "atratividade"do cliente está relacionado com o seu valor para a empresa de telecomunicações, sendo maior se o cliente for mais ativo e usufruir de um maior leque de serviços. Assim identificaram 6 classes de clientes com diferentes níveis hierarquicos de atração.

Em (64) foi desenvolvido um modelo com o XGB para prever a falência do banco (classificação binária) com base em indicadores financeiros. Este problema tem um conjunto de dados muito desbalanceado, pelo que foi efetuado um pré-processamento cuidadoso, desde a seleção de variáveis preditoras a testes de hipótese ANOVA. Este modelo obteve uma *accuracy* de 96%, superando os restantes modelos de *boosting*. As 21 variáveis utilizadas para efetuar a classificação representam a rentabilidade, a qualidade dos bens, capitais, depósitos e liquidez bancária de cada banco.

2.6 Descoberta de padrões

Visto que os dados do trabalho exploram as transações bancárias, uma ferramenta de PD muito poderosa de implementar para descobrir os padrões de utilização dos clientes bancários é a construção de *itemsets*, i.e., padrões e regras de associação, que fazem parte da esfera da aprendizagem au-

tomática semi-supervisionada. O propósito das regras de associação é encontrar relações significativas entre itens ou variáveis que ocorram frequentemente numa base de dados transacional.

Em (53) foi efetuada a segmentação de clientes com base em transações na área da indústria de fibras, utilizado o método *Fuzzy C-means*. Em seguida, os autores escolheram o grupo com os clientes mais valiosos para a empresa com o objetivo de analisar os seus comportamentos de consumo recorrendo a regras de associação para perceber que tipo de produtos costumam ser comprados em conjunto, tendo encontrado uma relação entre 5 produtos. Foi efetuado um estudo comparativo para perceber se existiam diferenças significativas nas regras de associação obtidas para os outros grupos, tendo sido encontrados padrões de compra distintos característicos de cada grupo, o que permite compreender melhor as preferências dos clientes.

Á semelhança ao estudo anterior, em (49), foram utilizados dados transacionais e de contas bancárias para segmentar clientes em 3 grupos diferentes com o *k-means*. O método *Apriori* determinou as regras de associação dos grupos para caracterizar os grupos de clientes e criar os perfis de cada grupo. Como pré-processamento, além do tratamento de dados incompletos, foram agregado os comportamentos e feita uma análise RFM. Foram construídas regras de associação com as variáveis mais importantes para a classificação dos grupos, onde cada regra representa um perfil de cliente que esteve predominante no seu grupo.

Em (50), foi construída uma metodologia para obter regras de associação de âmbito mais geral, através de uma grande quantidade de dados. Primeiro foram descobertas as regras individuais de todos os clientes de forma a capturar o comportamento pessoal, com dados sociodemográficos e transacionais. Depois, os clientes são agrupados com base em preferências similares de forma a reduzir a complexidade das regras. Por fim, obteve-se as regras gerais dentro de cada grupo, combinando regras similares, reduzindo a especificidade. Assim, foi possível obter segmentos de clientes que revelam o comportamento do grupo. A metodologia foi aplicada a um conjunto de dados com os serviços de um hotel.

Outra forma de determinar regras entre variáveis, nomeadamente, de forma a perceber as associações para uma dada classe é utilizar árvores de decisão como modelo de classificação. Por base, estes são modelos que constroem regras de decisão para classificar os dados, pelo que podem gerar *insights* interessantes sobre os mesmos. Este tipo de técnica é bastante utilizado em conjuntos de dados de cariz comercial ou industrial (65).

2.7 Discussão

Nas secções 2.1 e 2.2, contextualizámos o sistema financeiro atual e algumas das suas limitações e tendências, nomeadamente, a elevada concorrência, devido à entrada de novos adversários, como também o rápido avanço tecnológico com a integração de IA nos serviços bancários. Os bancos têm vindo a alterar a sua estratégia para ser cêntrica na experiência do cliente e assim garantir a sua satisfação e fidelização, recorrendo à personalização. Percebemos, assim, a importância de conhecer os clientes através da análise dos seus dados. Relativamente à República Checa, compreendemos que os dados utilizados foram originados de um período de transformação e desafios

na indústria bancária e população checa que pode revelar uma realidade nos dados também fora do normal. Em seguida, apresentamos a pesquisa do estado de arte quanto ao tipo de técnicas de AA aplicadas a dados bancários, ou análises semelhantes noutras indústrias, com o objetivo de compreender o que foi trabalhado previamente e como estender essas análises na nossa metodologia.

Percebemos que as tecnologias IA estão assim a começar a ser integradas nos produtos e serviços bancários, em vários tipos de aplicações, desde segurança a *marketing*, elevando a experiência bancária dos seus clientes. Para garantir a personalização e identificar perfis financeiros, é necessário estudar o comportamento dos clientes através da análise dos seus dados bancários.

Em primeiro lugar, o banco deverá determinar quais os grupos mais relevantes de clientes, numa segmentação inicial caso ainda não exista. Para tal, recorre-se a algoritmos de agrupamento, uma forma de AA não supervisionada, pois não conhecemos os grupos *a priori*. Academicamente, existem diversos trabalhos com o objetivo de segmentar entidades usando diferentes tipos de algoritmos, desde os baseados em centróides a métodos hierarquicos, onde cada um possui as suas vantagens e desvantagens (ver secção 2.4). Para o problema em questão, sabendo que os dados do trabalho têm um tamanho razoável mas que não justifica o uso de um algoritmo mais efieciente todavia complexo, como o *BIRCH*, o *k-means* demonstra ser o mais indicado, pois é fácil de implementar e simples de otimizar, com dados nunca antes explorados neste contexto. É interessante, ainda assim, testar um método que consiga lidar com *outliers* ou identificar grupos com base em regiões densas, o caso do *DBSCAN*, com o objetivo de perceber qual o mais indicado para o presente problema.

Tendo identificado os grupos, é necessário treinar um modelo supervisionado que, com as classificações já efetuadas do grupo obtidas do modelo anterior, consiga aprender o que distingue cada grupo com o objetivo de classificar um novo cliente, ou reclassificar um cliente, periodicamente. Sendo à partida um problema multiclasse, i.e., com mais de dois grupos, após revisão da literatura (ver secção 2.5), destacaram-se os algoritmos , *Random Forests* e regressão logística, pois são robustos e eficientes. Testaremos também um algoritmo simples, o *Naive Bayes*, para nos fornecer uma base de comparação das métricas resultantes dos testes. O melhor será escolhido com base em métricas de avaliação.

Por último, tendo identificado os grupos e sendo possível atribuir rapida e automaticamente um grupo a qualquer cliente através do modelo de classificação, resta estudar mais profundamente os grupos obtidos para lá de uma análise de dados simples com base em visualizações e estatísticas. Os algoritmos de descoberta de padrões (ver secção 2.6), nomeadamente o *A priori* e, por extensão, a deteção de padrões sequenciais, são vias interessantes de explorar dado existerem dados de créditos e de transações com informação temporal. Estes métodos permitem, por exemplo, perceber que padrões existem em cada grupo e quais as diferenças entre si em relação a capacidade financeira, estudando o balanço da conta e os montantes de transações disponíveis no conjunto de dados. Aliado às variáveis demográficas e os rótulos dos grupos, podemos delinear diferentes perfis para cada grupo. Ainda mais, examinando as sequências de transações de cada grupo podemos identificar de que forma os clientes bancários estão a gerir o seu dinheiro.

Estas 3 vertentes da metodologia, permitem uma análise de grupos de um conjunto de dados por explorar, uma caracterização detalhada de cada grupo recorrendo a descoberta de padrões e, por último, um algoritmo de classificação que permite acompanhar a jornada de evolução de um cliente, ao longo do tempo.

Capítulo 3

Dados

Para desenvolver a metodologia de identificação e caracterização de perfis de clientes proposta foram utilizados dados bancários *open-source* da República Checa (66). Neste capítulo descrevemos a fonte de dados, as etapas do pré-processamento e a análise exploratória dos dados.

3.1 Descrição dos dados

Os dados foram disponibilizados no âmbito de um desafio, *Discovery Challenge*, no contexto da 3ª Conferência Europeia sobre Princípios e Práticas de Descoberta de Conhecimento em Bases de Dados (conhecido por PKDD'99). Dois conjuntos de dados com dados reais foram disponibilizados para este fim, sendo que trabalhámos com o relativo ao domínio financeiro, também conhecido por *Berka Dataset* (66).

Tabela 3.1: Descrições das tabelas utilizadas do conjunto de dados *Berka*.

Nome da Tabela	Descrição
CONTAS	Cada registo descreve as características estáticas de uma conta bancária, num total de 4500 contas.
CLIENTES	Cada registo descreve as características de um cliente, num total de 5369 clientes.
DISPOSIÇÃO	Cada registo relaciona um cliente com uma conta, ou seja, essa relação descreve os direitos dos clientes para utilizar as contas, podendo este ser primário ou secundário.
TRANSAÇÕES	Cada registo descreve as características de uma transação, com 1.056.320 de registos.
TRANSAÇÕES PERMANENTES	Cada registo descreve as características de uma ordem de pagamento.
EMPRÉSTIMOS	Cada registo descreve as características de um empréstimo aprovado para uma dada conta, num total de 682 empréstimos.
CARTÕES DE CRÉDITO	Cada registo descreve o cartão de crédito atribuído a cada conta.
DISTRITOS	Cada registo descreve algumas variáveis demográficas dos respetivos distritos.

Dado o objetivo deste trabalho, i.e., a segmentação de clientes no contexto da banca com base em dados de transações, o *Berka Dataset* foi o mais adequado dos disponíveis à data publicamente (outras opções incluíram (67; 68; 69; 70)). Este conjunto de dados contém transações de clientes reais num período de tempo de 5 anos, longo o suficiente para encontrar padrões significativos. O conjunto de dados é composto por cerca de 5.369 clientes bancários com aproximadamente 1M de transações no total. Além disso, o banco representado no conjunto de dados concedeu cerca de 700 empréstimos e emitiu quase 900 cartões de crédito, todos representados nos dados. Descrevemos cada uma na tabela 3.1.

O Modelo Entidade-Relacionamento é resumido na fig. 3.1, onde se destacam as ligações entre tabelas e as chaves primárias.

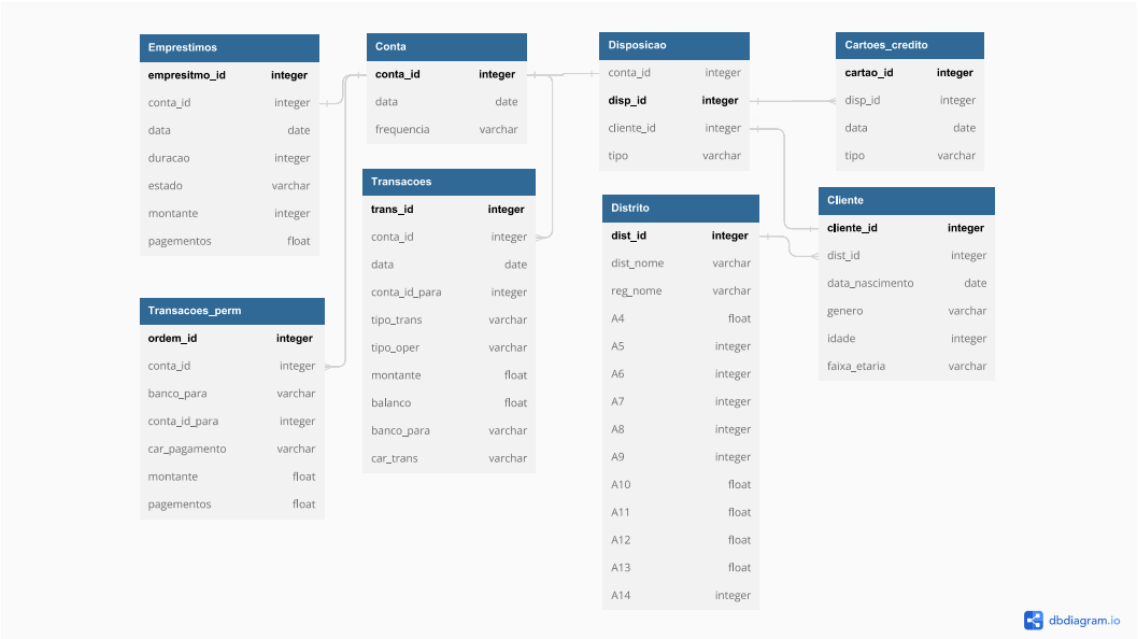

Figura 3.1: Diagrama Entidade-Relacionamento que apresenta o esquema dos dados da República Checa com as tabelas e respetivas conexões obtido com a aplicação dbdiagram.io (1).

De seguida, apresentamos alguns aspetos que ajudam a compreender as relações entre as tabelas:

- Cada conta tem características estáticas (por exemplo, data de criação) dadas em relação à tabela CONTA e características dinâmicas (por exemplo, pagamentos debitados ou creditados, saldos) dadas nas tabela TRANSAÇÕES PERMANENTES e TRANSAÇÕES;
- A tabela CLIENTES descreve características de pessoas que podem manipular as contas;
- Um cliente pode ter mais que uma conta, clientes e contas estão relacionados entre si na tabela DISPOSIÇÃO;
- As tabelas EMPRÉSTIMOS e CARTÕES DE CRÉDITO descrevem alguns serviços que o banco oferece aos seus clientes;

- Mais de um cartão de crédito pode ser emitido para uma conta;

- No máximo um empréstimo pode ser concedido por conta.

Em seguida resumimos, para cada tabela, as variáveis existentes e os seus tipos. Esta análise é importante para tomar decisões em relação ao pré-processamento tanto na etapa de limpeza de dados, como também, para a preparação de dados de treino nos modelos de AA. Além disso, aproveitamos para esclarecer alguns aspetos, após cada tabela, das respetivas variáveis.

Tabela 3.2: *Schema* da tabela CONTAS.

Colunas	Descrição	Tipo	Exemplo
`Conta_id`	Número de identificação da conta.	Inteiro	576
`Dist_id`	Número de identificação do distrito do sucursal.	Inteiro	576
`Data`	Data em que a conta foi criada.	Inteiro	930101
`Frequencia`	Frequência de emissão de extratos.	Texto	Emissão Mensal

Na tabela 3.2 apresentamos a informação disponível em relação às contas bancárias dos clientes, temos acesso à data em que a conta foi criada no banco bem como a frequência com que o cliente deseja que os seus extratos bancários sejam emitidos. A identificação do distrito diz respeito à sucursal no qual se procedeu a criação da conta, dado na altura não ser possível fazer online.

Tabela 3.3: *Schema* da tabela CLIENTES.

Colunas	Descrição	Tipo	Exemplo
`Cliente_id`	Número de identificação do cliente	Inteiro	1
`Dist_id`	Número de identificação do distrito do sucursal.	Inteiro	576
`Data_nascimento`	Data de nascimento do cliente (ano, dia, mês).	Inteiro	701213
`Genero`	Género do cliente.	Texto	Feminino
`Idade`	Idade do cliente.	Inteiro	29
`Faixa_etaria`	Faixa etária do cliente.	Texto	Adulto

Na tabela 3.3 é apresentada toda a informação relativa ao cliente, nomeadamente variáveis demográficas importantes para a segmentação como a sua data de nascimento, idade, género e distrito. A faixa etária foi criada com base nos seguintes critérios: Jovem (0-24 anos), Adulto (24-35 anos), Meia-idade (36-64 anos) e Sénior ($\geq$ 65 anos).

Tabela 3.4: *Schema* da tabela DISPOSIÇÃO.

Colunas	Descrição	Tipo	Exemplo
Cliente_id	Número de identificação do cliente	Inteiro	1
Disp_id	Data de nascimento do cliente (ano, mês, dia).	Inteiro	1
Conta_id	Género do cliente.	Inteiro	1
Tipo	Tipo de disposição.	Texto	Primário

A tabela 3.4 serve para conectar o cliente à respetiva conta, dado que vários clientes podem estar associados à mesma conta, sendo a distinção feita em termos da disposição (i.e., se é o dono principal da conta ou secundário). O que distingue os dois são as restrições no uso da conta, onde o primeiro tem mais privilégios.

Tabela 3.5: *Schema* da tabela EMPRÉSTIMOS.

Colunas	Descrição	Tipo	Exemplo
Emprestimo_id	Número de identificação do empréstimo.	Inteiro	5314
Conta_id	Número de identificação da conta.	Inteiro	1787
Data	Data em que foi concedido o empréstimo.	Inteiro	930705
Montante	Montante de dinheiro do empréstimo.	Inteiro	96396
Duração	Duração em semanas do empréstimo.	Inteiro	12
Pagamentos	Pagamentos mensais do empréstimo.	Decimal	8033.0
Estado	Estado em relação ao pagamento do empréstimo.	Texto	B

Na tabela 3.5 é possível consultar a informação relativa aos empréstimos dos clientes, como, por exemplo, a data em que foi atribuído, o montante, o número de prestações e o valor a pagar. Temos também acesso ao estado em que se encontra o cliente relativamente ao crédito, que pode tomar 4 valores possíveis: **A**- Contrato terminado sem problemas, **B** - Contrato terminado, o cliente está em incumprimento, **C**- Contrato ativo, sem problemas e **D** - Contrato ativo, o cliente está em incumprimento. Este atributo é interessante para a segmentação, dado ser fundamental para os bancos identificarem os grupos de clientes que podem ser mais problemáticos.

Tabela 3.6: *Schema* da tabela TRANSAÇÕES PERMANENTES.

Colunas	Descrição	Tipo	Exemplo
Ordem_id	Número de identificação da ordem.	Inteiro	29401
Conta_id	Número de identificação da conta que faz a ordem.	Inteiro	1
Banco_para	Banco recipiente da transação.	Texto	YZ
Conta_id_para	Número de identificação da conta recipiente da transação.	Inteiro	87144583
Montante	Montante debitado.	Decimal	2452.0
Car_pagamento	Caracterização do pagamento.	Texto	Pagamento doméstico

A tabela 3.6 apresenta informações relativas às a transações permanentes de cada conta, no-

meadamente, a informação dos recipientes, como o banco e o identificador da conta. Além disso, possui também um descritivo do pagamento efetuado, o que permite obter algum contexto em relação à natureza do pagamento.

Tabela 3.7: *Schema* da tabela TRANSAÇÕES.

Colunas	Descrição	Tipo	Exemplo
Trans_id	Número de identificação da transação.	Inteiro	695247
Conta_id	Número de identificação da conta que fez a transação.	Inteiro	2378
Data	Data em que foi efetuada a transação (ano, mês, dia).	Inteiro	930101
Tipo	Tipo de transação.	Texto	Crédito
Operacao	Tipo de operação.	Texto	Levantamento em dinheiro
Montante	Montante da transação.	Decimal	700.0
Balanco	Balanço da conta após a transação.	Decimal	700.0
Car_trans	Caracterização da transação.	Texto	Juros creditados
Banco_para	Banco da conta recipiente.	Texto	QR
Conta_id_para	Número de identificação da conta recipiente.	Inteiro	99994199

A tabela 3.7 apresenta as transações efetuadas por cada conta, bem como as suas características, destacando a data em que ocorrem, o tipo de transação (de crédito ou débito), o montante e o balanço respetivo da conta após a transação. De destacar, que a cada transação está associada uma única conta, não sendo possível distinguir qual dos clientes associado a essa conta efetuou uma dada transação.

Tabela 3.8: *Schema* da tabela CARTÕES DE CRÉDITO.

Colunas	Descrição	Tipo	Exemplo
Cartao_id	Número de identificação do cartão.	Inteiro	1005
Disp_id	Disposição da conta a que o cartão foi atribuído.	Inteiro	9285
Data	Data em que foi atribuído o cartão (ano, mês, dia).	Inteiro	931107
Tipo	Tipo do cartão.	Texto	Clássico

A tabela 3.8 tem informação relativa aos cartões de crédito, sendo possível saber a que cliente da conta se encontra associado o cartão. O tipo de cartão pode tomar 3 valores possíveis: Júnior, Clássico e *Gold*.

Tabela 3.9: *Schema* da tabela DISTRITO.

Colunas	Descrição	Tipo	Exemplo
Dist_nome	Nome do distrito.	Texto	Praha
Reg_nome	Nome da região.	Texto	Prague
A4	Nº de habitantes.	Inteiro	1204953
A5	Nº de Munícipios com menos de 499 habitantes.	Inteiro	0
A6	Nº de Munícipios com 500-1999 habitantes.	Inteiro	0
A7	Nº de Munícipios com 2000-9999 habitantes.	Inteiro	0
A8	Nº de Munícipios com mais 10.000 habitantes.	Inteiro	1
A9	Nº de Cidades.	Inteiro	1
A10	Ratio de habitantes nas cidades.	Decimal	100
A11	Salário Médio.	Decimal	12541.0
A12	Taxa de desemprego em 1995.	Decimal	0.29
A13	Taxa de desemprego em 1996.	Decimal	0.43
A14	Nº de empresários por 1000 habitantes.	Inteiro	167

A tabela 3.9 contém para além do nome dos distritos e regiões da República Checa, um conjunto de variáveis demográficas que nos podem ajudar a interpretar os clientes em cada distrito e os grupos. A tabela 3.6 está contida na tabela 3.7, pelo que não será usada no trabalho.

3.2 Pré-processamento

Nesta secção, apresentamos o pré-processamento efetuado ao conjunto de dados utilizado. Esta etapa inclui o tratamento de valores em falta, deteção de anomalias, seleção de colunas e criação de novas variáveis.

Originalmente, algumas colunas possuíam os valores na língua materna do conjunto de dados pelo que foi necessário substituir pela sua respetiva tradução, existente no dicionário de dados em anexo (ver anexo A.1), para uma interpretação mais fácil.

Em seguida, verificou-se quais as variáveis que continham mais valores em falta, e com base nisso retiraram-se as seguintes colunas da análise: Car_trans com 50% (tabela TRANSAÇÕES) e Operação com 17% (tabela TRANSAÇÕES). As datas encontram-se com o formato YYMMDD pelo que foi necessário reconverter para um um objeto *datetime*.

A deteção de *outliers* foi realizada com o método *z-score*, calculado através da eq. 3.1, com μ e σ a corresponder à média e ao desvio padrão da variável e x à observação, respetivamente. Esta é uma medida estatística que quantifica quantos desvios-padrão uma observação está afastada da média.

$$Z_{scores} = \frac{(x - \mu)}{\sigma} \tag{3.1}$$

Escolheu-se esta técnica dada a sua simplicidade e base estatística, recorrendo ao desvio padrão. Apesar de o método escolhido poder ser influenciado pelos valores extremos e assumir a existência de uma distribuição normal, permite manter o intervalo dos valores. A alternativa

seria utilizar um método baseado em percentis, que apesar de ser mais robusto, limita bastante o intervalo dos dados, o que não reflete realisticamente os clientes.

Definiu-se um *theshold* de 3σ, o que significa que qualquer observação fora do intervalo $[-3\sigma, 3\sigma]$ é considerado um *outlier*. Com isto, fez-se a análise para a tabela TRANSAÇÕES nas variáveis montante e balanço, sendo que se retirou, respetivamente, 24.466 e 7547 observações, descendo o total de transações disponíveis para 953.965.

Por último, procedeu-se à criação de novas variáveis:

- `montante_sinal`: junta as variáveis do montante e tipo de transação referentes à tabela TRANSAÇÕES, acrescentando um sinal + ou - consoante a transação é crédito ou débito, respetivamente;
- `racio_montante_balanço`: rácio entre o montante da transação e o balanço da conta, da tabela TRANSAÇÕES, de forma a perceber a dimensão do gasto/crédito em relação ao valor da conta do cliente. Será necessário aplicar um *smoothing* (adição por um valor infinitesimal), para o caso em que o balanço toma o valor 0 no divisor;
- `flag_conta_conjunta`: variável binária que identifica se a conta é conjunta ou não;
- `antiguidade_cliente`: utilizando a variável `Data` da tabela CONTA, é possível perceber há quanto tempo o cliente está inserido no banco através de uma variável categórica. Esta possui a quantidade de anos em que o cliente está inserido no banco tomando os valores de 0 a 5;
- `zona_geográfica`: existem 77 distritos distintos, sendo uma cardinalidade bastante elevada para os modelos e analisar. Foi assim necessário reduzir este número através de uma nova variável que agrupa todos os distritos em zonas geográficas de acordo com a sua proximidade geográfica, sendo esta nova variável mais compreensiva que a variável região (toma 5 valores), mas menos exaustiva que a variável distrito (toma 77 valores). A variável `zona_geográfica` tem os seguintes valores possíveis: Prague Metropolitan Area, South Bohemia, West Bohemia, North Bohemia, East Bohemia, South Moravia, Vysočina, Zlin Region e North Moravia.

3.3 Análise exploratória

Nesta secção, apresentamos a análise exploratória que realizámos ao conjunto de dados com recurso a visualizações, como histogramas, gráficos de barra, gráficos de pizza, entre outras. Avaliámos a distribuição de clientes, contas, transações e as variáveis demográficas disponíveis para os distritos.

Como variam os perfis das contas conjuntas?

O primeiro problema com que nos deparámos foi o facto de poder existir mais do que um cliente por conta. De facto, só sabemos qual a conta que efetuou uma transação e não qual o cliente em

específico associado a essa conta é que realizou essa transação. Para podermos proceder com a nossa análise, assumimos que o cliente que tem a disposição de dono é quem efetua as transações e por isso associamos as suas variáveis demográficas à conta, ficando um total de 4500 clientes em vez dos 5369 clientes iniciais. No entanto, é interessante analisar a distribuição em relação às variáveis demográficas destes pares de clientes com conta conjunta. Em grande parte das contas conjuntas, está associado um homem e mulher com a mesma faixa etária, podendo ser um casal. Mais em específico, as 3 combinações que surgem mais vezes são:

- {utilizador principal, mulher, meia idade} com {utilizador secundário, homem, meia-idade}, em 23% das contas conjuntas;
- {utilizador principal, homem, meia-idade} com {utilizador secundário, mulher, meia-idade}, em 20% das contas conjuntas;
- {utilizador principal, homem, sénior} com {utilizador secundário, mulher, sénior}, em 12% das contas conjuntas.

Como variam as variáveis demográficas dos clientes bancários? Existe balanço em relação às transações de crédito/débito?

De seguida foi importante perceber as distribuições das variáveis demográficas por forma a analisar se existe algum tipo de enviesamento na amostra de clientes (ver fig. 3.2).

A distribuição de género encontra-se bastante equilibrada. A faixa etária tem uma predominância de pessoas na meia idade. No que toca a geografia, as regiões relacionadas com Moravia são as mais presentes, no entanto Praga apesar de mais pequena em dimensão espacial tem a maior presença dado ser onde fica a capital da República Checa (com maior densidade populacional).

No gráfico (D) percebemos que a proporção entre crédito e débito nos tipos de transação não é equilibrada, com uma diferença de 13%.

Será que a faixa etária do cliente influencia os seus rendimentos em termos do balanço da conta e do número de transações?

Como podemos observar na fig. 3.3, à medida que a idade aumenta existem mais pessoas com balanço da conta altos, sendo que são também estes os que têm dívidas, como se verifica pelos valores negativos. Na faixa etária sénior sénior, a tendência dos rendimentos aumentarem inverte-se, diminuindo o balanço da conta e o montante das transações, na generalidade. Dado cerca de 50% da amostra ser de clientes com média idade, verifica-se que estes também dominam o histograma (faixa laranja).

Será que o género do cliente influencia os seus rendimentos em termos do balanço da conta e do número de transações?

Podemos observar na fig. 3.4, que o género dos clientes não parece apresentar uma discrepância em relação a valores monetários.

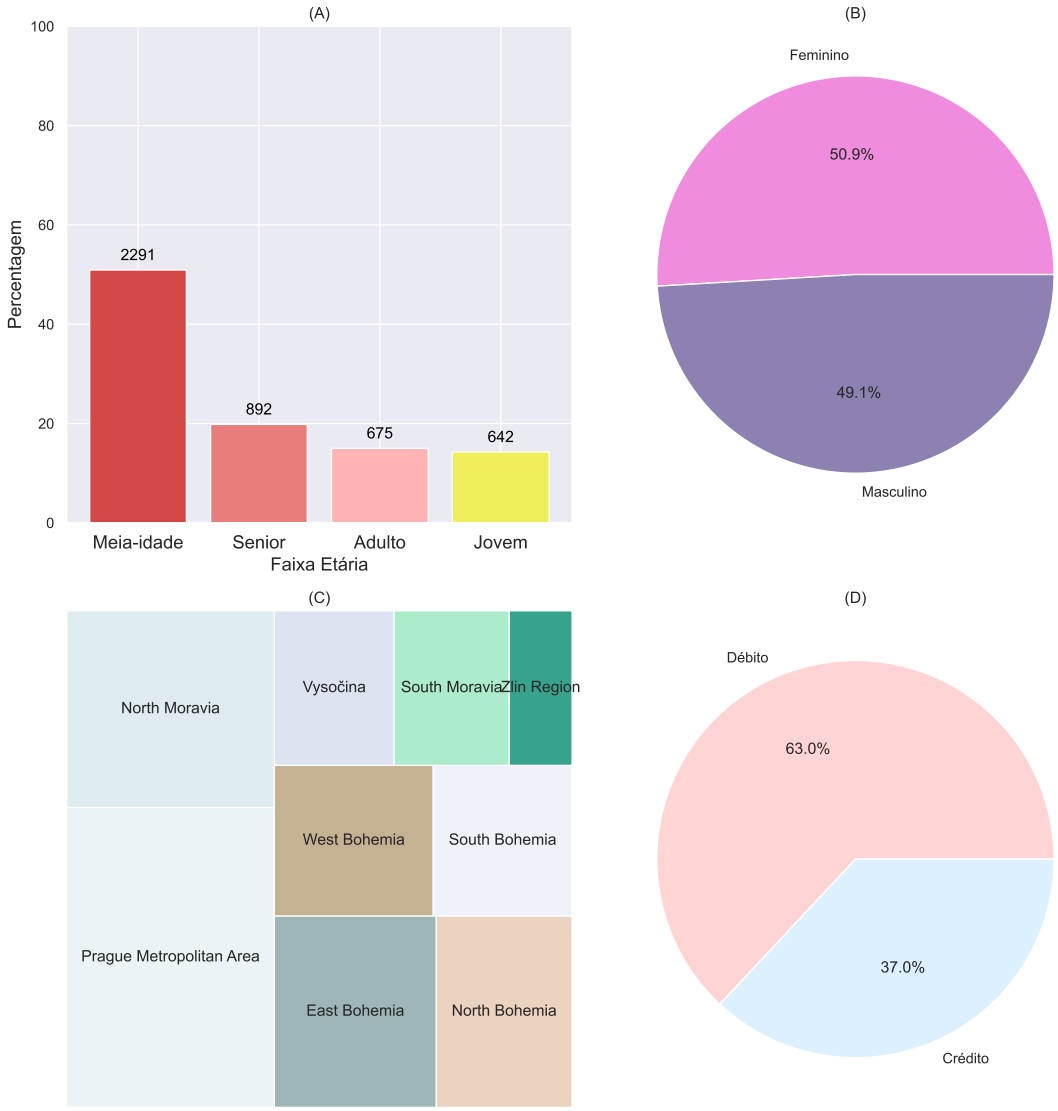

Figura 3.2: Distribuição de clientes em relação às seguintes variáveis: faixa etária (A), género (B), região (C) e tipo de transação (D).

Como variam o número de transações efetuadas pelos clientes bancários? E na dimensão temporal?

Foi importante compreender quantas transações em média existem por conta e se existe alguma sazonalidade na série temporal das transações. Verificamos no gráfico do lado esquerdo da fig. 3.5 que existem dois picos por volta das 100 a 400 transações, no espaço de 5 anos, e na mesma fig. do lado direito verifica-se que a série temporal tem de facto sazonalidade anual (no final/início de ano), e mais uma pequena sazonalidade de caráter semestral, com uma tendência geral de aumentar com o tempo.

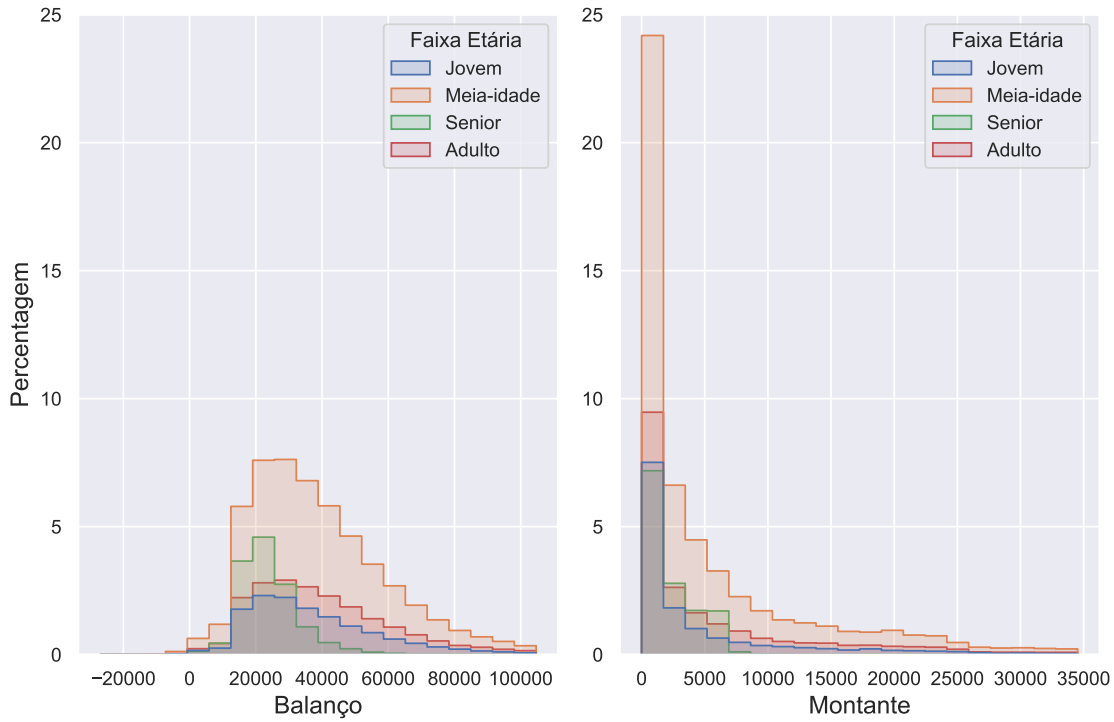

Figura 3.3: Histograma das variáveis montante de transação e balanço da conta bancária por faixa etária. A unidade da moeda é a coroa checa, cujo símbolo é Kč.

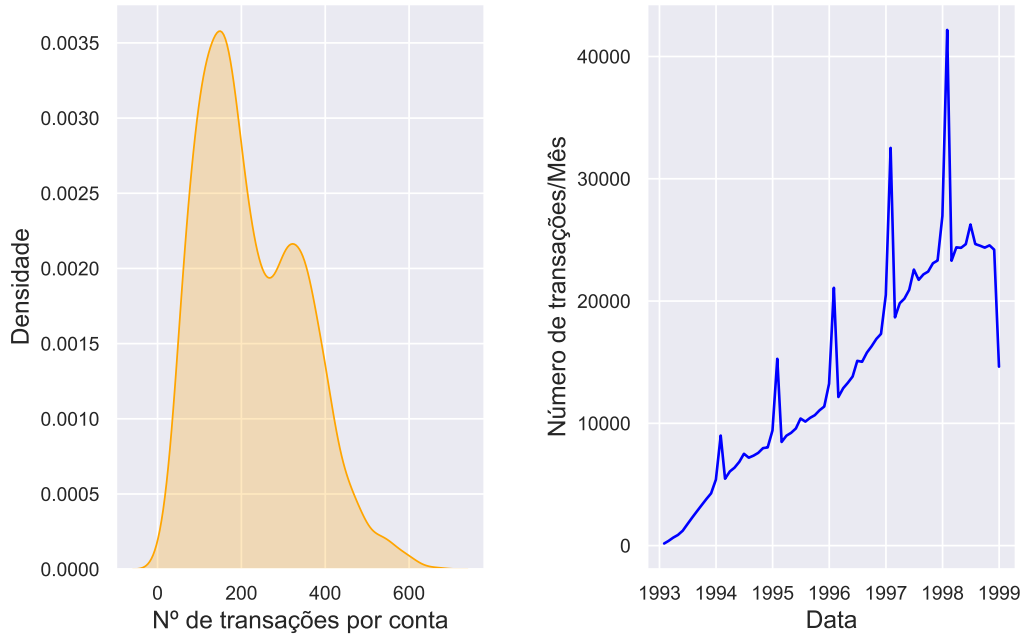

Figura 3.5: No lado esquerdo apresentamos um gráfico da densidade do nº de transações por conta, e no lado direito, a série temporal do nº de transações agregadas ao mês.

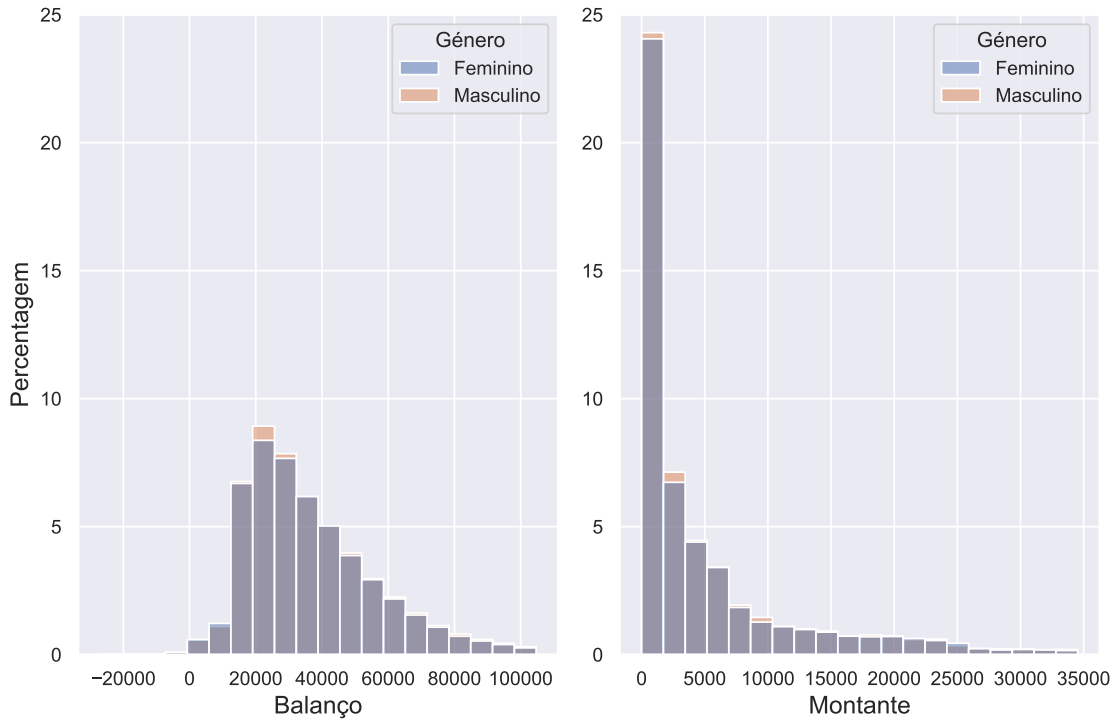

Figura 3.4: Histograma das variáveis montante de transação e balanço da conta bancário por género. A unidade da moeda é a coroa checa, cujo símbolo é Kč.

Como se resumem as variáveis numéricas, relativas às transações e empréstimos, utilizando estatística descritiva?

Observa-se com base na tabela 3.10 que as médias e medianas são bastante diferentes indicando que existe enviesamento na distribuição. Usámos dois testes estatísticos para estudar a normalidade: o Teste de *Shapiro-Wilk* e o Teste de *Kolmogorov-Smirnov*, com *alpha* de 5%. O teste de *Shapiro-Wilk* calcula uma estatística de teste com base nas diferenças entre os valores observados e os valores esperados numa distribuição normal. O teste de *Kolmogorov-Smirnov* compara a distribuição empírica dos dados com a distribuição teórica (normal neste caso) e calcula a maior discrepância entre elas (estatística de teste). Como em ambos os testes a hipótese nula, de normalidade, foi rejeitada em todas as variáveis, podemos afirmar que existe evidência estatística que as distribuições não são normais.

Comparando os valores máximos com as médias observa-se também a existência de valores extremos (*outliers*). A detecção de valores extremos em variáveis numéricas é importante para identificar pontos de dados que possam distorcer os resultados do modelo e afetar negativamente a precisão das previsões. Além disso, a verificação da normalidade das variáveis é essencial, pois muitos algoritmos de AA assumem a distribuição normal dos dados para funcionarem corretamente.

Tabela 3.10: Tabela com a estatística descritiva de algumas variáveis numéricas. A unidade da moeda é a coroa checa, cujo símbolo é Kč.

Tabela	Coluna	Min.	Máx.	Mediana	Média	Desvio Padrão
Transações	Montante	0	3452	192	476	686
Transações	Balanço	-27019	104661	32241	36581	19425
Empréstimos	Montante	4980	590820	116928	151410	113372
Empréstimos	Pagamentos	304	9910	3934	4190	2215

Qual a distribuição das variáveis categóricas relativas ao número de amortizações e estado do empréstimo no conjunto de dados?

Em relação aos empréstimos, as amortizações podem tomar os valores de 12, 24, 36, 48 e 60, com proporções semelhantes no conjunto de dados. No que toca ao estado dos empréstimos, este tem a seguinte distribuição: C (59%), A (30%), D (7%) e B (5%).

Existem variáveis correlacionadas, positivamente ou negativamente?

Por último, foi calculada a correlação entre variáveis utilizando o coeficiente de *Pearson*, após a normalização destas e codificação com *one-hot encoding*, destancando-se as seguintes correlações:

- Forte correlação positiva entre as variáveis montante do empréstimo com duração e pagamentos;
- Forte correlação positiva com variáveis estado de empréstimo A e C com a maior duração do empréstimo;
- Correlação positiva entre o montante da transação e balanço da conta.

A avaliação da correlação entre variáveis em AA é fundamental e traz diversas vantages: permite reduzir a dimensionalidade do conjunto de dados, melhorar a interpretabilidade do modelo, evitar sobreajuste (*overfitting*) (71),lidar com a multicolinearidade, e otimizar o desempenho geral do modelo.

Como podemos resumir as zonas geográficas em função das suas informações demográficas?

Para o leitor melhor visualizar as diferentes regiões foram colocados em anexo dois mapas complementares da República Checa (ver anexo A.2).

Por fim, analisamos as zonas geográficas existentes nos dados de acordo com as variáveis demográficas disponibilizadas na tabela DISTRITO, resumidas na tabela 3.11.

Tabela 3.11: Regiões e características demográficas.

Região	Características demográficas
Boémia Oriental	População de tamanho médio, dispersa pelo território, salário médio de cerca de 8611 Kč, baixa taxa de desemprego de 2,98 e uma presença notável de empresários.
Boémia do Norte	População ligeiramente maior em relação à norma, dispersa pelo território, salário médio de aproximadamente 9334 Kč, taxa de desemprego moderada de 5,86, a mais alta.
Morávia do Norte	População maior do que à norma, maior concentração populacional, salário médio de cerca de 9063 Kč, taxa de desemprego moderada de 5,71 e a presença mais baixa de empresários.
Área Metropolitana de Praga	População bastante acima da norma, dispersa pelo território, salário médio mais alto de 9602 Kč, baixa taxa de desemprego de 2,69 e a mais alta proporção de empresários.
Boémia do Sul	População de tamanho médio, dispersa pelo território, salário médio de aproximadamente 8831 Kč, taxa de desemprego de 2,81.
Morávia do Sul	População maior do que à norma, maior concentração populacional, salário médio de cerca de 8874 Kč, taxa de desemprego moderada de 3,49.
Vysočina	População de tamanho médio, dispersa pelo território, salário médio de aproximadamente 8542 Kč, a mais baixa, taxa de desemprego de 4,02.
Boémia Ocidental	População de tamanho médio, salário médio de cerca de 9015 Kč, baixa taxa de desemprego de 2,65.
Região de Zlin	População de tamanho médio, com alguma dispersão no território e concentração com um número significativo de pequenos municípios, salário médio de aproximadamente 8978 Kč, taxa de desemprego moderada de 4,53 e a terceira presença mais alta de empresários.

A região central, com Praga no seu centro, experimentava um crescimento populacional mais rápido, impulsionado pela migração interna e pela concentração de oportunidades económicas. Essa região era marcada por uma população mais jovem e dinâmica, com um ambiente urbano desenvolvido e uma economia em crescimento. Por outro lado, as regiões do norte, sul e leste do país apresentavam uma dinâmica demográfica diferente, com declínio populacional em função do êxodo rural de jovens e da falta de oportunidades económicas. Além disso, essas áreas eram caracterizadas por uma população mais envelhecida e uma economia menos desenvolvida, dependente principalmente da agricultura e da indústria tradicional.

3.4 Discussão

O capítulo 3 foi dedicado à descrição, processamento e análise exploratória dos dados utilizados neste trabalho. Dado o objetivo de identificar perfis de clientes bancários foi necessário encontrar dados disponíveis online com informações de transações e dados dos clientes para implementar a metodologia proposta. Foi identificado um conjunto de dados, o *Berka Dataset* que continha 1M de transações, de 5369 clientes, como também informações relativas a 700 empréstimos, num período temporal que abrange 5 anos. Um dos principais problemas com que nos deparámos foi a transação estar associada a uma conta, o que torna difícil de perceber quem foi o cliente que efetuou essa transação, pois as contas podem ser conjuntas. Sendo assim tomou-se a decisão de avaliar somente os utilizadores principais considerando que foram eles que efetuaram as transações.

Em relação ao pré-processamento, traduziram-se as variáveis para português, retiraram-se as colunas com uma elevada percentagem de valores em falta, e retiraram-se transações que foram consideradas *outliers* com o método *z-score*, pois existem modelos de AA sensíveis a estes. Procedeu-se à criação de algumas variáveis novas, como o `flag_conta_conjunta`, para perceber quais contas são partilhadas, entre outras.

A análise exploratória serviu para compreender melhor os dados existentes. Por exemplo, percebemos que a amostra de clientes é enviesada para a faixa etária de meia-idade, existe um equilíbrio entre os géneros e existe também uma maior proporção de indíviduos na cidade capital de Praga. Observou-se através de histogramas que a faixa etária influencia bastante as variáveis financeiras como o balanço da conta e os montantes das transações, já em relação ao género não verificamos a mesma influência. O número de transações médias por conta varia entre circa de 150 e 400, na maioria, o que nos permite obter uma imagem clara dos comportamentos dos clientes. Em relação aos créditos existe um elevado desvio padrão no montante emprestado, indicando que estes podem ser utilizados para os mais variados fins e que cerca de 89% dos créditos estão a ser cumpridos.

Com esta análise inicial dos dados e pré-processamento podemos começar a aplicar os métodos propostos com uma perspetiva mais clara e precisa das características dos dados.

Capítulo 4

Métodos e resultados

Este capítulo tem como objetivo sistematizar os métodos aplicados para efetuar a segmentação de clientes e caracterização dos respetivos grupos. Iniciamos com a análise RFM, uma técnica que utiliza três principais indicadores: Recência, Frequência e Valor Monetário, para identificar diferentes segmentos de clientes com base no seu comportamento. Em seguida, abordamos os algoritmos de agrupamento, uma técnica que permite identificar grupos semelhantes de clientes com base nas suas características e comportamentos.

Além da segmentação, também exploramos métodos para descoberta de padrões para criar perfis e caracterizar mais profundamente cada grupo encontrado na etapa de agrupamento. Começamos com a construção de *itemsets*, onde identificamos variáveis frequentemente encontradas em conjunto em cada grupo. Em seguida, utilizamos o *sequence mining*, uma técnica de PD que permite identificar padrões sequenciais de comportamento dos clientes, revelando tendências e preferências ao longo do tempo.

Posteriormente, concentramo-nos na classificação de clientes, que envolve a atribuição de clientes a grupos específicos com base nas suas características e comportamentos. Analisamos também a explicabilidade do modelo para obter uma melhor compreensão de quais as variáveis preditoras mais importantes.

Em resumo, este capítulo apresenta uma visão geral dos métodos e técnicas utilizadas para a tarefa de descoberta de conhecimento em dados de clientes bancários e os respetivos resultados.

4.1 Segmentação de clientes

A segmentação de clientes é uma abordagem que utiliza técnicas estatísticas, algoritmos de AA e PD para identificar grupos de clientes com características semelhantes. Esta abordagem baseia-se em dados e análises estatísticas para compreender e agrupar os clientes de forma mais precisa e eficiente. As principais vantagens de utilizar métodos orientados a dados na segmentação de clientes em comparação com os métodos tradicionais (como campanhas e comunicações generalizadas) são:

- **Precisão**: permitem uma segmentação mais precisa e pessoal, tendo em consideração múltiplas variáveis que permitem identificar padrões complexos que podem não ser perceptíveis

por meio de outras abordagens;

- **Eficiência**: Ao utilizar técnicas automatizadas de análise de dados, é possível processar grandes volumes de informações de forma rápida e eficiente, economizando tempo e recursos na segmentação de clientes.

Alguns dos métodos utilizados com mais frequência na literatura nesta etapa são a análise RFM e algoritmos de agrupamento, como o *k-means* ou *DBSCAN*.

4.1.1 Análise *Recency-Frequency-Monetary* (RFM)

A análise RFM é uma técnica amplamente utilizada na segmentação de clientes, que se baseia em três principais componentes para avaliar o valor e comportamento de um cliente num determinado período de tempo: a recência das compras, a frequência destas e o seu valor monetário. Neste caso, estando a lidar com dados bancários, consideramos em vez das compras, a análise baseada em transações. De seguida, apresentamos sumariamente as suas definições:

- **Recência** (*Recency*): Refere-se ao tempo decorrido desde a última compra de um cliente. Geralmente, quanto mais recente foi a compra, maior é o compromisso do cliente.

- **Frequência** (*Frequency*): Refere-se ao número de compras realizadas por um cliente num determinado período de tempo. Clientes com maior frequência de compras tendem a ser mais valiosos.

- **Monetário** (*Monetary*): Refere-se ao valor monetário total gasto por um cliente nas suas compras. Clientes que gastam mais dinheiro são considerados mais valiosos para o negócio.

A análise RFM pode ser utilizada para classificar os clientes em segmentos com base nestas três componentes. Além disso, é uma abordagem eficaz, pois é simples de entender e aplicar, oferecendo *insights* imediatos sobre o valor e comportamento dos clientes, ao contrário de algoritmos de agrupamento que necessitam de uma análise mais detalhada. Permite ainda personalizar a análise com base no que as empresas valorizam em relação a cada variável.

Para calcular estes valores utilizámos os seguintes métodos. Para a recência, atribuímos como valor a diferença de dias entre a última data presente no dataset (1998-12-12) e a última data da transação do cliente. Para a frequência, agrupámos as transações por clientes e agregámos o valor com uma operação de contagem, representado o número de ocorrência de transações em cada conta. Por último, o valor monetário é a soma do montante de todas as transações dos clientes, onde se recorreu à variável `montante_sinal` que tem em conta os débitos *versus* créditos dos clientes, dado que queremos priorizar os fluxos de dinheiro nas suas contas. Escolhemos ainda o período temporal de 5 anos para a análise, ou seja, para criar as três variáveis utilizamos todas as transações dos clientes. A distribuição das 3 métricas podem ser visualizadas através do diagrama de bigodes e histogramas apresentados na fig. 4.1.

Pela fig. 4.1 observamos que em relação à recência, gráfico (A), a maioria dos clientes são ativos tendo valores perto do valor 0 (o mais recente), existindo ainda assim bastantes *outliers*. Em relação à frequência, gráfico (B), vemos a distribuição do número de transações dos clientes com dois picos distintos em, aproximadamente, 150 e 350, como verificado anteriormente dado estarmos a considerar o período de cálculo dos valores RFM igual ao histórico total de transações, os 5 anos. Por último, o histograma (C) apresenta maioritariamente um equilíbrio no fluxo de dinheiro, com alguma dispersão para valores negativos, que significa que os clientes gastam mais em débito do que recebem em crédito na generalidade.

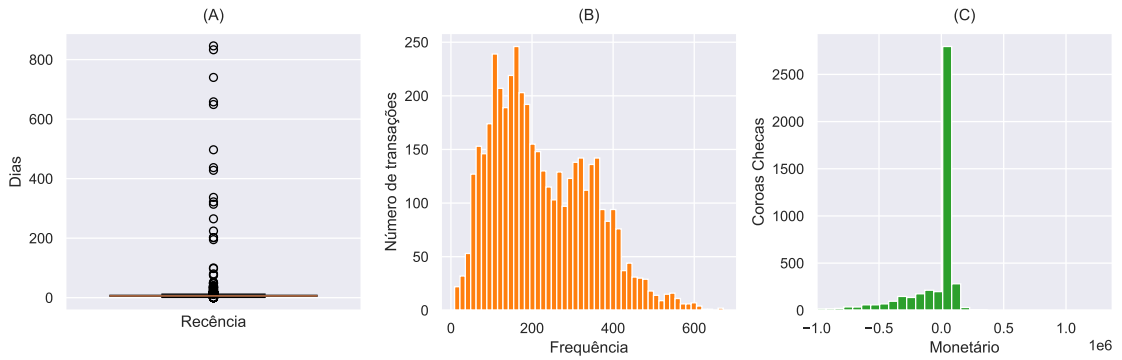

Figura 4.1: Distribuição do comportamento do cliente em relação a três métricas-chave: Recência (A), Frequência (B) e Valor Monetário (C). O número de *bins* para os gráficos (B) e (C) foi 50.

Em seguida, queremos segmentar os clientes por forma a criar uma nova variável que reflita o seu valor e comportamento geral num só *score*. Existem duas formas de efetuar este cálculo: através do uso da média, ou através da concatenação de um valor normalizado. A primeira utiliza um método de normalização como *ranking* ou percentis e efetua a média ponderada dos valores. A última utiliza uma normalização de 0 a 5 para cada variável, e concatena numa *string* as três, por exemplo, "555", é um cliente com o valor máximo em todas as variáveis RFM ([72]). A média ponderada é uma abordagem simples e intuitiva, vantajosa pela sua simplicidade e interpretabilidade, é também flexível e adequada para os cenários em que é necessário considerar as características específicas de cada componente e ajustar a análise de acordo com as necessidades estratégicas da empresa. Por outro lado, a concatenação mantém a granularidade dos três componentes RFM, permitindo uma segmentação mais detalhada, no entanto muito mais complexa dado o número de possíveis combinações. Dado queremos manter uma metodologia geral e personalizável a qualquer necessidade, vamos usar a 1ª abordagem. No entanto seria interessante como trabalho futuro, utilizar a 2ª abordagem para estudar casos mais detalhados.

Foi necessário normalizar as 3 variáveis que possuem ordens de grandeza diferentes. Para tal utilizou-se a técnica de *ranking* de clientes, onde se ordena os mesmos de acordo com o seu *ranking* e normaliza-se pelo valor máximo dessa variável. A partir daqui utilizou-se a seguinte fórmula para se obter o *score*:

$$RFM_{score} = \omega_R \times Recencia_{rank} + \omega_F \times Frequencia_{rank} + \omega_M \times Monetario_{rank} \quad (4.1)$$

Os pesos ω_i são os respetivos pesos de cada variável que podem ser ajustados de acordo com a sua importância, para análise.

Utilizámos os seguintes pesos: $\omega_R = 0.15$, $\omega_F = 0.35$, $\omega_M = 0.50$. Estes são normalmente atribuídos com base no conhecimento do setor ou a experiência no domínio. O importante é garantir que os pesos atribuídos estejam alinhados com as metas e prioridades do negócio. Neste caso, vamos atribuir aproximadamente os valores para o exemplo de um banco cujo foco seja a fidelização dos clientes e otimizar o lucro obtido com essa relação, o caso mais comum e geral. Assim, a recência, com um peso de 0.15, é considerada importante, mas não dominante, indicando a valorização de clientes que realizaram transações recentes. A frequência, com um peso de 0.35, é altamente valorizada, sugerindo que o banco recompensa clientes que transacionam com frequência para aumentar a fidelidade. O componente monetário, com peso de 0.50, é significativamente valorizado, indicando um foco na maximização do valor dos clientes, especialmente aqueles que realizam transações de alto valor.

Tendo este *score*, que se encontra normalizado no intervalo [0,10], podemos segmentar os clientes de acordo com a seguinte regra para enriquecer mais o conjunto de dados:

$$Segmento = \begin{cases} \texttt{Clientes de topo}: \text{RFM}_{\text{score}} \in]9, 10] \\ \texttt{Clientes de alto valor}: \text{RFM}_{\text{score}} \in]7, 9] \\ \texttt{Clientes de alto valor médio}: \text{RFM}_{\text{score}} \in]5, 7] \\ \texttt{Clientes de alto baixo}: \text{RFM}_{\text{score}} \in]3, 5] \\ \texttt{Clientes perdidos}: \text{RFM}_{\text{score}} \in [0, 3] \end{cases} \quad (4.2)$$

Em seguida descrevemos mais detalhadamente as características de cada segmento:

- **Clientes de topo** (*Top Customers*): Estes são os clientes mais valiosos para o negócio, tendo um `RFM_score` entre 9 e 10, o que significa que eles têm a maior pontuação possível. Estes clientes fazem transações com frequência, gastam muito dinheiro e fizeram transações recentemente.

- **Clientes de alto valor** (*High-Value Customers*): Este grupo inclui clientes que têm um `RFM_score` entre 7 e 9. Ainda são muito valiosos, embora não estejam no topo, mas acima da norma.

- **Clientes de médio valor** (*Medium Customers*): Estes clientes têm um `RFM_score` entre 5 e 7. São considerados valiosos, mas talvez não transacionam com tanta frequência quanto os grupos anteriores ou não fazem transações com montantes elevados.

- **Clientes de baixo valor** (*Low-Value Customers*): Este segmento é composto por clientes com um `RFM_score` entre 3 e 5. Não são tão valiosos quanto os grupos anteriores, podem fazer transações menos frequentes e gastar menos em cada transação, ou ser mais ativos.

- **Clientes perdidos** (*Lost Customers*): Estes são os clientes menos valiosos ou inativos. Eles têm um `RFM_score` entre 0 e 3. Isso sugere que eles não fizeram transações recentes, não transacionam com frequência ou gastam muito pouco quando o fazem. Estes clientes são considerados em risco de serem perdidos.

Dado o período para o cálculo das variáveis RFM ser 5 anos, utilizando o histórico todo disponível, é previsível que um cliente que esteja há mais tempo do banco irá, à partida, ter valores de RFM melhores. Foi assim importante perceber como é que a duração da relação entre o cliente e o banco inluenciaram a escolha do segmento em que o cliente se encontra. Verificou-se que o segmento `Clientes perdidos` é na verdade composto na sua maioria por clientes recentes, com 2 ou menos anos de antiguidade. Assim subdividiu-se ainda mais esta variável de forma a classificar os clientes do segmento prévio de `Clientes perdidos` com 2 ou menos anos como `Novos clientes`.

A fig. 4.2 sumariza os resultados da análise RFM obtidos em função das variáveis `RFM_score` e `segmento`.

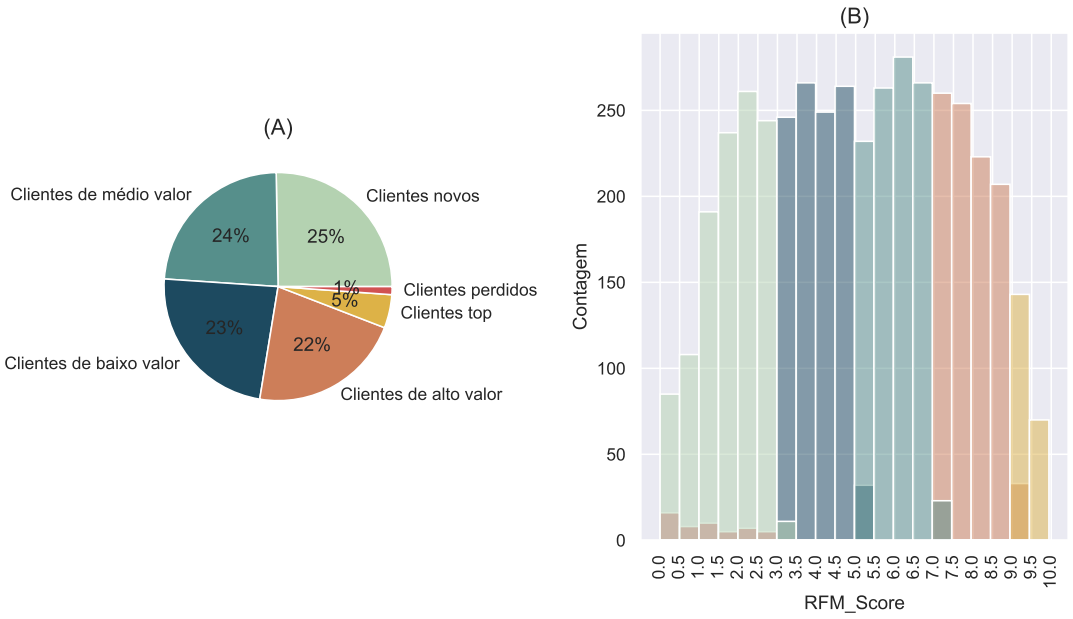

Figura 4.2: Distribuição do segmento (A) e histograma *score* dos clientes (B), obtido com recurso às equações 4.2 e 4.1, respetivamente.

A partir da fig. 4.2 (A), concluímos que os clientes no conjunto de dados se encontram distribuídos em proporções semelhantes nos diversos segmentos exceto no segmento `Clientes de topo` que contém menos clientes em comparação com os restantes (apenas 5%), e os `Clientes perdidos` que possuem uma parcela muito pequena de clientes (somente 1%). Como se observa na fig. 4.2 (B), mais especificamente comparando valores, dentro da classe `Clientes de baixo valor` e `Clientes de médio valor`, a distribuição é aproximadamente uniforme

no seu respetivo intervalo. Nas extremidades do `RFM_score`, em `Clientes perdidos` existem algumas pessoas com valores bastante baixos que aumentam gradualmente com o `RFM_score`, e ao contrário para a classe `Clientes de elevado valor` que desce gradualmente.

Na fig. 4.3 vemos como a antiguidade do cliente se relaciona com o segmento. Em relação ao `Cliente de valor médio` este tem uma distribuição praticamente uniforme ao longo dos anos, já `Cliente de topo` contém maioritariamente clientes que já estão há pelo menos 4 anos com o banco.

Após esta análise RFM e com as respetivas novas variáveis, inclusive o segmento do cliente, podemos proceder para análise com grupos, onde enriquecemos a informação dos clientes com outras variáveis interessantes de forma a encontrar grupos semelhantes.

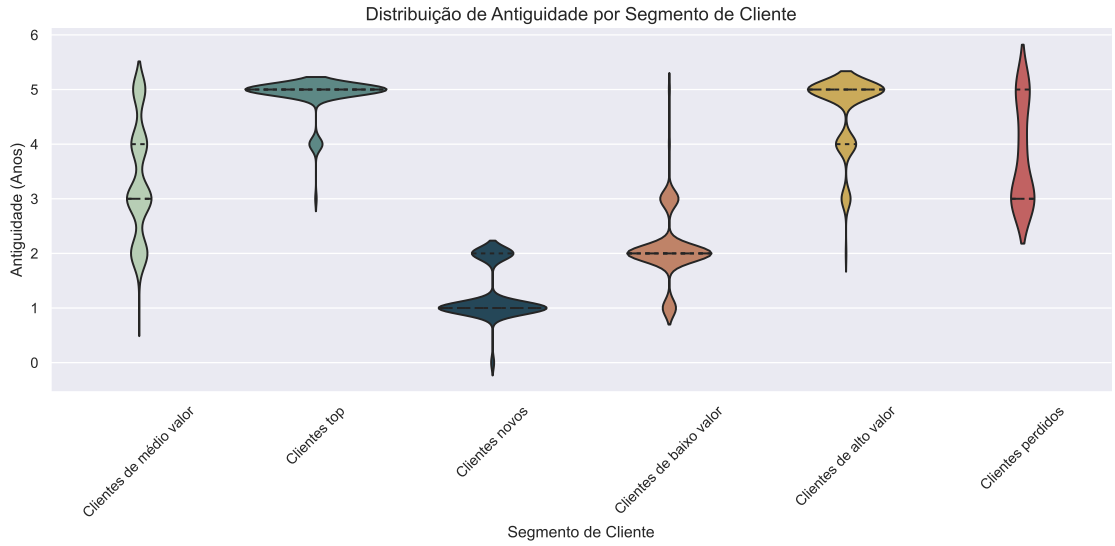

Figura 4.3: Distribuição da antiguidade do cliente no banco dentro de cada segmento.

4.1.2 Algoritmos de agrupamento

De seguida, apresentamos os resultados da aplicação de dois algoritmos de agrupamento, o *k-means*, da família dos algoritmos baseados em centróides, e *DBSCAN*, da família de algoritmos baseados em densidade. Iremos usar as mesmas variáveis nos dois algoritmos, para podermos comparar ambas abordagens.

Começamos por enriquecer ainda mais o conjunto de dados com novas variáveis relevantes para os bancos, num processo de engenharia de variáveis (ou *feature engineering*), útil na criação de KPIs. Apresentamos as suas definições em seguida.

Valor Vitalício do Cliente (VVC) (ou *Customer Lifetime Value)*, é um KPIs muito utilizado em CI, concentrando-se em estimar o potencial valor ou lucratividade de um cliente ao longo de todo o seu relacionamento com uma empresa (73) (ver eq. 4.3 (74)). O VVC fornece uma visão abrangente do valor a longo prazo de um cliente e ajuda a identificar clientes de alto valor que provavelmente contribuirão mais para o negócio ao longo do tempo.

$$\text{VVC} = \frac{\text{VMM} \times \text{FT} \times \text{TVP}}{\text{1 - TPC}} \tag{4.3}$$

Onde VMM é o valor do montante de transação médio desse cliente, FT a frequência das transações, TVP o tempo médio de vida do cliente (a diferença entre a primeira e última transação deste) e TPC a taxa de perda de clientes. Neste último, procedeu-se ao cáculo de quantos clientes já não faziam transações há mais de 90 dias, revertendo num total de 18 clientes em 4500. A interpretação desta variável é a seguinte: um VVC de um cliente hipotético de 150.000 Kč, significa que, em média, esse cliente gera 150.000 Kč de receita para a empresa ao longo de seu relacionamento com o cliente, considerando os valores e as frequências médias deste.

Em seguida, criámos uma variável denominada Taxa de Pagamento Mensal, `TPM`, que é calculada com a fórmula 4.4. Esta é uma métrica que expressa a percentagem do valor total do empréstimo que é paga como prestação a cada mês. Isto permite-nos não só entender quanto tempo um cliente pode demorar a pagar o empréstimo, como também permite comparar entre diversos clientes o quanto pagam do seu empréstimo por mês. Valores baixos indicam prestações baixas, ou uma duração elevada do empréstimo. Por um lado, esta métrica reflete se o cliente está a comprometer uma parcela significativa do seu rendimento para pagar o empréstimo que pode aumentar o risco de incumprimento, ou, por outro lado, por ter maior duração, pagará mais juros ao banco.

$$TPM = \frac{Pagamentos}{Montante} \times 100 \tag{4.4}$$

Por último, por forma a incluir na análise variáveis que reflitam o comportamento do cliente em relação às suas transações utilizámos as variáveis de RFM obtidas na secção 4.1.1 de forma a obter a inserir mais informação nos dados. Dado terem diferentes escalas decidimos normalizar as mesmas usando quartis, denominada normalização por quartis ou normalização robusta. Existem várias vantagens inerentes a esta técnica: a redução do impacto de valores extremos, a comparação justa entre variáveis a preservação das características originais dado que mantém a ordem relativa dos dados e maior facilidade na interpretação. Assim dividimos as três variáveis R, F e M, pelos 4 quartis (0%-25%, 25%-50%, 50%-75% e 75%-100%), nas variáveis respetivas: `Recencia_Quartil`, `Frequencia_Quartil` e `Monetario_Quartil`.

Tendo estas variáveis, incluímos também na análise a faixa etária, variável `Faixa_etaria`, o estado do empréstimo, variável `Estado_empr`, e o segmento do cliente, variável `Segmento`. Como são variáveis categóricas foi necessário codificá-las utilizando um *encoder* numérico, dado que é possível atribuir uma hierarquia entre os seus valores. As variáveis com os quartis também foram codificadas com um *encoder* numérico, pela mesma razão.

Não utilizámos a variável demográfica género por duas razões, primeiro, dado ser uma variável binária teria bastante impacto no modelo não sendo interessante dividir os clientes por género, e segundo poderia ser considerada uma decisão descriminatória. A variável zona geográfica não contém uma forte correlação com as restantes variáveis criadas, nem entre si (após *one-hot encoding*), logo é de baixo valor informativo. De notar que na tarefa de segmentação não é interessante

usar demasiadas variáveis pois tornaria a interpretação mais complexa, ainda assim, na etapa de descoberta de padrões posterior é possível caracterizar a distribuição dos grupos em relação às restantes variáveis não utilizadas nesta etapa através dos *itemsets*.

Resumindo as variáveis utilizadas para o agrupamento foram: `Faixa_etaria`, `TPM`, `Recencia_Quartil`, `Frequencia_Quartil`, `Monetario_Quartil`, `VVC`, `Segmento` e `Estado_empr`.

Por último, foi necessário normalizar todas as variáveis, e para tal utilizámos o algoritmo *minmax*.

Algoritmo *k-means*

Para decidir qual o valor de k ideal é recomendado testar vários métodos e escolher tendo em consideração todos os resultados obtidos. Neste trabalho foram usados 4 métodos diferentes de análise: Métricas de avaliação da qualidade de grupos; gráficos de silhueta; dendograma resultante da aplicação do método HCA; e análise da relação entre a magnitude e cardinalidade do grupo.

Começámos por testar para vários valores de k quais os resultados do modelo face a diferentes métricas de avaliação mais utilizadas na literatura: coeficiente de silhueta, índice de Calinski Harabasz e inércia.

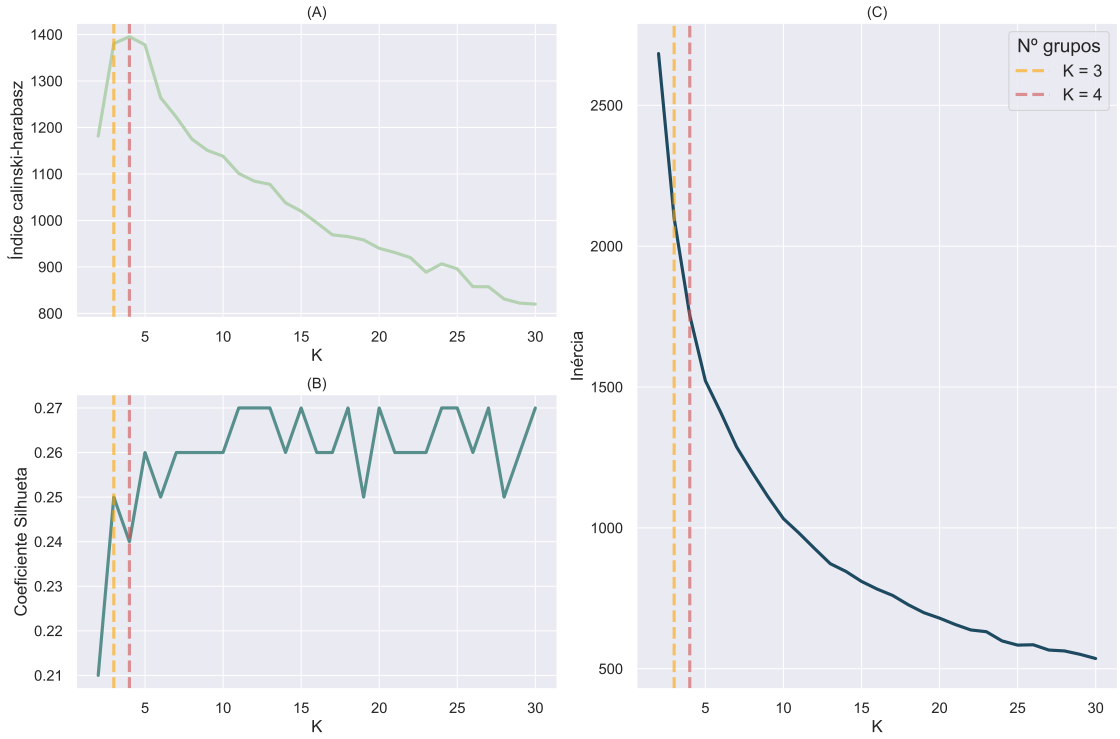

Figura 4.4: Resultados das diferentes métricas de avaliação para vários valores de k, no algoritmo *k-means* com inicialização "k-means++".

O coeficiente de silhueta é uma métrica que avalia a qualidade dos grupos, medindo o quão bem um ponto está agrupado em relação aos outros pontos no mesmo grupo, levando em consi-

deração a distância média para os pontos do mesmo grupos (a coesão) e a distância média para os pontos dos grupos vizinhos (a separação) (75). O valor do coeficiente de silhueta varia de -1 a 1, onde valores mais próximos de 1 indicam grupos mais esféricos e de melhor qualidade. O índice de Calinski-Harabasz, também conhecido como critério de variação entre e dentro dos grupos, é uma métrica que avalia a separação entre estes. Este índice compara a dispersão (variância) entre os grupos com a dispersão dentro dos grupos (76). Quanto maior o valor do índice, melhor a separação entre os grupos. A inércia é uma métrica que mede a soma das distâncias quadráticas entre cada ponto e o centro do grupo ao qual pertence (77). Noutras palavras, é uma medida da compactação dos grupos. Quanto menor a inércia, mais compactos e similares são os grupos. A inércia é frequentemente usada para escolher o número ideal de grupos, pois diminui à medida que estes aumenta, e geralmente é utilizada em conjunto com o método do cotovelo (*elbow method*).

A partir da fig. 4.4 vemos que $k = 4$ é a melhor escolha segundo o gráfico (A) e $k = 3$ é a melhor escolha segundo o gráfico (B). Pelo gráfico (C) é difícil de perceber onde fica o cotovelo, sendo inconclusivo.

Em seguida foi importante perceber dado o algoritmo de *k-means* depender da inicialização dos centróides e não garantir que a solução encontrada é a global como essa aleatoriedade poderia influenciar as métricas de avaliação. Para tal corremos 100 vezes o mesmo algoritmo *k-means* com diferentes *seeds* geradas por um gerador de números aleatórios, para $k = 3$ e $k = 4$. Concluímos que não existe variação, implicando que o algoritmo consegue facilmente encontrar os mesmos grupos e convergir para a mesma solução.

Olhámos também para os gráficos de silhueta (ver fig. 4.5) e o dendograma (ver fig. 4.6) para determinar qual o melhor k entre 3 e 4.

Os gráficos de silhueta são uma técnica de avaliação de grupos que medem quão bem cada amostra se encaixa no seu grupo atribuído. Cada amostra é representada por uma barra horizontal no gráfico, cujo comprimento é proporcional ao seu valor de silhueta. Além disso, as barras são agrupadas por grupo, com uma cor diferente para cada grupo. Um valor de silhueta próximo de +1 indica que a amostra está bem ajustada ao seu grupo e relativamente longe dos restantes grupos, um valor próximo de 0 indica que a amostra está na fronteira entre dois grupos. Valores negativos sugerem que a amostra pode ter sido atribuída ao grupo errado. Idealmente, procuramos obter uma solução com grupos de aproximadamente o mesmo tamanho sem pontos com silhueta negativa, todos os grupos devem ainda estar acima da linha do valor médio de silhueta.

Um dendrograma é uma representação gráfica que apresenta a estrutura de agrupamento dos dados em forma de uma árvore, onde as amostras são agrupadas com base na sua similaridade em diferentes níveis de hierarquia. A avaliação de um dendrograma envolve a identificação visual de linhas de corte. Utilizámos duas métricas de conexão, a completa, que calcula a distância entre dois grupos como a maior distância entre quaisquer duas amostras. Isso significa que ela considera a distância máxima entre todos os pares de amostras dos grupos, preservando a distribuição global dos dados. A conexão completa tende a formar grupos compactos e bem separados, mas pode ser sensível a *outliers* e ruídos nos dados. Já a conexão de *Ward* calcula a distância entre dois

grupos com base na soma dos quadrados das diferenças entre as amostras de cada grupo. Esta medida tenta minimizar a variância dentro de cada grupo ao fundi-los e tende a formar grupos mais equilibrados em termos de tamanho e densidade, sendo por isso menos sensível a *outliers*. Tentamos, assim, visualmente encontrar ramos isolados e distintos que denotam um grande nível de dissimilaridade com os restantes grupos, aí se definem as linhas de corte.

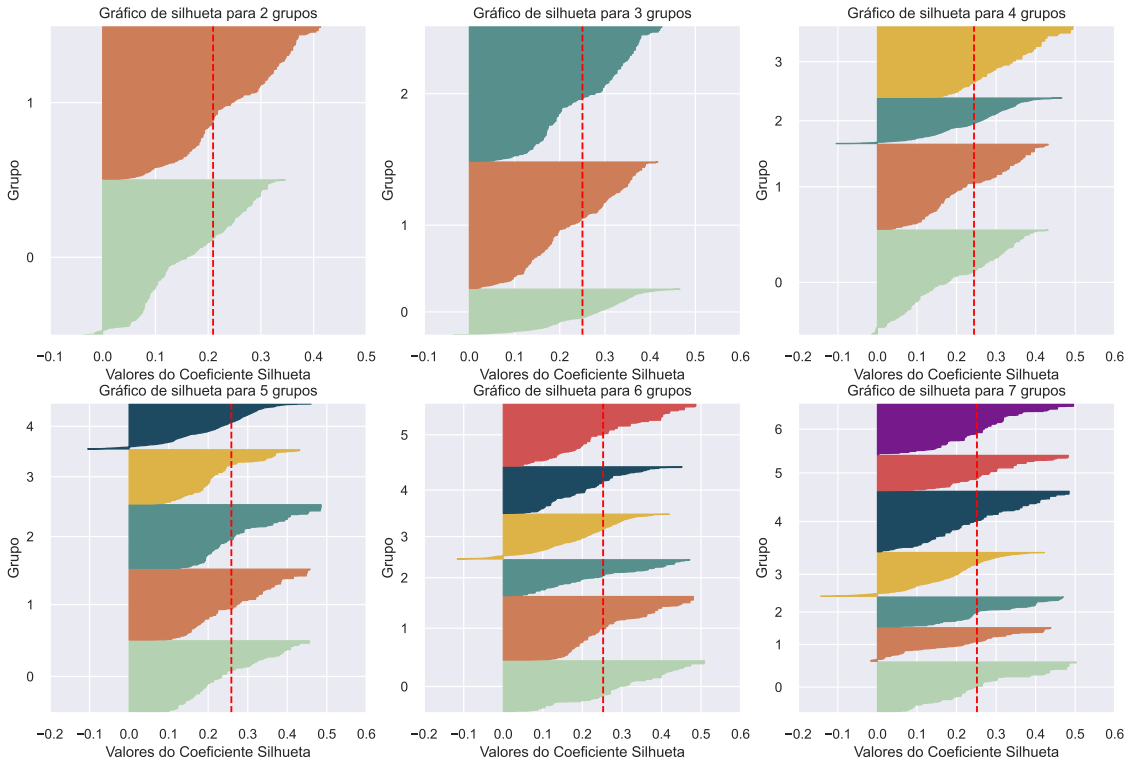

Figura 4.5: Gráficos de silhueta do algoritmo *k-means* com $k \in [2, 7]$.

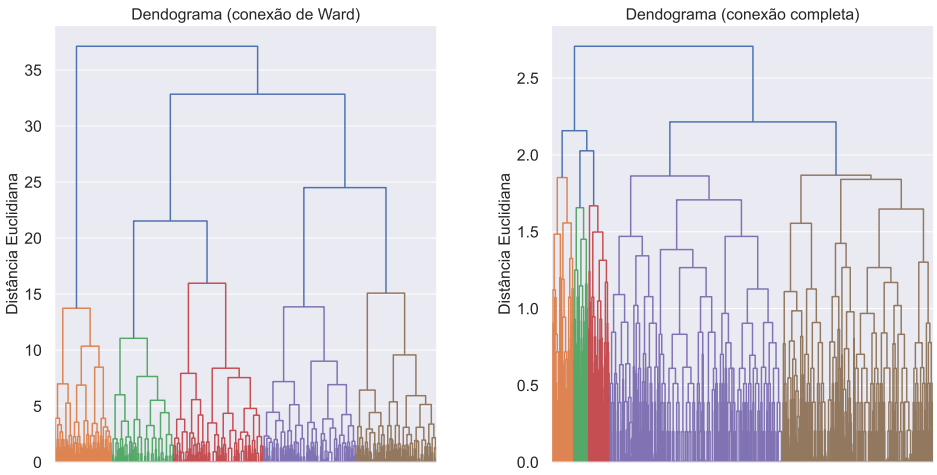

Figura 4.6: Dendograma do HCA para a conexão de *Ward*, na esquerda, e completa, na direita.

Observamos que o melhor valor de k é 3, segundo a fig. 4.5, dado que é o modelo mais simples que tem silhuetas positivas e acima da linha vermelha, apesar de ter um grupos mais pequeno (com a etiqueta 2) em relação aos restantes. Segundo a fig. 4.6 o dendograma com a conexão completa não tem uma linha de corte óbvia, já o de ward é fácil obervar que $k = 3$ é o melhor valor. Assim confirmamos novamente a escolha de $k = 3$.

A última análise feita, foi em relação à cardinalidade e magnitude dos grupos (ver fig. 4.7).

O gráfico da cardinalidade do grupo mostra o número de pontos pertencentes a cada grupo. Podemos verificar se alguns grupos contêm significativamente menos ou mais pontos do que os restantes. Neste caso, observamos que tanto $k = 3$ como $k = 4$ contém grupos com tamanhos diferentes, sendo que para $k = 3$, a diferença é mais desproporcional. O gráfico da magnitude do grupo mostra a distância total dos pontos em relação ao seu centróide. Permite observar o quão dispersos estão os pontos em cada grupo e se existem grupos mais densos ou mais dispersos. Por fim, o gráfico da Cardinalidade *versus* Magnitude, que reúne os anteriores, compara a relação entre as duas métricas. Grupos com maior cardinalidade tendem a ter também uma maior magnitude. Idealmente queremos que os pontos se encontrem na linha dos 45 graus a azul, ou muito próximos dela. As anomalias tendem a estar mais distantes desta linha. Para $k = 4$, apesar de poder haver mais informação existem dois grupos (2 e 0), claramente *outliers*.

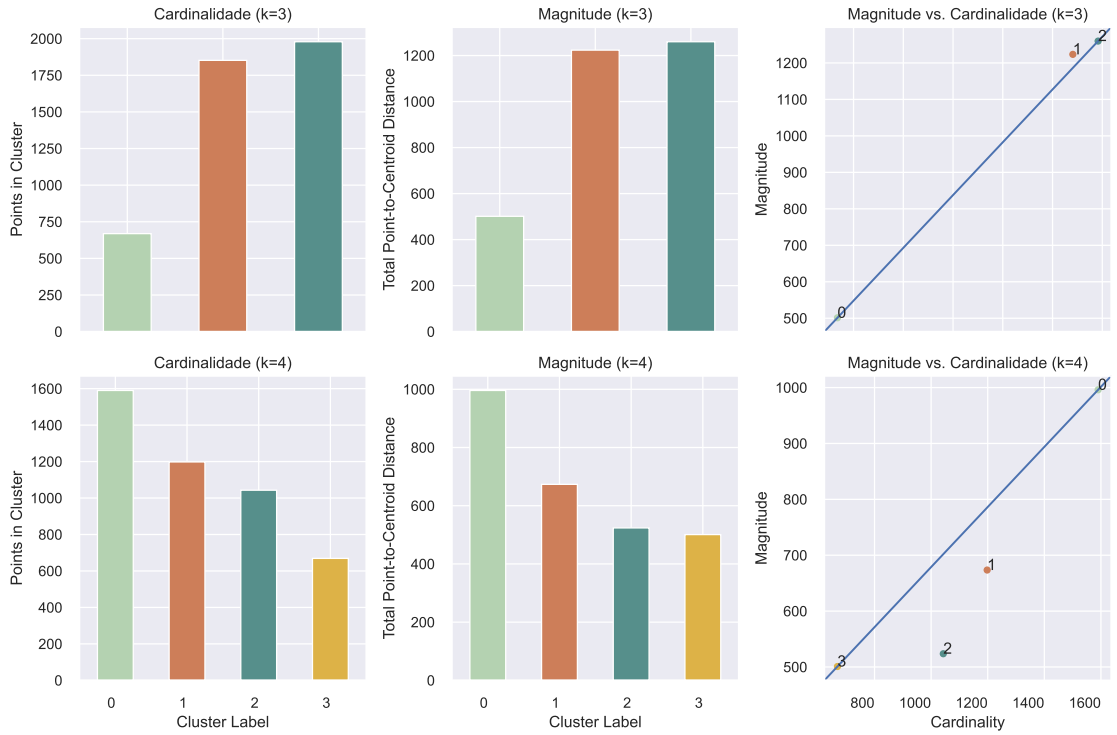

Figura 4.7: Gráficos da cardinalidade, magnitude e cardinalidade *versus* magnitude, para os valores de $k = 3$ (1ª linha) e $k = 4$ (2ª linha). Para construir os gráficos recorreu-se à biblioteca "ds_utils.unsupervised"de python.

Tendo em conta os resultados de todos os métodos, treinamos o *k-means* para $k = 3$ como

modelo de segmentação final, para o qual obtivémos as seguintes estatísticas descritivas resumidas na tabela 4.1.

Tabela 4.1: Estatísticas descriptivas para cada grupo e variável utilizada na tarefa de segmentação. As medidas são $\overline{x}$ média, M_d mediana e σ desvio padrão. Relembramos que os valores categóricos foram codificados, apresentando aqui o resumo dos seus valores numéricos.

Grupo	**0**			**1**			**2**		
	$\overline{x}$	M_d	σ	$\overline{x}$	M_d	σ	$\overline{x}$	M_d	σ
Faixa_etaria	1.49	2.00	0.71	1.68	2.0	0.92	1.71	2.0	0.93
TPM	0.20	0.08	0.25	-1.00	-1.0	0.05	-1.00	-1.0	0.07
Recencia_Quartil	1.05	1.00	1.09	1.25	1.0	1.23	1.35	1.0	1.26
Frequencia_Quartil	1.77	2.00	1.05	2.49	2.0	0.50	0.48	0.0	0.50
Monetario_Quartil	0.92	0.00	1.20	1.67	2.0	1.14	1.54	1.0	1.00
VVC	85.42	89.99	12.58	89.95	92.2	7.56	91.43	91.8	1.54
Segmento	2.91	3.00	1.19	3.55	4.0	0.71	1.46	1.0	0.62
Estado_empr	2.16	2.00	0.72	-1.00	-1.0	0.05	-1.00	-1.0	0.06

Para interpretar melhor os grupos podemos visualizar o desvio de cada variável em cada grupo da média geral dessa variável. O gráfico radar também é uma visualização eficaz para observar diretamente as diferenças entre os diferentes grupos (ver fig. 4.8).

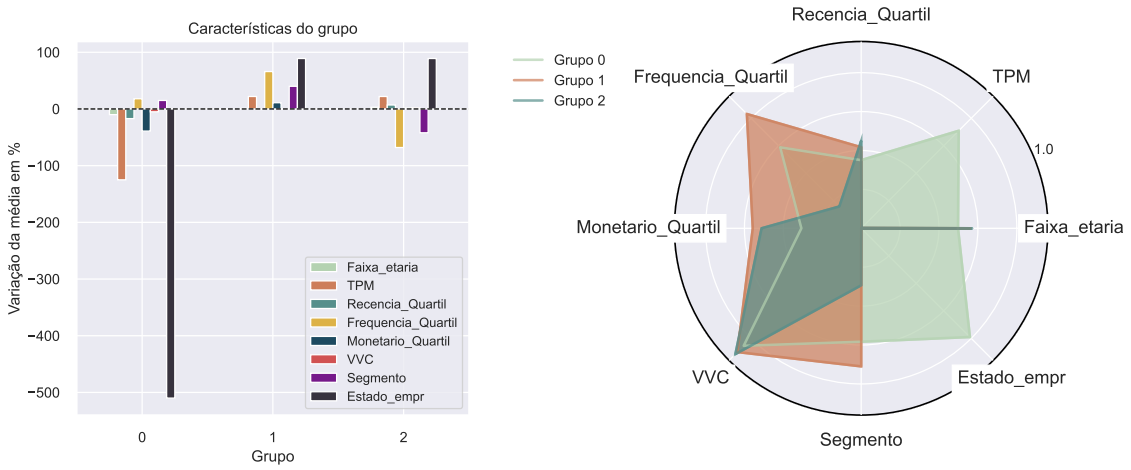

Figura 4.8: Na esquerda, o gráfico de barras apresenta o desvio percentual em relação à média de cada variável em cada grupo, e no gráfico radar da direita, é apresentada a distribuição normalizada de cada variável em cada grupo.

De seguida verificámos como a distribuição das restantes variáveis demográficas como o género e a zona geográfica estão presentes nos grupos e se existe alguma segmentação notável a acrescentar à análise. A caracterização demográfica, presente na fig. 4.9, permite-nos visualizar qual a distribuição das variáveis demográficas de género e zona geográfica e perceber como variam entre grupos. A zona metropolitana de Praga é a que tem maior percentagem dado metade dos clientes do conjunto de dados pertencerem a esta zona. Verifica-se também que todos os grupos contêm clientes de todas as regiões demográficas, sendo que os grupos 1 e 2 têm percentagens

muito semelhantes. Não existem diferenças significativas em relação ao género.

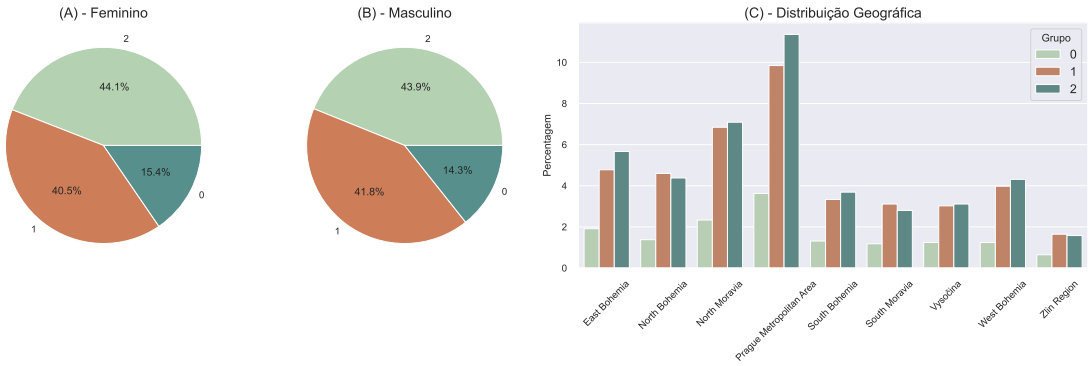

Figura 4.9: Os gráficos (A) e (B) apresentam a distribuição dos clientes nos grupos do género feminino e masculino, respetivamente. O gráfico (C) apresenta a distribuição dos clientes em cada grupo em função da sua zona geográfica.

Combinando a análise realizada, e tendo como referência a tabela 4.1 e a fig. 4.8 podemos inferir as seguintes descrições dos grupos:

- **Grupo 0 - Clientes com empréstimos (CE)**: representa um grupo de clientes, em média, ligeiramente mais jovem comparando com os restantes grupos, e possuem atividade de empréstimo. Fizeram na maioria transações recentemente, com frequências médias mas valores monetários baixos, ainda que com um desvio padrão grande. O valor médio do tempo de vida do cliente para este grupo é o mais baixo comparando com os outros grupos, mas ainda assim elevado. Estes clientes possuem a maior variação em relação ao segmento enquadrando-se, em média, em `Clientes de médio valor`. O número de clientes deste grupo é 669 (15% dos clientes totais);

- **Grupo 1 - Clientes de alto valor (CAV)**: consiste em clientes, em média, com idade semelhante ao grupo 1 mas ligeiramente mais jovens que o grupo 2 não possuindo empréstimos. Estes fizeram transações recentes, e possuem uma frequência de transação bastante elevada, em comparação com outros grupos. Os valores das transações variam dentro desse grupo sendo o mais elevado dos três, e o valor médio do tempo de vida do cliente é relativamente alto. Estes clientes enquadram-se especialmente no segmento de `Clientes de alto valor`. O número de clientes deste grupo é 1852 (41% dos clientes totais);

- **Grupo 2 - Clientes com potencial (CP)**: representa clientes em média, mais velhos, comparando com os restantes, sem atividade de empréstimo. Estes fizeram transações recentemente, em média, com maior desvio padrão. Exibem baixíssimas frequências de transações, em média. No entanto, os valores das transações são os mais elevados, sendo, consequentemente, o valor médio do tempo de vida do cliente o mais elevado, entre os 3 grupos. Em média, fazem parte do segmento `Clientes de baixo valor`. O número de clientes deste grupo é 1979 (44% dos clientes totais).

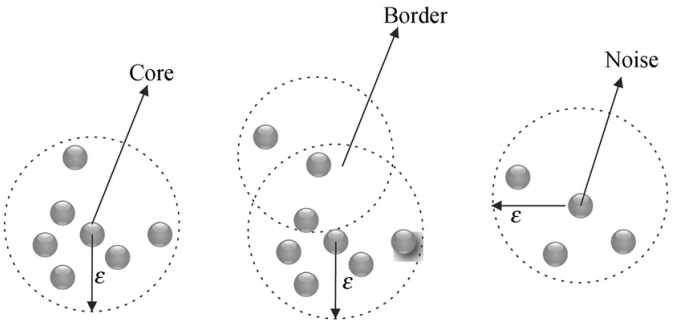

Figura 4.10: Exemplo da classificação dos pontos nas três categorias possíveis (central, fronteira e ruído, respetivamente), com base no algoritmo DBSCAN (2).

Para analisar mais detalhes da distribuição de cada variável por grupo pode-se recorrer ao anexo A.3.

Algoritmo DBSCAN

Um segundo algoritmo bastante utilizado na literatura para segmentação de clientes é o DBSCAN. Este é um algoritmo de agrupamento que se baseia na densidade dos pontos para segmentar o conjunto de dados. Este algoritmo é útil para quando não conhecemos o número de grupos *apriori* e quando os dados têm distribuições não lineares com presença de ruído e *outliers*. A principal diferença entre os dois algoritmos está na forma como os grupos são formados: o *k-means* atribui um ponto a um grupo com base na proximidade a um centróide, enquanto que o DBSCAN baseia-se não na distância entre pontos mas antes na densidade.

O algoritmo DBSCAN classifica cada ponto como um de três possíveis categorias: central, de fronteira e de ruído (ver fig. 4.10). Os pontos centrais são aqueles que possuem um número mínimo de pontos (denominado como "*min_samples*", um dos parâmetros do modelo), dentro de uma determinada distância (denominada "*eps*", o segundo parâmetro do modelo). Os pontos de fronteira estão próximos dos pontos centrais, mas não possuem o número mínimo de pontos na sua vizinhança, determinada por "*eps*". Os pontos de ruído não estão próximos de nenhum ponto central.

O algoritmo DBSCAN, em contraste com o *k-means*, não necessita que se defina *apriori* o número de grupo. Este começa por escolher um ponto aleatoriamente do conjunto de dados e expande o grupo até que não seja possível adicionar mais pontos. Em seguida outro ponto é escolhido, que ainda não tenha sido visitado, e o processo é repetido até que todos os pontos do conjunto de dados tenham sido visitados. Entre os dois algoritmos, o DBSCAN é preferível quando os grupo podem não ser esféricos ou a densidade de pontos varia entre as diferentes regiões do espaço dos dados.

Apesar desta aparente vantagem, a principal desvantagem do DBSCAN é ser muito sensível a hiperparâmetros, poder classificar uma parte significativa dos dados como ruído, que por vezes resulta em métricas melhores às do *K-Means*, mas em grupos pouco relavantes. Tem também tendência consoante o valor de "eps" a formar grandes quantidades de possíveis grupos pequenos,

que torna a análise bastante mais complexa. A análise do método DBSCAN é mais complexa, necessitando de se estudar os resultados das métricas de avaliação para vários parâmetros.

Para determinar os possíveis valores dos parâmetros existem algumas abordagens:

- Em relação ao "*min_samples*", é recomendado, quando os dados não são 2D, utilizar pelo menos $2 \times \#dim$ onde dim o número de dimensões. Este valor determina o tamanho do grupo mínimo, e regulariza o quanto o modelo classifica os dados como ruído (78);
- Para estimar o "eps", é recomendado usar um *k-Distance Plot*. Esta técnica calcula a distância média entre cada ponto e os seus k vizinhos mais próximos, recorrendo ao algoritmo *K Nearest Neighbors*, sendo k o valor $min\_samples - 1$. Estas k-distâncias médias são colocadas num gráfico em função do índice do ponto, por ordem crescente. Para identificar o melhor valor, procura(m)-se o(s) ponto(s) com máxima curvatura ou "*knee(s)*", onde há maior variação do declive. Estes pontos representam valores de corte onde existe uma fronteira entre diferentes regiões de densidades (79).

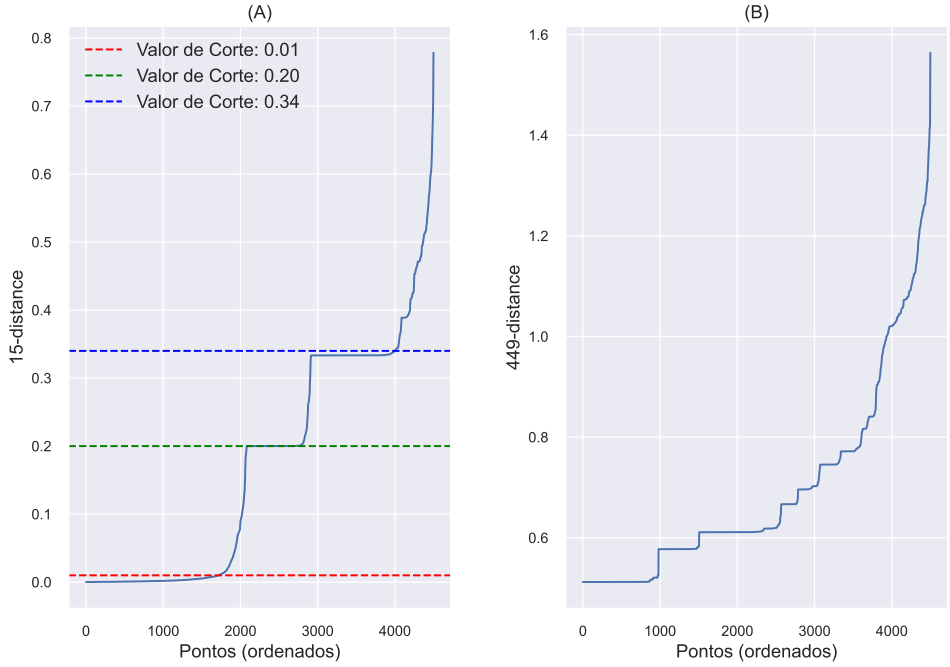

Figura 4.11: No gráfico (A), o *15-Distance plot* e no gráfico (B), o *449-Distance plot*. A tracejado identificam-se os pontos de maior curvatura, aproximadamente, obtidos através da análise visual dos gráficos. No eixo y, encontra-se as k-distâncias médias.

Seguindo o protocolo explicado acima, começamos com o valor "*min_samples*" igual a 16, dado a dimensão, i.e. o número de variáveis, ser 8. Construímos também um segundo gráfico com $min\_samples$ igual a 450. Isto vai permitir comparar os resultados dos gráficos para grupos mais pequenos e específicos ou maiores e gerais (pois com "*min_samples*" definimos o tamanho

mínimo dos grupos encontrados). No caso de $min\_samples = 450$, estabelecemos a regra que cada grupo deve conter no mínimo 10% dos clientes totais.

Na fig. 4.11, podemos observar o *15-Distance plot* e *449-Distance plot*.

Na fig. 4.11 observamos vários possíveis valores para eps, ou seja vários "*knees*". Isto implica que, com as variáveis selecionadas neste conjunto de dados, existem vários níveis diferentes de densidade ou seja, é possível identificar grupos em diferentes escalas, tornando-se inconclusivo o melhor valor de eps com este método.

Sendo assim, testámos um leque de vários valores para combinações de $eps \in [0.01, 0.1, ..., 0.45, 0.5]$ e $min\_samples \in [16, 38, ..., 468, 484]$ (ou seja de 20 em 20). Para avaliar os vários grupos obtidos, utilizaram-se as métricas do coeficiente de silhoueta e o indíce de Calisnki-Harabasz.

Para restringir o espaço de resultados, foram descartados os resultados que obtiveram um número de grupos superior a 10, pois formavam grupos demasiado específicos, e os que consideraram mais de 50% dos dados como ruído.

Com estes critérios, obtivémos o melhor resultado nas métricas de avaliação com coeficiente de silhueta 0.32 e índice de Calinski-Harabasz de 1147, sem ruído, cujos parâmetros utilizados foram $eps = 0.6$ e $min\_samples = 32$.

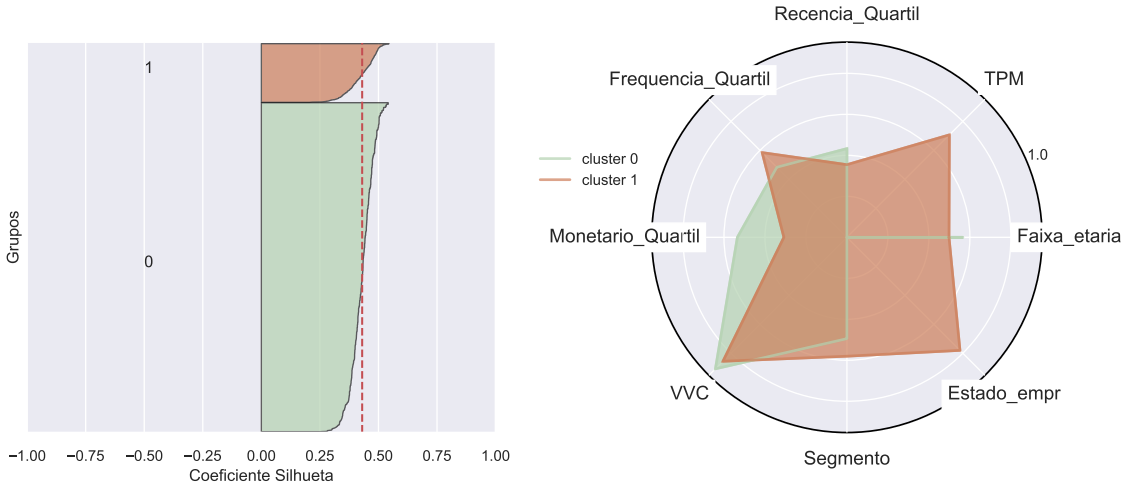

Figura 4.12: No gráfico (A), o gráfico silhoueta, e no gráfico (B) a distribuição de cada variável no respetivo grupo.

Comparando ambas abordagens, concluímos que o algoritmo DBSCAN obteve um valor de silhoueta maior em relação ao *k-means*, no entanto obteve somente 2 grupos com tamanhos bastante desproporcionais. Em relação aos grupos encontrados pelo modelo DBSCAN, analisando o gráfico (B) da fig. 4.12, percebemos que o grupo 0 não tem clientes com atividade de empréstimo, no entanto possuem um valor monetário, em média, superior. Não existe grande distinção em relação ao segmento do cliente e valor *VVC*. Os grupos obtidos com o DBSCAN foram demasiado gerais para as melhores métricas, já o *k-means* apesar de uma qualidade inferior, foi capaz de encontrar 3 grupos mais distintos entre si e interessantes. Sendo assim, prosseguimos a análise com

os resultados do algoritmo *k-means*.

4.2 Descoberta de padrões

A descoberta de padrões surge neste trabalho com o objetivo de perceber que variáveis aparecem mais frequentemente no conjunto de dados para assim criarem-se os perfis de cada grupo. A partir desta técnica é possível construir os perfis de cada grupo através da criação de *itemsets*. Estes têm diferentes níveis de especificidades, i.e., à medida que o comprimento do *itemset* aumenta, mais específico é o perfil encontrado, mas também mais raramente surge no conjunto de dados. Existe além disso, a vantagem de ser possível categorizar as variáveis numéricas e analisar diretamente a sua distribuição nos dados de forma simples de interpretar através dos *itemsets*. Iremos analisar apenas os *itemsets* com comprimento um para construir perfis mais generalizados, no entanto, é possível obter perfis específicos consoante a necessidade.

4.2.1 Construção de perfis

O algoritmo *Apriori* é um dos algoritmos mais utilizados para encontrar padrões, apesar de não ser o mais eficiente para regimes de grandes quantidades de dados. Para o nosso caso, não será muito diferente de um algoritmo mais eficiente pois o conjunto de dados é pouco complexo. Este algoritmo baseia-se no conceito de "propriedade *a priori*", que afirma que, se um conjunto de itens é frequente, todos os seus subconjuntos também devem ser frequentes. O algoritmo alavanca esta propriedade para localizar com mais eficiência do que um algoritmo *brute-force* conjuntos de itens frequentes. Em seguida encontra-se uma visão geral do algoritmo *Apriori*:

1. Inicialmente, o algoritmo varre o conjunto de dados para determinar a frequência de cada item individual (conjuntos de itens de tamanho um). É a etapa de "geração de candidatos";

2. O algoritmo gera conjuntos de itens candidatos de tamanho dois combinando conjuntos de itens frequentes da etapa anterior. Esses conjuntos de itens candidatos são criados pela união de dois conjuntos de itens frequentes, garantindo que, pela propriedade *apriori*, todos os subconjuntos do conjunto de itens candidatos também sejam frequentes;

3. Em seguida, o algoritmo varre o conjunto de dados novamente para contar o suporte (frequência) de cada conjunto de itens candidato. O suporte é o número de transações em que o *itemset* aparece;

4. O algoritmo remove os conjuntos de itens candidatos que não correspondem ao limite mínimo de suporte. Esta etapa reduz o número de conjuntos de itens que precisam ser considerados nas iterações subsequentes;

5. As etapas 2 a 4 são repetidas iterativamente para gerar conjuntos de itens candidatos de tamanhos maiores (3, 4 e assim por diante) até que nenhum conjunto de itens mais frequente possa ser encontrado.

Para o nosso caso de estudo procedeu-se à partição do conjunto de dados original em três conjuntos mais pequenos com características específicas para facilitar a procura de padrões:

- **P1**: contém as variáveis RFM criadas originalmente e as variáveis demográficas dos clientes, para cada grupo;
- **P2**: este conjunto de dados tem as transações dos clientes por cada grupo, inclusive uma nova variável que é o ratio entre o montante de transação e o balanço, em percentagem (`PCT_Racio`);
- **P3**: esta partição contém informação em relação aos empréstimos e clientes, como o seu segmento e variáveis demográficas. Neste caso, só relativo ao grupo Clientes com empréstimos, o único grupo com atividade de empréstimo.

Para aplicar o algoritmo *Apriori* é necessário criar uma base de dados transacional. Para tal, procedeu-se à categorização e codificação de todas as variáveis. No caso das variáveis categóricas, foi realizado o *one-hot encoding*, no caso das variáveis numéricas contínuas é necessário dividir por diferentes intervalos. Dado as variáveis serem enviesadas, a partição destas foi feita com base na sua distribuição, isto é, percentis. De forma a não perder a informação relativa aos valores, renomeou-se estas categorias obtidas para conter o limite inferior e superior dos respetivos intervalos, por exemplo `Recencia_0-10`.

A nova variável criada para a partição P2, o rácio entre o montante da transação e o balanço da conta após o débito/crédito (`PCT_Racio`), permite perceber a personalidade do cliente em relação a ser mais conservador com o dinheiro ou precipitado. No entanto, após criação e visualização da estatística descritiva percebeu-se que esta variável tinha *outliers* que não faziam sentido, pelo que se utilizou o método *z-score*, o mesmo na etapa de pré-processamento, e retirou-se um total de 536 observações.

Por fim, para as partições P1 e P2, dividiram-se novamente as observações por cada grupo obtido na etapa de agrupamento. Isto deveu-se em grande parte à diferença da magnitude de cada grupo que influenciaria o suporte dos *itemsets* encontrados, pois grupos maiores possuem mais instâncias e por isso um suporte maior que obscura os restantes padrões.

Para construir a tabela 4.2 selecionou-se os top 5 *itemsets* com comprimento 1 de cada resultado. Através da análise da tabela 4.2, podemos delinear os perfis mais comuns em cada uma das partições e grupos. No caso do P1, percebemos que a zona geográfica da área metropolitana de Praga sobressai em dois grupos, o grupo Clientes de alto valor (CAV) e o grupo Clientes com potencial (CP), com suporte superior a 20%. A faixa etária de meia idade também sobressai em todos os grupos, com suporte superior a 40%, aparecendo adulto somente no grupo Clientes com empréstimos (CE). Existe uma diferença significativa na variável frequência, sendo que o CP possui clientes com poucas transações, e o CAV com mais transações. Em relação ao valor monetário, o CAV é o único que não possui valores negativos, sendo que o CE contém os clientes que mais débitos efetuam.

Tabela 4.2: Perfis obtidos através do algoritmo *Apriori*, para o nível mínimo de suporte de 5%. O prefixo k é o anglicismo que se refere à ordem dos milhares e M à ordem de milhões, para facilitar a comparação entre valores numéricos. Relembra-se que os respetivos grupos são: Clientes com empréstimos (CE), Clientes de alto valor (CAV) e Clientes com potencial (CP).

Partição	Grupo	Itemset	Suporte (%)
P1	CE	Meia Idade	62
		Monetário -3.2M a -161.9k	51
		Frequencia 341 a 672	28
		Adulto	26
		Frequencia 168 a 242	25
	CAV	Meia Idade	50
		Frequencia 242 a 341	42
		Frequencia 341 a 672	38
		Monetario 53.3k a 371.9k	26
		Prague Metropolitan Area	24
	CP	Meia Idade	48
		Frequencia 9 a 112	42
		Frequencia 112 a 168	40
		Monetario -161.9k a 18.7k	28
		Prague Metropolitan Area	26
P2	CE	Montante -35k a -5k	26
		Balanço 55k a 105k	25
		Montante -1.0k a -15	24
		PCT_ratio -880% a -16%	22
		Balanço 42k a 55k	21
	CAV	Montante -1.9k a -15	31
		PCT_racio -6% a 0%	18
		Balanço 25.2k a 32.2k	18
		Balanço 18.9k a 25.2k	18
		PCT_ratio 0% a 11%	17
	CP	Montante -1.9k a -15	28
		Balanço -27k a 19k	24
		PCT_racio 11% a 869%	19
		Montante 2.9k a 34.5k	19
		Balanço 19k a 25k	18
P3	CE	Meia Idade	62
		Clientes de valor médio	27
		Adulto	26
		Clientes de valor baixo	25
		Prague Metropolitan Area	24

Em relação a P2, observamos que, relativamente ao valor monetário das transações, todos os grupos têm montantes negativos mais frequentes, exceto o CP que possui um *itemset* com montante positivo. O CE é quem faz transações de débito mais elevadas com frequência. Em relação ao balanço, o CE é quem possui o balanço mais elevado, seguido do CAV, e o CP é o único com um *itemset* de balanço negativo frequente. A variável `PCT_Racio` surge no CE com valores de percentagem negativa elevados, que significa que a variável montante tinha valores elevados de

débito comparado com o saldo após a transação. Já o CAV possui percentagens de `PCT_Racio` reduzidas tanto positiva como negativa. Estes resultados revelam várias informações comportamentais em relação aos clientes. No CE, clientes com empréstimo, verificamos que os montantes são negativos com frequência apesar de balanços na maioria positivos, isto significa que estes clientes são menos conservadores com o dinheiro dispendendo-o em quantias elevadas e são financeiramente mais instáveis em relação à sua saúde financeira. O CAV, contém percentagens de `PCT_Racio` reduzidas revelando uma frequência maior de transações reduzidas em relação ao balanço da conta e de forma equilibrada dado a presença de `PCT_Racio` positivo e negativo nos *itemsets*, isto revela uma relação mais conservadora com o dinheiro e equilibrada sendo financeiramente mais estáveis e saudáveis. O CP é um caso interessante, pois possui tanto valores de montante como de balanço positivos e negativos. O `PCT_Racio` elevado indica que o montante da transação é muito maior comparado com o balanço após essa transação. Além disso, por `PCT_Racio` ser positivo, significa que o montante e o balanço assumiam valores positivos, i.e. transações de crédito com balanço positivo, ou valores negativos, i.e. transações de débito com balanço negativo. Isto pode revelar que estes clientes têm mais dívidas para com o banco e uma saúde financeira agravada.

Por último, observamos os resultados em relação aos empréstimos em si, na partição P3, representados somente no CE. Percebemos que os segmentos `Clientes de valor médio` e `Clientes de baixo valor` possuem percentagens semelhantes, e compõem juntos mais de 50% dos clientes com empréstimos. Além disso, observa-se uma grande proporção de clientes de meia idade com empréstimo (cerca de 60%).

4.2.2 *Sequence mining*

A prospeção de sequências é uma técnica de Prospeção de Dados utilizada para descobrir padrões sequenciais ou subsequências num conjunto de dados, identificando padrões onde certos eventos ou itens ocorrem numa ordem ou sequência específica frequentemente. A prospeção de sequências é comumente aplicada a vários domínios, incluindo análise de mercado, análise de *logs* da *web*, sequências biológicas e análise de comportamento do cliente. No nosso trabalho, tendo dados de transações de clientes que podem constituir sequências, exploramos esta técnica para analisar quais as sequências mais frequentes de cada grupo e compará-las entre si, por exemplo em relação ao montante e tipo de transação. Isto permite complementar a análise dos perfis efetuada anteriormente e pode ser extendida a estudar, por exemplo, as sequências de canais (internet, máquinas de multibanco, TPA ou telemóvel) utlizados pelos clientes.

O algoritmo *PrefixSpan* é um dos algoritmos mais utilizados para a prospeção de sequências. Este descobre com eficiência padrões sequenciais frequentes utilizando o conceito de um prefixo. O algoritmo funciona da seguinte forma:

1. O algoritmo *PrefixSpan* começa por identificar itens frequentes no conjunto de dados. Verificando as sequências para determinar o suporte (frequência) de cada item individual;

2. Em seguida, seleciona cada item frequente como um prefixo e estende-o para encontrar padrões sequenciais frequentes. Realiza uma pesquisa em profundidade no conjunto de dados projetado, que consiste na parte restante das sequências que começam com o prefixo escolhido;

3. Ao estender o prefixo, o algoritmo acompanha o suporte das sequências estendidas. Se o suporte de uma sequência exceder um limite mínimo de suporte, ela será considerada uma sequência frequente;

4. O algoritmo aplica recursivamente o mesmo processo ao conjunto de dados projetado, considerando cada sequência frequente como um prefixo. Este passo é repetido até que não sejam encontradas sequências mais frequentes.

Ao utilizar o conceito de prefixo, o algoritmo *PrefixSpan*, à semalhança do *Apriori*, explora com eficiência o espaço de busca e descobre padrões sequenciais frequentes de maneira profunda. Evita assim a geração de sequências candidatas desnecessárias, construindo sobre os prefixos que já são conhecidos por serem frequentes. O resultado do algoritmo *PrefixSpan* é um conjunto de padrões sequenciais frequentes.

Para aplicar o algoritmo *PrefixSpan* ao nosso problema, procedeu-se ao pré-processamento necessário. Começou-se por criar, primeiramente, as sequências com a forma:

$$[(item1, item2), (item1, item2), (item1, item2)]$$

Onde, entre os parêntesis retos, estão as sequências de transações de um único cliente, em que *item1* e *item2*, correspondem às respetivas características dessas transações e entre parêntesis curvos a informação de uma transação. As características de transações escolhidas foram três: o tipo de transação (crédito ou débito), a categoria do montante de transação que se divide em baixo, médio e alto valor, e um indicador de fim de semana através da variável correspondente à data da transação, que toma valor binário verdadeiro no caso positivo. A categoria do montante da transação foi definida de acordo com a eq. 4.5, onde $Q1$, $Q2$ e $Q3$ se referem aos quartis da variável `Montante`.

$$Categoria = \begin{cases} \text{Alto: Montante} > Q3 \\ \text{Médio: } Q1 < \text{Montante } < Q3 \\ \text{Baixo: Montante } < Q1 \end{cases} \tag{4.5}$$

Assim, por exemplo, uma sequência para dois clientes diferentes passa a ter a forma indicada no exemplo em baixo. Estes dois clientes pertencem por exemplo ao CP (delimitado pelo 1º grupo de parêntesis retos), o primeiro com duas transações e o segundo com uma (delimitado pelo 2º grupo de parêntesis retos).

$$[[(\text{'Crédito'}, \text{'Baixo'}, 0), (\text{'Crédito'}, \text{'Baixo'}, 1)], [(\text{'Crédito'}, \text{'Baixo'}, 0)]]$$

O algoritmo *PrefixSpan* tem uma complexidade computacional de $O(m \times n)$, sendo m o comprimento médio das sequências com e n o número total de sequências. Para o nosso conjunto de dados, foi necessário fazer uma amostra dos clientes dos grupos CAV e CP (com quase 2000 clientes). Para tal foi utilizada a fórmula 4.6 para calcular um tamanho de amostra representativo (80):

$$n = \frac{\frac{z^2 \cdot p(1-p)}{e^2}}{\left(1 + \frac{z^2 \cdot p(1-p)}{e^2 \cdot N}\right)} \tag{4.6}$$

Para um intervalo de confianças de 95%, margem de erro 3% e variabilidade $p = 0.5$ (o pior caso), substituindo os tamanhos das populações obtemos o tamanho da amostra 675 para CP, e 679 para CAV. No caso CE usaremos o grupo todo, pois sendo o mais pequeno, inclui 669 clientes no total.

Em seguida foi necessário definir os hiperparâmetros do algoritmo. Selecionou-se um suporte mínimo de 50% (que significa que as sequências só são consideradas frequentes se aparecerem em pelo menos 50% dos clientes). Em relação ao tamanho das sequências frequentes a serem encontradas calculámos o número médio de transações que um cliente faz numa semana, resultando no comprimento de 5. Para ser mais fácil de estudar as sequências obtidas, dado serem um número bastante elevado, procedeu-se à criação de novas métricas que resumem os resultados obtidos na descoberta de sequências:

- **TTD - Tipo de Transação Dominante**: Tipo de transação dominante, isto é se na sequência obtida surge mais vezes débito ou crédito;
- **MTD - Montante de Transação Dominante**: Montante de transação dominante, isto é se na sequência obtida surge mais vezes baixo, médio ou alto;
- **CDR - Rácio Crédito/Débito**: Distribuição do tipo de transação, se aparecer na sequência por exemplo, 4 vezes crédito e 0 vezes débito, o resultado é 4:0;
- **MTR - Rácio Montante de Transação**: Distribuição do montante da transação, se aparecer na sequência por exemplo, 3 vezes baixo, 1 vez médio e 0 vezes alto, o resultado é 3:1:0;
- **FDSC - Contagem Fim de Semana**: Conta o número de vezes em que na sequência aparece o valor binário positivo no indicador fim de semana.

Na tabela 4.3 encontram-se resumidos os resultados da análise realizada com as novas métricas criadas, permitindo estudar com mais detalhe os padrões de transações de cada grupo de clientes.

Observamos que na generalidade todos os grupos gastam com mais frequência do que recebem, em especial o grupo CP que possui maior desproporção. Este também é o único grupo que não possui transações de altos valores dominantes nas suas sequências, revelando ter um poder económico mais baixo. O grupo CE é o que possui montantes de altos valores dominantes com mais frequência, superior a 50% de suporte nas sequências encontradas. Sendo o único grupo com

atividade de empréstimo, é interessante ser o que efetua transações mais elevadas revelando gerar mais poder económico ou mais dívida.

Analisando a tabela 4.2, em relação à métrica CDR os grupos CE e CAV têm fluxos de dinheiro mais equilibrados sendo semelhantes, já o grupo CP é o único com sequências frequentes compostas por 5 débitos e 1 único crédito, revelando ter uma gestão financeira mais fraca que os restantes grupos. O grupo CAV é o que apresenta um comportamento mais equilibrado com o seu dinheiro, sendo também o que possui maior poder económico analisando as sequências e montantes dominantes. É seguido de perto pelo grupo CE. Em relação à métrica MTR, destacamos as com um suporte superior a 10%, sendo o grupo CAV o que possui mais sequências com elevados montantes e o grupo CP sequências frequentes com maioritariamente baixos valores. Em relação aos fim de semanas todos os grupos têm comportamentos semelhantes não existindo grandes diferenças. Percebemos que numa média semanal para este conjunto de clientes de cinco transações por semana, sendo que 2 a 3 transações são feitas no fim de semana (com frequência superior a 50%) existe alguma preferência para transacionar neste período.

Tabela 4.3: Resultados obtidos com as métricas criadas para as sequências frequentes encontradas pelo algoritmo *PrefixSpan*, com os respetivos valores de suporte. Relembra-se que os respetivos grupos são: Clientes com empréstimos (CE), Clientes de alto valor (CAV) e Clientes com potencial (CP).

Grupo/Métricas	CE	CAV	CP
TTD Crédito	34%	37%	27%
TTD Débito	66%	63%	73%
MTD	Baixo (32%) Médio (14%) Alto (53%)	Baixo (41%) Médio (18%) Alto (41%)	Baixo (79%) Médio (20%) Alto (0%)
CDR	2:3 (31 %) 1:4 (23%) 3:2 (21%)	2:3 (30%) 3:2 (22%) 1:4 (21%)	2:3 (27%) 1:5 (26%) 3:2 (15%)
MTR	2:1:2 (14%)	2:1:2 (15%) 3:1:1 (12%) 3:0:2 (10%)	4:1:0 (26%) 3:2:0 (26%) 2:3:0 (12%) 5:0:0 (10%)
FDSC	2 (32%) 3 (30%) 1 (17%) 4 (14%)	2 (34%) 3 (28%) 1 (20%) 4 (11%)	2 (32%) 3 (28%) 1 (21%) 4 (12%)

4.3 Classificação de clientes

Após se ter identificado quais os perfis de clientes bancários e estudado os respetivos grupos em função do seu comportamento, pretendemos automatizar a (re)classificação dos clientes. O objetivo é treinar um modelo de AA que classifique o novos clientes, ou os já existentes após algum tempo, no grupo em que mais se enquadrarem.

4.3.1 Avaliação de modelos de Aprendizagem Automática

Com base na revisão da literatura desenvolvida no capítulo 2, testámos os seguintes modelos: *Random Forest Classifier* (RFC), *XGBoostClassifier* (XGB) e Regressão Logística (RL). Além disso, acrescentou-se a este conjunto o modelo *Gaussian Naive Bayes* (GNB) para servir como modelo base de comparação, dado ser simples.

Foram testadas inicialmente diferentes partições dos dados para perceber como o uso de certos conjuntos de variáveis relacionadas influenciam a qualidade do modelo. Assim, é possível perceber e quantificar qual conjunto de variáveis possuí informação mais discriminativa do grupo e contribui para um modelo mais eficaz. As partições testadas foram as seguintes:

- **P1**: Dados demográficos, contém variáveis como o género, a faixa etária e a zona geográfica;
- **P2**: Em adição às variáveis demográficas contém também as variáveis RFM obtidas na secção 4.1.1, inclusive: Recência, Frequência, valor Monetário, Score RFM e segmento de cliente;
- **P3**: Variáveis de transações, inclusive: número da conta, montante de transação, balanço e tipo de transação;
- **P4**: Variáveis de transações e variáveis demográficas;
- **P5**: Variáveis demográficas e de empréstimos, como o número de prestações, valor do empréstimo e estado;
- **P6**: Variáveis demográficas, de RFM e de empréstimos.

Após selecionar os dados para cada partição procedeu-se à codificação das variáveis categóricas, onde se efetuou o mesmo processamento e raciocínio que nos capítulos anteriores, e normalização dos dados com dois métodos, o *minmax* e o *standard scaler*. O primeiro é útil para o modelo RFC e o último para os restantes modelos.

As métricas escolhidas para avaliar os modelos foram a precisão, *recall* e *f1-score*. As médias micro e macro são duas abordagens diferentes de obter medidas de desempenho do modelo em problemas de classificação multiclasse. A média micro tem em consideração o desempenho de cada exemplo individualmente, ou seja, o desempenho global do classificador em todas as classes. Já a média macro, por outro lado, calcula as métricas separadamente para cada classe e no fim calcula a média dessas métricas, ponderando igualmente a classe independentemente do número de exemplos. Como existem 3 grupos, o problema a resolver é multiclasse, além disso as proporções de cada grupo no conjunto de dados não são iguais. Pelo que se deve ter isto em conta nas métricas de avaliação dos modelos e parâmetros. Assim sendo, queremos utilizar a média macro pois tem em consideração a distribuição dos exemplos em cada classe ao calcular métricas, permitindo que as classes minoritárias tenham uma maior contribuição no cálculo final das métricas, sendo a mais adequada para problemas *imbalanced* (classes com tamanhos diferentes).

Queremos em primeiro lugar, comparar os vários modelos e perceber qual o que na sua generalidade obtém melhores resultados nas diferentes partições de dados. Na tabela 4.4 encontra-se um resumo dos hiperparâmetros usados no treino dos modelos, na grande maioria utilizaram-se os hiperparâmetros pré-definidos.

Tabela 4.4: Hiperparâmetros alterados em relação aos pré-definidos na documentação, que foram utilizados no treino dos modelos de Aprendizagem Automática. Os restantes hiperparâmetros, denominados nesta tabela por "Pré-definidos" que diferem entre modelos, podem ser consultados na respetiva documentação: XGB (6), RFC (7), GNB (8) e RL (9).

Modelo	Parâmetros
XGB	max_delta_step = 5 num_class = 3 objective: multi:softmax eval_metric: mlogloss Pré-definidos
RFC	class_weight = balanced Pré-definidos
GNB	Pré-definidos
RL	class_weight = balanced Pré-definidos

No RFC e RL, o hiperparâmetro *class_weight* permite-nos dar o mesmo peso a todas as classes, sendo que o algoritmo vai tentar a aprender todas as classes e não somente a que tem mais exemplos. Em relação ao modelo XGB, escolheram-se os hiperparâmetros mais adequados a um problema multiclasse. O hiperparâmetro `max_delta_step` controla o ritmo de aprendizagem do modelo, pelo que colocámos o valor 5 para ajudar a estabilizar o treino e prevenir sobreajuste. O hiperparâmetro `num_class` é o número de classes, neste problema são 3 grupos. O hiperparâmetro `objective` define a função de custo que o modelo tenta otimizar, o *multi:sofmax* (81) é comumente utilizado em problemas multiclasse. O hiperparâmetro `eval_metric` define como o erro é calculado, neste caso, *mlogloss* (82) mede a perda de entropia para várias classes sendo a mais indicada para o problema.

Após a implementação dos modelos e respetivo treino nas partições dos dados definidas anteriormente, é possível resumir na tabela 4.5 os resultados dos modelos no conjunto de teste que corresponde a 20% dos exemplos disponíveis. A divisão entre conjunto de treino e de teste foi feita com o módulo *train_test_split* (83) do *sklearn.selection*.

A partir da tabela 4.5 concluímos que a partição 6, que conjuga as variáveis RFM, demográficas e de empréstimos é a que obtém os melhores resultados, chegando a uma classificação perfeita pois os dados eram simples e os grupos facilmente distinguíveis. Com as variáveis demográficas (P1), não obtemos discriminação suficiente entre as classes obtendo métricas semelhantes aos de um modelo aleatório (1/3 de probabilidade de acertar). Daqui podemos concluir que estas variáveis não possuem poder explicativo. Já as variáveis RFM conseguem introduzir maior poder explicativo aumentando até 0.4 as métricas, dado terem sido utilizadas para efetuar o agrupamento. As

partições que contêm a informação de transações (P3 e P4) conseguem aumentar a precisão para os modelos mais complexos (o XGB e o RFC), pois existe um maior número de exemplos disponíveis para o treino. Em geral, o XGB e o RFC são os que obtêm melhores resultados, exceto na partição P5, onde se observa o oposto.

Tabela 4.5: Resumo da performance dos modelos de Aprendizagem Automática selecionados, para as 6 partições de dados, no conjunto de teste. As métricas apresentam o resultado obtido com média macro. A negrito, nas colunas numéricas, destaca(m)-se o(s) melhor(es) modelo(s) em cada métrica. Sublinhado a cinza destaca(m)-se o(s) modelo(s) com melhor resultado nas três métricas, para a partição específica.

Partição	Modelo	Precisão	Recall	F1-Score
P1	XGB	**0.40**	0.33	0.30
	RFC	0.34	**0.35**	0.30
	RL	0.36	**0.35**	**0.33**
	NB	0.35	**0.35**	0.32
P2	**XGB**	**0.73**	**0.70**	**0.70**
	RFC	0.72	**0.70**	0.68
	RL	0.70	**0.70**	**0.70**
	NB	0.67	0.66	0.65
P3	XGB	**0.90**	0.74	0.77
	RFC	0.86	**0.79**	**0.82**
	RL	0.52	0.56	0.51
	NB	0.45	0.51	0.47
P4	XGB	0.91	0.77	0.80
	RFC	**0.96**	**0.93**	**0.94**
	RL	0.52	0.56	0.51
	NB	0.58	0.52	0.58
P5	XGB	0.65	0.65	0.65
	RFC	0.65	0.66	0.65
	RL	0.67	0.67	**0.67**
	NB	**0.68**	**0.68**	0.66
P6	**XGB**	**1.00**	**1.00**	**1.00**
	RFC	**1.00**	**1.00**	**1.00**
	RL	0.99	0.99	0.99
	NB	0.96	0.97	0.97

Com base nos resultados obtidos, escolhemos o RFC como o melhor modelo a utilizar no resto da análise a realizar.

4.3.2 Treino do melhor modelo

Após a escolha do melhor modelo, os próximos passos são a seleção de variáveis, treino com recurso a validação cruzada e hiperparametrização.

Seleção de variáveis

É recomendado utilizar mais que um método de seleção de variáveis de forma a possuir maior variedade nas abordagens utilizadas, assim recorremos a 4 métodos diferentes para comparar os resultados e nos auxiliar na escolha das melhores variáveis deste problema concreto. Os métodos escolhidos são o *Boruta*, o *Lasso*, o *Extra Trees Classifier* e o *SelectKBest*, com naturezas distintas. Fez também sentido a criação de uma variável binária `Tem_credito?`, dado um dos grupos ter somente clientes com créditos. De seguida, apresentamos o seu modo de funcionamento, bem como as principais vantagens e desvantagens de cada um:

- ***Boruta*** (84; 85): O método é baseado num algoritmo *Random Forest*. Este compara as variáveis originais com variáveis "sombra" criadas aleatoriamente para determinar a importância de cada variável. Cada variável sombra é uma réplica de uma variável original, mas os valores das variáveis sombra são aleatoriamente baralhados ou permutados. Essa aleatoriedade garante que as variáveis sombra não possuam nenhuma relação real com a variável alvo nem com as outras variáveis. O Boruta atribui importância às variáveis com base na sua capacidade de superar as variáveis sombra. Este método é robusto em relação a ruído e consegue lidar com interações complexas entre as variáveis. Também lida bem com conjuntos de dados com um grande número de variáveis.

- ***Lasso* (Regularização L1)** (86): É um método que utiliza a regularização L1 para estimar os coeficientes das variáveis num modelo linear. Este penaliza os coeficientes menores que zero, o que leva à seleção automática das variáveis mais relevantes. O coeficiente *alpha* é um hiperparâmetro do Lasso que controla o nível de regularização aplicado ao modelo. A regularização é uma técnica utilizada para evitar o sobreajuste e melhorar a capacidade de generalização do modelo. Valores mais altos de *alpha* resultam em mais restrições e menos variáveis selecionadas, enquanto valores mais baixos de *alpha* relaxam as restrições e permitem que mais variáveis sejam mantidas no modelo. O Lasso é eficaz na seleção de variáveis nos problemas de alta dimensionalidade, onde há muitas variáveis com pouca relevância. Também é capaz de lidar com multicolinearidade entre as variáveis.

- ***Extra Trees Classifier*** (87): É um algoritmo de Aprendizagem Automática que se baseia no modelo de árvores de decisão aleatória. Este utiliza várias árvores de decisão construídas a partir de diferentes subconjuntos de variáveis e calcula a importância de cada variável com base na média das importâncias obtidas em cada árvore. A importância de uma variável é calculada com base na redução média da impureza (por exemplo, o índice Gini ou a entropia) obtida ao usar essa variável para fazer divisões nos nós. Após a construção das árvores, o *Extra Trees Classifier* calcula a importância média de cada variável ao longo de todas as árvores. Esta importância é um *proxy* da relevância geral de cada variável para o problema. É um modelo rápido e eficaz na seleção de variáveis relevantes, sendo menos suscetível a sobreajuste em comparação com métodos como o *Random Forest*. As pontuações de

importância podem ser normalizadas de forma que a soma de todas as pontuações seja igual a 1, permitindo uma comparação relativa entre as variáveis.

- ***SelectKBest*** (88): É um método que utiliza testes estatísticos univariados para atribuir um *score* de importância a cada variável. Este seleciona as K melhores variáveis com base em critérios estatísticos, como análise de variância (ANOVA) ou teste qui-quadrado. O *SelectKBest* é útil quando se deseja selecionar um número específico de variáveis mais relevantes. Tem a vantagem de ser relativamente simples de implementar e interpretar.

Começando pelo **Método Boruta**, utilizaram-se todas as variáveis da dimensão do cliente, que se destacaram na tabela 4.5. Foi necessário dividir estes dados em treino e teste, numa composição 80% e 20%, respetivamente. Criou-se uma instância do modelo *Random Forest* com os parâmetros pré-definidos. Em seguida, aplicámos este modelo no método selecionador do algoritmo *Boruta* com os parâmetros pré-definidos, exceto o número de estimadores que foi selecionado como "auto". Após treino, conseguimos perceber as variáveis que foram aceites ou rejeitadas e ainda um *ranking* de importância das variáveis. Os resultados demonstram que as variáveis relativas à zona geográfica, o género e a idade foram rejeitadas pelo algoritmo. Já as relativas ao RFM, o segmento do cliente e empréstimos foram aceites. O modelo coloca-as todas no rank 1.

Em relação à **Regularização Lasso**, para o mesmo conjunto de variáveis e um valor de *alpha* de 0.1, são selecionadas 4 variáveis: o `RFM_score` (com coeficiente 0.23), o `Segmento` (com coeficiente 0.05) e 2 variáveis referentes a créditos, o `TPM` (com coeficiente 0.06) e o `Tem_Credito?` (com coeficiente 0.34). O valor de *alpha* foi testado iterativamente de forma a que o modelo escolha algumas variáveis, neste caso foi necessário uma regularização suave, para escolher mais que uma variável com coeficientes diferentes de 0.

No **Método *SelectKBest*** iteramos em k, i.e. o número de variáveis que queremos selecionar, para perceber o *rank* de cada variável. Os testes utilizados foram o *f_classif*, que é um teste F-valor ANOVA utilizado em tarefas de classificação. O segundo teste utilizado foi o do qui-quadrado também bastante utilizado em tarefas de classificação, mas que requer variáveis não negativas. Assim usamos a normalização do *standard scaler* para o primeiro teste, e do *min max*, pois resulta num valor positivo nas variáveis, para o segundo. Ambos resultaram num ranking e seleção de variáveis bastante similar. As 5 variáveis selecionadas por estes métodos pela ordem de mais importante a menos importante são: `Tem_Credito?`, `TPM`, `Estado_empr`, `Pagamentos`. A útlima variável difere entre os métodos, no caso do teste *f_classif* foi selecionado o `RFM_score`, no caso do teste do qui-quadrado, o `Montante`.

Por fim, em relação ao último modelo de seleção de variáveis, o ***ExtraTreesClassifier***, o resultado pode ser resumido num gráfico de barras (ver fig. 4.13). Como se observa, a variável `RFM_score` é a que apresenta maior relevância, seguida da `Frequência`, `Segmento`, `VVC` e `TPM`. Todas têm peso maior que 0.1 nos scores demonstrando importâncias relevantes. Pelo contrário, as variáveis demográficas como género, faixa etária e zona geográfica não demonstram ser importantes, reconfirmando os resutados obtidos nos restantes modelos de seleção.

Com base nos resultados dos diferentes modelos iremos utilizar as seguintes variáveis para treinar o modelo final: `RFM_score`, `Frequência`, `Tem_Credito?`, `Segmento` e `TPM`.

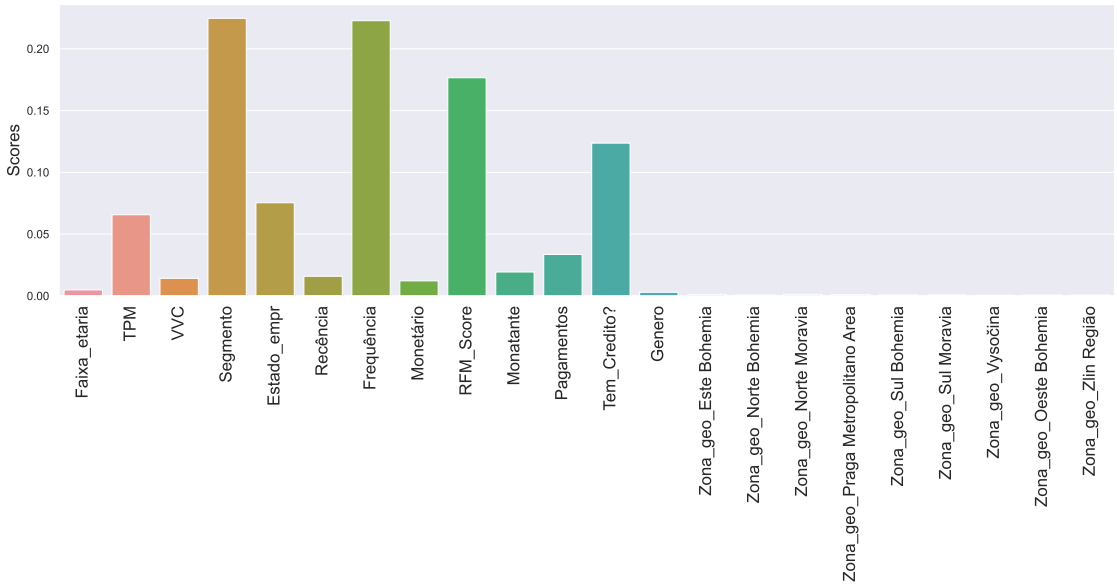

Figura 4.13: Gráfico de barras das variáveis e respetivos scores de performance obtidos pelo algoritmo *ExtraTreesClassifier*.

Treino com validação cruzada e hiperparametrização

A validação cruzada é uma técnica utilizada em Aprendizagem Automática (AA) para avaliar o desempenho de um modelo e a sua capacidade de generalizar. Nesta, o conjunto de dados disponível é dividido em várias partições denominadas "dobras"(*folds*). O modelo é treinado numa parte dos dados (k-1 *folds*) e avaliado na restante. Esse processo é repetido k vezes, alternando a *fold* de teste em cada iteração. Ao final, obtém-se uma média das métricas de desempenho calculadas em cada iteração com diferentes partições de teste e treino. Assim, é importante fazer validação cruzada no treino do nosso modelo para ser apresentado vários perfis de clientes diferentes para melhor generalizar.

Em seguida, destacam-se as vantagens de utilizar a validação cruzada:

1. **Utilização eficiente dos dados**: Ao dividir o conjunto de dados em várias partes, podemos utilizar todas as amostras tanto para treino quanto para teste, se não foram selecionadas aleatoriemente. Isto evita desperdício de dados, principalmente em conjuntos de dados pequenos;

2. **Estimativa mais precisa do desempenho**: A validação cruzada permite obter uma estimativa mais confiável do desempenho do modelo, pois leva em consideração a média das métricas calculadas em várias iterações. Assim é possível evitar estimativas enviesadas que podem ocorrer ao usar apenas uma única divisão de treino/teste;

3. **Avaliação do desempenho em diferentes configurações**: A validação cruzada permite avaliar o desempenho do modelo em diferentes configurações de treino e teste. É particularmente útil na seleção de hiperparâmetros do modelo, pois fornece uma visão geral do desempenho em diferentes cenários;

4. **Identificação de sobreajuste**: Ao avaliar o modelo em diferentes conjuntos de teste, a validação cruzada pode ajudar a identificar se o modelo sofre de *overfitting* (ajuste excessivo aos dados de treino) ou *underfitting* (incapacidade de capturar os padrões nos dados). Se o desempenho variar significativamente entre as dobras, é um indicador de *overfitting*.

Hiperparametrizar um modelo refere-se ao processo de selecionar os valores ideais dos hiperparâmetros do modelo de Aprendizagem Automática. Os hiperparâmetros são parâmetros que não são aprendidos diretamente a partir dos dados, mas sim definidos pelo utilizador antes do treino do modelo. O nosso modelo já tinha um *score* bastante bom, no entanto a hiperparametrização permite testar a sensibilidade do modelo a diferentes hiperparâmetros e regularizá-lo, ajudando a mitigar *overfitting*, verificada no nosso caso. Além disso, estas duas etapas, validação cruzada e hiperametrização podem ser feitas ao mesmo tempo. As vantagens de hiperparametrizar o modelo resumem-se a:

1. **Otimização do desempenho**: Ao ajustar os hiperparâmetros, é possível encontrar a combinação que maximiza o desempenho do modelo nos dados de teste ou validação;

2. **Controlo da complexidade do modelo/regularização**: Os hiperparâmetros podem controlar a complexidade do modelo, como a profundidade de uma árvore de decisão num algoritmo de árvore de decisão ou o número de neurónios numa rede neural. Ajustar estes hiperparâmetros permite encontrar um equilíbrio entre um modelo muito simples (que pode subajustar os dados) e um modelo muito complexo (que pode sobrejustar os dados).

O *Grid Search Cross Validation* (GDCV) e *Bayes Search Cross Validation* (BDCV) são duas técnicas utilizadas na otimização de hiperparâmetros em algoritmos de Aprendizagem Automática. Ambas as abordagens têm o objetivo de encontrar a combinação ideal de valores para os hiperparâmetros de um modelo, a fim de obter o melhor desempenho possível. A diferença fundamental reside na forma como exploram o espaço de busca dos hiperparâmetros.

O GDCV é uma técnica simples, mas exaustiva. Este requer que se especifique uma grade (ou lista) de valores para cada hiperparâmetro a otimizar. Em seguida, o GDCV realiza uma busca exaustiva em todos os valores possíveis desses hiperparâmetros, treinando e avaliando o modelo para cada combinação. Por fim, este retorna a combinação de hiperparâmetros que obteve o melhor desempenho. Por exemplo, tendo três hiperparâmetros a otimizar, cada um com cinco valores possíveis, o GDCV irá treinar e avaliar 125 modelos diferentes para encontrar a melhor combinação (5x5x5).

O BDCV utiliza uma abordagem mais inteligente e eficiente, recorrendo a um algoritmo de otimização Bayesiana para encontrar a combinação ideal de hiperparâmetros. Em vez de explorar exaustivamente todo o espaço de busca, o BDCV constrói uma função de probabilidade

(modelo probabilístico) com base nas observações anteriores de desempenho do modelo em diferentes combinações de hiperparâmetros. Essa função de probabilidade é usada para decidir quais combinações de hiperparâmetros devem ser avaliadas em seguida, com base na probabilidade de melhor desempenho. O BDCV utiliza a informação acumulada a cada iteração para refinar progressivamente a busca e se concentrar nas áreas mais promissoras do espaço de hiperparâmetros. Isto resulta em menos avaliações de modelos que o GDCV, tornando-o mais eficiente para espaços de hiperparâmetros de grande dimensão.

Realizámos a etapa de treino com validação cruzada e hiperparametrização, utilizando o algoritmo BDCV da biblioteca *Scikit-learn* (89) dado ser mais eficiente e o RFC ter 6 parâmetros. Utilizamos uma validação estratificada, com 5 *folds*. O nosso *scorer*, i.e. a métrica que queremos otimizar com a hiperparametrização será o *f1-score*, com média macro (calculado através da eq. 4.7).

$$f1_{score} = 2 \times \left(\frac{precision \times recall}{precision + recall} \right) \tag{4.7}$$

O *f1-score* é uma média harmónica entre as métricas *precision* e *recall*, sendo que esta métrica dá mais peso a classes pequenas e recompensa modelos com valores de *precision* e *recall* similares (90). Ao otimizar o *f1-score* obtemos um equilíbrio no desempenho do modelo para todas as classes e dá-se igual peso às métricas *precision* e *recall*. Isto é particularmente útil em problemas multiclasse, nos quais cada classe pode ter um papel significativo e queremos garantir que o modelo seja capaz de identificar corretamente exemplos positivos em todas as classes, independentemente de sua frequência, que é o nosso caso (91).

A hiperparametrização através de métodos de busca em rede, como o GDCV e BDCV, é uma abordagem heurística, i.e., a escolha dos parâmetros é definida pelo utilizador para o problema em questão, utilizando o seu conhecimento do domínio, e o método escolhe a combinação que produz os melhores resultados. O dicionário de parâmetros que define o espaço de busca do RFC é o seguinte:

- `n_estimators`: 5 e 10;
- `criterion`: “gini”e “entropy”;
- `max_depth`: 1 e 2;
- `min_samples_split`: 40, 50 e 60;
- `min_samples_leaf`: 30, 40 e 50;
- `class_weight`: “balanced” e “balanced_subsample”;
- `warm_start`: True;
- `bootstrap`: False;
- `max_leaf_nodes`: 4, 5 e 6.

Dada a simplicidade do conjunto de dados e do problema de classificação, o número de estimadores não necessita de ser elevado, quanto mais reduzido menor a probabildidade de *overfitting*. É comum testar-se diferentes critérios de divisão, sendo o "gini" e "entropy" os mais utilizados. Como já discutido, o equilíbrio de classes é importante em problemas de classificação desequilibrados, os valores "balanced" e "balanced_subsample" ajustam automaticamente os pesos das classes de acordo com a frequência das classes no conjunto de dados, ajudando a lidar com classes desequilibradas. Os restantes hiperparâmetros, foram escolhidos e testados para diminuir *overfitting*, por exemplo, *max_depth* define a profundidade da árvore, valores reduzidos podem diminuir *overfitting* porém o modelo pode não conseguir capturar relações complexas nos dados.

Após aplicação do método BDCV para o dicionário de hiperparâmetros, obteve-se uma média de *score* do conjunto de teste de 0.97 com um desvio padrão de 0.03, sendo o menor *score* igual 0.82. Podemos afirmar existe alguma sensibilidade aos hiperparêmetros, ainda que reduzida, dado o desvio padrão de 3%.

Os melhores parâmetros obtidos foram: *class_weight* = balanced, *criterion* = "entropy", *max_depth* = 1, *min_samples_leaf* = 50, *min_samples_split* = 50, *n_estimators* = 10 e `max_leaf_nodes` = 5. Estes resultaram nas métricas *recall*, *f1-score* e *precision* com média macro 0.97 no conjunto de teste. No conjunto de validação, obteve-se 0.95 para *precision* e 0.96 para *f1-score* e *recall*. É pertinente analisar, adicionalmente, as médias micro pois conseguimos perceber onde houve uma maior dificuldade para classificar, o grupo CAV e CP, têm *f1-score* de 0.94 e 0.95, respetivamente, existindo um total de 37 exemplos mal classificados na matriz de confusão do conjunto de validação.

Os resultados obtidos nas métricas de avaliação foram satisfatórios, as variáveis selecionadas, na sua maioria, corresponderam às utilizadas na tarefa de segmentação de clientes, pelo que serão excelentes preditores, e o modelo conseguiu identificar a sua importância e discernir entre grupos. De facto, pode-se afirmar a existência de *overfitting*, ainda que tenhamos conseguido reduzir um pouco a complexidade do modelo, pois a performance foi elevada em todos os conjuntos de dados, mesmo que ligeiramente menor no conjunto de validação, devido à natureza do problema e simplicidade dos dados utilizados. Como este é *open-source* não é possível obter mais exemplos, idealmente testaria-se o modelo contra novos dados para verificar a sua capacidade de generalização mais eficazmente.

Ainda assim, o importante nesta tarefa foi automatizar a classificação dos clientes nos grupos encontrados e caracterizados, o qual o modelo foi capaz de aprender as características específicas a cada classe. Na próxima secção, analisamos a explicabilidade do modelo.

4.3.3 Explicabilidade do modelo

Nesta secção queremos perceber melhor como o modelo funciona e chega às decisões na tarefa de classificação. Estudar a explicabilidade dos modelos é especialmente importante quando as decisões têm impacto significativo em vidas humanas, como em casos médicos, ou sistemas de tomada de decisão automáticos. Além disso, esta análise permite perceber se os modelos de

Aprendizagem Automática têm algum tipo de viés ou discriminação na tomada de decisão. Por último, a explicabilidade ajuda a aumentar a confiança nos modelos de Aprendizagem Automática. Existem várias formas possíveis de o fazer, para este trabalho abordamos duas em praticular: calculando os valores SHAP (92) e construindo uma árvore de decisão (93).

O *Shapley Additive Explanations* (SHAP) é um método baseado na teoria dos jogos cooperativos. A ideia central é atribuir um valor de SHAP a cada variável, representando a sua contribuição marginal para a previsão do modelo. Para calcular estes valores, o SHAP considera todas as combinações possíveis de variáveis e calcula a diferença na previsão do modelo quando uma variável específica está incluído ou excluído. Primeiro, o SHAP requer um valor de referência que representa uma linha base para calcular as contribuições, geralmente, é uma instância média dos dados de treino. Em seguida, o SHAP realiza uma permutação sistemática de todos as variáveis, calculando a diferença nas previsões quando uma variável é incluída ou excluída. Para cada combinação de variáveis, são avaliadas todas as possíveis atribuições de valores e calculada a contribuição marginal da variável. Com base nestas contribuições, é criado um modelo de explicação, que é uma simplificação linear dos efeitos das variáveis. Este modelo é capaz de capturar as contribuições de cada variável individualmente e como se combinam para formar a previsão final do modelo.

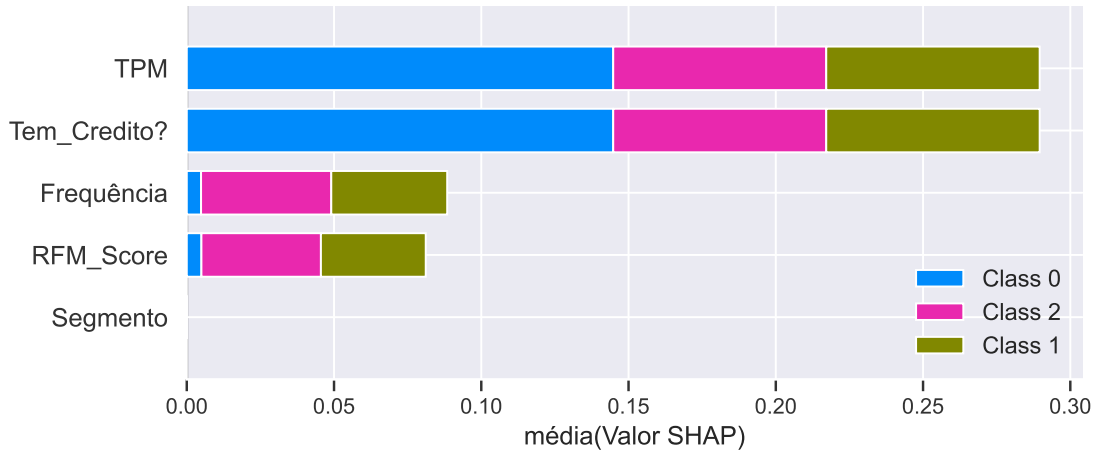

Figura 4.14: Gráfico de barras das variáveis e respetivos valores médios de SHAP, para cada grupo.

Com base na fig. 4.14 compreendemos que a variável `TPM` é a que tem mais influência, sendo utilizada para distinguir entre todos os grupos, principalmente o CE. Em segundo lugar, surge a variável `Tem_credito?`, semelhante à primeira. A variável `Segmento` revelou-se ser a menos importante na classificação, não tendo sido utilizada.

O modelo treinado originalmente foi o RFC, sendo difícil de analisar cada árvore individualmente de forma a estudar a explicabilidade do modelo. Esta é uma dificuldade inerente de um modelo de *ensemble*, que tal como as redes neuronais são considerados modelos *black-box*. No entanto, podemos estudar a versão mais simplificada, uma árvore de decisão. É possível extrair facilmente a explicabilidade de uma árvore de decisão pois tem uma estrutura intuitiva e facilmente

compreensível baseada em regras. Esta é composta por nós (representando testes em variáveis) e ramos (representando as ramificações de decisões). Cada caminho da raiz até uma folha numa árvore de decisão representa uma sequência de testes nas variáveis e as suas condições. Estes caminhos podem ser facilmente interpretados como regras *if-else*. De notar que variáveis que aparecem próximas da raiz da árvore são consideradas mais importantes, pois têm um impacto maior na divisão dos dados (ou seja uma maior redução da impureza).

Através da fig. 4.15, percebemos que a variável `RFM_score` é utilizada em primeiro lugar, com o valor 5.235 (ou seja, um cliente de médio valor). Em seguida é utilizado o `TPM` com o valor -0.486, para distinguir entre posse de crédito, e por fim as variáveis de `RFM_Score` e `Frequência`, não tendo sido utilizada a variável `Tem_crédito?`. Através dos gráficos pizza observamos algumas folhas com impureza onde o modelo classificou incorretamente algumas instâncias. Esta visualização permite-nos perceber através de regras e caminhos na árvore como o modelo está a classificar e tomar as suas decisões com base nas variáveis disponíveis. De notar que este método visual não será viável para conjuntos de dados mais complexos (com mais variáveis) ou árvores de decisão mais complexas (mais profundidade ou maior número de folhas).

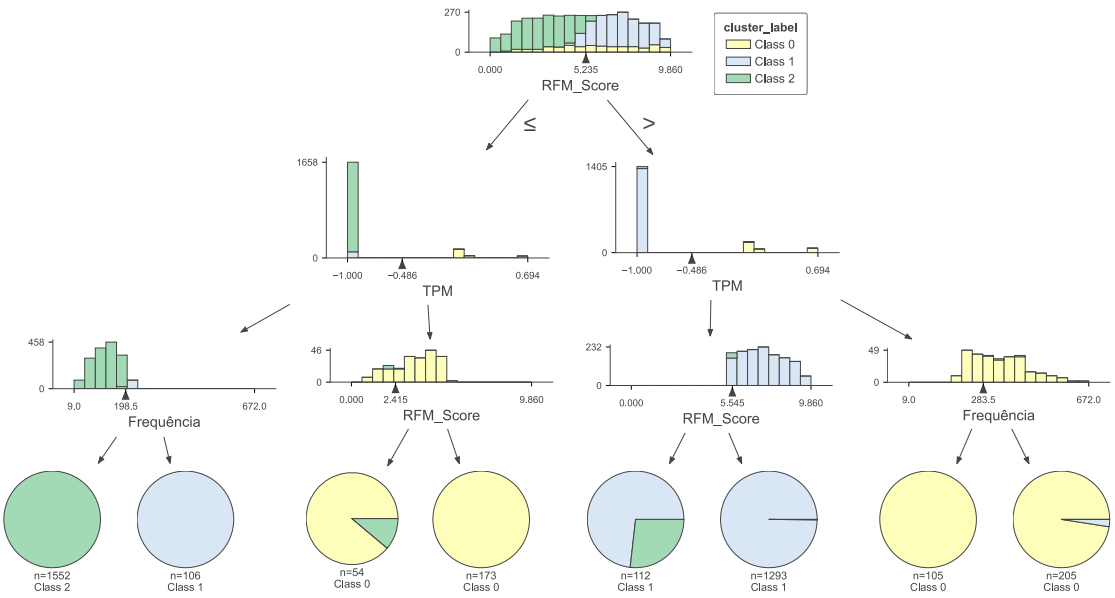

Figura 4.15: Visualização da árvore de decisão obtida com as variáveis selecionadas. É possível visualizar em cada folha as diferentes distribuições dos exemplos por classificar. Os gráficos de pizza no último nível permitem facilmente perceber a impureza das folhas. Esta árvore foi obtida utilizando a biblioteca *dtreeviz* (3).

Concluindo, através do método SHAP, nomeadamente pela análise do gráfico 4.14, percebemos que as variáveis têm diferentes importâncias na tarefa de classificação sendo a maior a da variável `TPM`, seguida de perto pela `Frequência` e a menor o `Segmento`. As variáveis relacionadas com o crédito são utilizadas principalmente para classificar o grupo CE. Recorrendo à árvore de decisão, com a fig. 4.15, conseguimos ter uma ideia de quais as regras de decisão e valores utilizadas para diferenciar clientes e grupos. A importância das variáveis revelam-se diferentes em

termos de *ranking* em relação ao método anterior; nomeadamente, a variável `RFM_Score` surge em primeiro lugar seguida de `TPM` e `Frequência`, as restantes duas nem sequer são utilizadas na classificação. Conseguimos também observar quais grupos foram mal classificados: possivelmente se tivéssemos aumentado a profundidade teríamos obtido uma classificação perfeita, no entanto, com *overfitting*.

4.4 Discussão

Ao longo deste trabalho foi desenvolvida uma metodologia, de forma exploratória, que conjuga várias técnicas de Ciência de Dados com o objetivo de identificar e caracterizar grupos de clientes bancários, através dos seus dados. Esta metodologia é composta por 4 partes distintas relacionadas entre si: 1) análise RFM, 2) agrupamento de clientes, 3) identificação de padrões nos grupos e transações e 4) automatização da classificação dos clientes nos grupos.

Na primeira fase, recorremos à análise RFM, muito utilizada na análise comportamental de clientes, para construir três variáveis (recência, frequência e valor monetário). Estas são as principais métricas que resumem o comportamento de transações de um cliente, numa determinada janela de tempo, neste caso 5 anos. Assim, foi possível conjugar as 3 variáveis num *score*, denominado *RFM_score*, que resume num só valor, entre 0 e 10, a valiosidade do cliente para o banco. Com este *score* procedeu-se à primeira segmentação dos clientes em 6 segmentos diferentes: clientes de topo, clientes de alto valor, clientes de médio valor, clientes de baixo valor, clientes perdidos e novos clientes, enriquecendo o conjunto de dados. Os clientes de topo corresponderam somente a 5% da amostra e os clientes perdidos e novos a menos de 1%. Os restantes grupos possuem um tamanho semelhante ($\sim$20%). Foi criada também uma nova métrica, de elevado interesse para o banco, o VVC que estima o potencial valor do cliente ao longo do relacionamento com o banco. Esta primeira análise serviu para a criação das variáveis a serem usadas na fase 2) e permitiu logo à partida uma segmentação fácil de executar e intuitiva de perceber.

Na segunda fase, procedeu-se ao agrupamento dos clientes utilizando modelos de Aprendizagem Automática não supervisionados (94). Para tal, testaram-se dois algoritmos diferentes: o *k-means* e o DBSCAN, que têm abordagens distintas na segmentação. No primeiro modelo, após testar vários valores de k e avaliar a qualidade dos grupos formados em função de métricas e visualizações, encontraram-se 3 grupos diferentes. O grupo CP representa clientes de baixo valor e sem atividade de empréstimo. O grupo CAV contém clientes de elevado valor sem atividade de empréstimo. O grupo CE contém clientes de variados segmentos e perfis, com atividade de empréstimo. Estes grupos estão distribuídos de forma equilibrada em relação às variáveis demográficas (género, faixa etária e zona geográfica), não existindo diferenças substanciais entre grupos. No segundo modelo, as melhores métricas resultaram em dois grupos que se dividem entre ter e não ter empréstimo. Dado os resultados do primeiro terem maior interesse procedeu-se com este para o desenvolver do restante trabalho.

A terceira fase de descoberta de padrões está dividida em duas partes: a construção de perfis e a análise sequencial de transações. Para construir os perfis recorreu-se ao algoritmo *Apri-*

ori; este encontra os *itemsets* mais frequentes no conjunto de dados. Assim, foi possível encontrar os valores que as variáveis em cada grupo tomam mais frequentemente. Escolhendo os 5 mais frequentes, com comprimento um, foi possível definir os perfis em função das variáveis demográficas, empréstimos e transações, com o respetivo suporte. Na análise sequencial de transações, construiu-se uma sequência para cada cliente composta por três variáveis (tipo de transação, categoria do valor da transação e um indicador de semana ou fim de semana). Desta forma, para cada grupo foi possível encontrar as sequências de transações mais frequentes para um dado suporte, e analisá-las em função de várias métricas bem como comparar as principais diferenças entre cada grupo.

Em todos os grupos, a zona geográfica da capital foi o mais prevalente nos perfis bem como a faixa etária da meia idade. Em termos de frequência e valor monetário conseguimos identificar os intervalos específicos de cada grupo permitindo conhecer em maior detalhe o cliente. Em relação aos balanços e montantes, na análise de transações, também se identificou uma significativa diferença entre os grupos. O grupo CE, apesar de mais valioso devido aos montantes elevados, é também o mais instável e menos conservador com o seu dinheiro, ao contrário do grupo CAV que também possui valores elevados de transações e balanços, mas menores que o grupo CE, sendo mais conservador e poupado com o seu dinheiro.

Em relação à análise sequencial de padrões encontraram-se diferenças substanciais entre transações de crédito vs débito, com o grupo CP o que apresenta a maior diferença na proporção (27% créditos e 73 % débitos). Este é o único grupo que sem transações de montante elevado a dominar as sequências frequentes encontradas e que apresenta um desiquilíbrio em relação à quantidade de crédito e de débito nas suas sequências. O grupo CE é o que efetua mais frequentemente transações de alto valor mas pode estar correlacionado com o facto de ser o único com empréstimos. Já o grupo CAV revela um forte poder económico e um perfil mais conservador de dinheiro com uma gestão financeira eficaz. Estudando as transações que os grupos efetuam nos fins de semana não se encontrou uma diferença substancial, revelando comportamentos semelhantes.

Na última fase, a classificação de clientes, testaram-se vários modelos de Aprendizagem Automática supervisionada mais utilizados na literatura com o objetivo de, dado um cliente, prever qual dos três grupos identificados e estudados pertence. Após testagem dos modelos em diferentes partições de variáveis, o melhor modelo foi o RFC, que conseguiu obter uma previsão perfeita. Com o objetivo de melhorar e regularizar o modelo final escolhido, procedeu-se à seleção de variáveis, treino com validação cruzada e hiperparametrização, chegando a um métrica *f1-score* de 0.96 no conjunto de validação, que indica existência de *overfitting*. No fim, estudou-se a explicabilidade do modelo, com recurso à análise dos valores SHAP e visualização de uma árvore de decisão. As variáveis que apresentam maior poder descriminativo são a `TPM` e `Tem_credito?`, para o primeiro modelo SHAP. Para a árvore de decisão, foram as variáveis `TPM` e `RFM_Score`.

Capítulo 5

Suporte à decisão baseada em dados

Após desenvolver a metodologia e aplicá-la ao conjunto de dados da República Checa resta disponibilizar os resultados obtidos e o modelo treinado de forma simples e efetiva. Os resultados, nomeadamente as métricas RFM, segmentos e grupos podem ser facilmente estudados por um administrador através de um *dashboard* analítico. Os nossos dados são estáticos, no entanto desenvolvemos a estrutura de um *dashboard* com esta metodologia para no futuro ser aplicada a outros conjuntos de dados semelhantes, em modo *streaming*. Uma *Application Programming Interface* (API) quando contém um modelo de AA é uma interface que permite que outros programas comuniquem com e utilizem o modelo para realizar tarefas de análise, previsão ou classificação de dados dependendo do problema, neste caso, classificação de clientes nos grupos encontrados.

5.1 *Dashboard* analítico

Um dashboard é uma interface gráfica que apresenta informações relevantes e atualizadas de forma visual e concisa sendo um painel que reúne dados de diversas fontes e os exibe em gráficos, tabelas e outros elementos visuais para facilitar a compreensão e a análise de informações complexas.

As vantagens de um *dashboard* são diversas. Este oferece uma forma eficiente de visualizar dados, permitindo identificar padrões, tendências e *insights* rapidamente. Ao consolidar informações, os tomadores de decisão podem obter uma visão panorâmica dos principais indicadores, auxiliando na definição de estratégias mais fundamentadas. Os *dashboards* podem ser conectados a fontes de dados em tempo real, garantindo que as informações estejam sempre atualizadas, refletindo assim a situação atual. A apresentação visual dos dados torna a comunicação mais clara e acessível a um público mais amplo, independentemente do nível de conhecimento técnico. Com a visualização dos dados em tempo real, é possível detectar problemas e anomalias rapidamente, permitindo uma resposta mais ágil a situações críticas.

O Power BI (95), desenvolvido pela Microsoft, é uma das ferramentas para a criação de dashboards mais utilizadas, por vários motivos. O Power BI permite a conexão com diversas fontes de dados, incluindo bancos de dados, serviços em nuvem, planilhas e muito mais, tornando-o altamente versátil. Tem também uma interface intuitiva, permitindo uma interação com esta através de movimentos de "arrastar e soltar", e tal como antes mencionado, o Power BI é acessível para uti-

lizadores sem conhecimentos avançados em programação ou análise de dados tornando-se numa das plataformas mais inovadoras e utilizadas (96). Aliado às ferramentas de visualização gráfica e ligações com dados em *real-time*, os *dashboards* criados no Power BI podem ser partilhados facilmente com outras pessoas, permitindo a colaboração e disseminação das informações de forma segura. Podem ainda ser integrados com outros produtos da Microsoft como o Power Automate ou Power Apps, é ainda possível colocar o *dashboard* num formato de ecrã de telemóvel para fácil consulta em qualquer lugar. Estas vantagens mencionadas fazem parte da característica de *self-service* do Power BI que permite que os utilizadores criem os seus próprios relatórios e *dashboards* de maneira autónoma, sem depender significativamente do departamento de TI ou especialistas em análise de dados, podendo ser facilmente alterados os filtros e métricas, consoante a necessidade, por um administrador sem conhecimentos técnicos.

O *dashboard* surge neste trabalho como uma forma de sintetizar os resultados mais importantes obtidos ao longo do desenvolvimento do trabalho e *insights* dos dados da República Checa, de uma forma clara e visualmente apelativa. Dividimos o *dashboard* em duas páginas diferentes. A primeira (ver fig. 5.1) contém uma visão mais global e generalizada do estado do banco em relação aos seus clientes com especial foco em diferentes métricas (como RFM, *churn*, entre outras), variáveis demográficas dos clientes e informação dos serviços bancários. Na segunda página (ver fig. 5.2), existe uma visão específica para um determinado cliente. O foco são as métricas específicas aos clientes e os seus comportamentos em relação às transações.

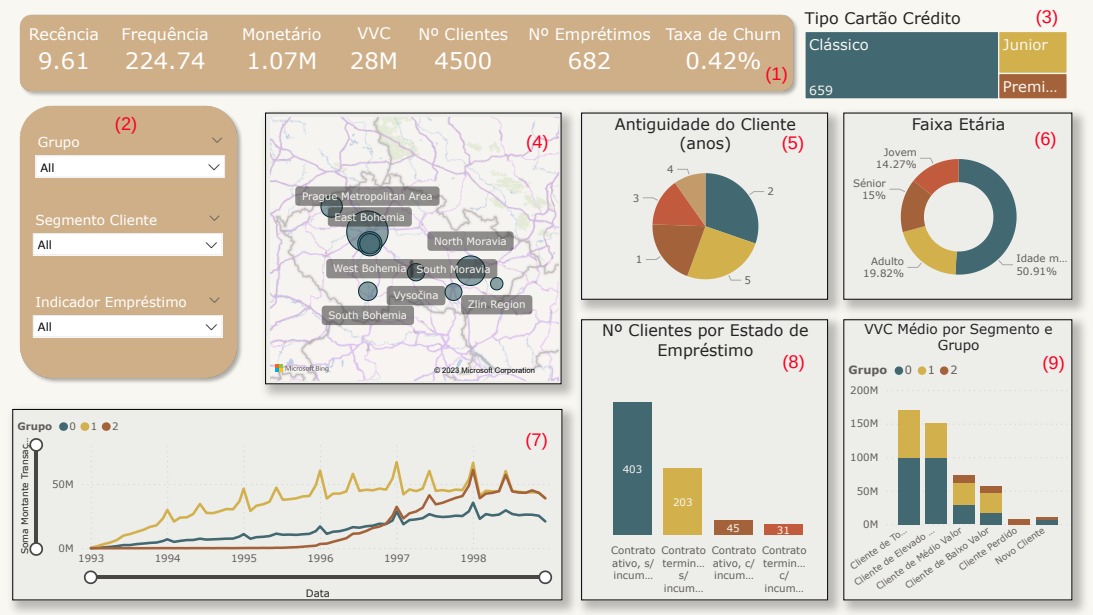

Figura 5.1: Página do *dashboard* com a uma vista geral administrativa do banco.

Na fig. 5.1 encontram-se diversos elementos como métricas, filtros e visualizações. No retângulo no canto superior **(1)** encontram-se as métricas, nomeadamente, as variáveis RFM e VVC criadas anteriormente, o número de clientes, o número de empréstimos e a taxa de *churn* de

clientes, i.e., clientes que deixaram de ser ativos com o banco. Estas métricas foram calculadas a partir da linguagem do DAX do Power BI, permitindo dinamismo nos filtros presentes na página do *dashboard*. Os filtros principais encontram-se em **(2)**, sendo que é possível filtrar pelo grupo, pelo segmento de cliente, ou um indicador de posse de empréstimo.

O primeiro gráfico **(3)** mostra a proporção de cada tipo dos cartões de crédito entre o clássico, junior e premium. É possível neste gráfico selecionar um dos retângulos e filtrar pelos clientes que possuem esse determinado cartão. Em seguida, encontra-se um mapa da República Checa **(4)**. O tamanho das bolhas é diretamente proporcional ao número de clientes em cada zona geográfica. Esta visualização foi criada utilizando uma variável hierarquica, com a zona geográfica no nível mais acima e a cidade no nível mais abaixo, sendo por isso possível utilizar a funcionalidade de *drill-down* (descer na hierarquia) ou *roll-up* (subir na hierarquia). Além disso, é possível, ao clicar em qualquer bolha, filtrar para os clientes específicos dessa zona. À direita, encontra-se um gráfico circular com a antiguidade do cliente em anos, sendo possível também filtrar ao clicar numa fatia **(5)**. À direita deste temos um gráfico donut com a distribuição de clientes em determinada faixa etária **(6)**, também passível de se filtrar. No canto inferior esquerdo, observamos o gráfico com as séries temporais do montante transacionado de cada grupo obtido **(7),** existe a possibilidade de selecionar um determinado intervalo de tempo, nos eixos do gráfico. Utilizou-se uma nova hierarquia com a variável do tipo data, criada automaticamente pelo Power BI, permitindo utilizar a funcionalidade de *drill-down* e *roll-up* com as variáveis temporais como o trimestre, mês, semana e dia. Seguidamente, à direita, o gráfico **(8)** resume o número de clientes em função do estado do seu empréstimo, sendo possível filtrar pelo estado do empréstimo. Por fim, o gráfico **(9)** mostra a distribuição da métrica `vvc`, de elevado interesse para o banco, em função do segmento e grupo, permitindo resumir a valiosidade futura dos clientes nestes grupos.

A fig. 5.2 contém a segunda página do *dashboard* com uma vista específica ao cliente. No primeiro retângulo **(10)** é possível escolher um cliente através do seu número de conta, esse número é utilizado para filtrar os restantes gráficos. Também é possível filtrar para uma janela temporal específica, usando o filtro da data que se encontra abaixo. Na imagem encontra-se o filtro com uma data relativa (sendo esta a data máxima presente nos dados),, sendo possível alterar a mesma. No retângulo à direita **(13)** encontram-se mais informações do cliente como o seu grupo e segmento bem como o valor específico das suas métricas RFM. Estes valores são, tal como já mencionámos, dinâmicos, e podem alterar-se consoante os filtros escolhidos. Seguidamente, o gráfico em cascata **(11)** foi escolhido pois mostra o fluxo dos movimentos do cliente ao longo do tempo, sendo possível visualizar facilmente descidas ou subidas nos seus montantes transacionados bem como descer ou subir na hierarquia da variável data. O retângulo central **(12)** contém mais algumas informações do cliente, nomeadamente a atividade de empréstimo bem como uma visualização de Gauge que representa o valor `VVC` do cliente escolhido, comparando este relativamente aos valores definidos como mínimo e máximo. O mínimo é o valor mínimo da variável `VVC` , já o valor máximo, dado existir extremos bastante distantes da norma, foi escolhido como o percentil 0.80, após análise de um gráfico de bigodes da variável.

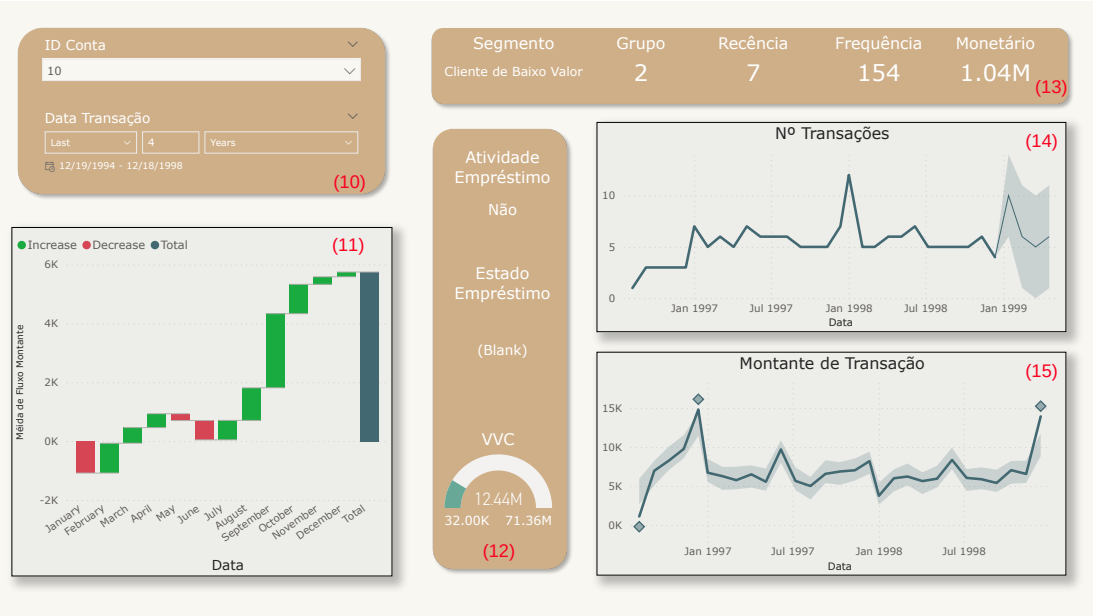

Figura 5.2: Página do *dashboard* com a uma vista específica a um cliente para um analista de negócio.

Os últimos dois gráficos contêm informações somente relativas às transações. O gráfico **(14)** é uma série temporal do número de transações. Este contém ainda uma cauda de previsão obtida utilizando o método do Power BI, com a opção de sazonalidade de 12 pontos (i.e. 12 meses), a previsão é dos próximos 4 pontos (i.e. 4 meses) com um intervalo de confiança de 95%. Diretamente abaixo, o gráfico **(15)** contém o montante (em módulo) transacionado do cliente. Este utiliza também o método do Power BI para a deteção de anomalias em séries temporais, com um parâmetro de sensibilidade de 75%. As anomalias são identificadas com os pontos em diamante.

Em conjunto, estas duas páginas que compõem o *dashboard* contêm 2 vistas que permitem não só conhecer mais detalhadamente o cliente e o seu comportamento transacional ao longo do tempo como também analisar a situação do banco através das métricas criadas.

5.2 Implementação do modelo

A última fase do ciclo de um projeto em CD é a implementação ou disponibilização do modelo de AA treinado. Existem várias formas de implementar um modelo, uma das principais é através da criação de uma API, que permite aceder ao modelo remotamente ou incorporá-lo diretamente em aplicações ou sistemas. Também podem ser disponibilizados em seviços na nuvem com *endpoints* específicos. No contexto de AA, uma API possibilita a disponibilização de modelos treinados sem precisar de os construir do zero. Assim podem fazer pedidos à API para enviar dados e receber previsões ou resultados, tornando o processo de integração de modelos muito mais rápido e eficiente. As principais vantagens de usar API's são a fácil integração com sistemas e ambientes e a abstração da complexidade.

Neste caso, criámos uma API usando a biblioteca *FastAPI* (97). Esta é uma biblioteca específica ao desenvolvimento de APIs em Python, projetada para facilitar a criação de serviços *web* rápidos, simples e práticos.

Tabela 5.1: Descrição dos *endpoints* da API criada e a respetiva descrição e especificação da entrada e saída.

Endpoint	**Descrição**	**Entrada**	**Saída**
/calculate_rfm	Calcula os valores RFM.	Dados em formato JSON.	Dados RFM calculados.
/predict_group	Realiza a previsão de grupo.	Dados da classe **PredictionData** em JSON.	ID da conta e grupo previsto como resposta.

Dado que as métricas de RFM são de elevado interesse para um banco criou-se a motivação para ter o primeiro *endpoint* na tabela 5.1, "/calculate_rfm". Partindo de um ficheiro json com as transações do cliente, inclusive o montante, tipo de transação e data da transação, é retornado o valor resultante dos cálculos das variáveis RFM, com base numa data de referência especificada no *endpoint*.

A classe definida pelo modelo **PredictionData**, especificada no segundo *endpoint* na tabela 5.1, especifica a entrada da API para efetuar a previsão com a seguinte estrutura que define os parâmetros esperados: `Frequência` (int), `RFM_Score` (float), `Segmento` (int), `Tem_credito?` (float), `TPM` (float) e `Id_conta` (int), as mesmas do modelo treinado na secção 4.3.

Na tabela 5.1, o *endpoint* "/predict_group"efetua a previsão do grupo utilizando o modelo RFC, com base nos dados do cliente. O modelo RFC foi guardado utilizando a biblioteca *pickle*. Esta consegue salvar modelos treinados, transformadores de dados e outros objetos presentes no fase de treino do modelo, sendo uma ferramenta útil para preservar o estado do modelo ou dos seus componentes para uso posterior. Além disso, armazenou-se também o transformador *min max* utilizado durante a fase de treino do modelo. Assim, este *endpoint* requer um ficheiro json com os valores das 5 variáveis utilizadas no modelo do cliente (adicionalmente, um identificador do mesmo), e procede, internamente, ao pré-processamento com reescalonamento dos dados e classificação do grupo, retornando o número do grupo previsto.

5.3 Discussão final e recomendações

Implementámos através de uma API o modelo treinado na fase 4) que permite classificar um cliente com base nas suas transações e calcular as variáveis RFM. Por último, construiu-se um *dashboard* que permite analisar os dados utilizados no desenvolvimento do trabalho com duas visões diferentes: uma global que reflete o estado do banco, e outra específica a um cliente com as métricas criadas em função dos segmentos construídos e grupos identificados na fase 1) e 2). De seguida, respondemos às perguntas de investigação colocadas na secção 1.2.

Quais são os principais atributos e características dos diferentes segmentos de clientes?

Concluímos que as variáveis RFM que resumem o comportamento do cliente relativo às suas transações bem como variáveis respetivas aos serviços/produtos bancários que estes clientes possuem permitem uma segmentação eficaz destes em diferentes grupos, com a identificação de segmentos de clientes com caracteríticas específicas e diferentes necessidades. Ao contrário dos métodos tradicionais de *mass market* ou segmentação baseada em variáveis demográficas como a zona geográfica e o género demonstraram não ser discriminativas para os grupos encontrados.

De que forma as técnicas aplicadas contribuem para a segmentação eficaz dos clientes e a criação de perfis financeiros?

Através da aplicação de modelos não supervisionados foi possível detetar no conjunto de dados dos clientes bancários 3 grupos de clientes que se diferenciam com base no seu valor, atividade de empréstimo e interesse para o banco. Com o modelo supervisionado e disponibilização deste através de uma API é possível automatizar o processo de classificação de clientes para um conjunto de dados não estático em que os clientes podem evoluir e alterar os seus comportamentos ao longo do seu relacionamento com o banco. A identificação de padrões, por sua vez, permitiu em primeiro lugar definir os perfis mais frequentes que resumem os grupos em 10 variáveis. A análise sequencial de padrões permitiu estudar a relação com o dinheiro dos diferentes grupos, se mais conservadores ou se mais dispendiosos, bem como em que altura da semana gastam mais dinheiro, se durante a semana ou durante o fim de semana.

Com a metodologia desenvolvida quais recomendações podem ser efetuadas para cada grupo identificado?

Com os grupos encontrados e análises efetuadas é possível inferir um conjunto de recomendações para o banco adoptar:

- **Grupo CE**: Focar no desenvolvimento de ferramentas *online* ou aplicações móveis que forneçam funcionalidades de gestão de empréstimos, como calculadoras de pagamento, planos de reembolso personalizados e lembretes de pagamento, integradas nos serviços do banco. Oferecer recursos educacionais e *workshops* sobre empréstimos responsáveis, gestão de dívidas e planeamento financeiro para ajudá-los a tomar decisões informadas e melhorar o seu bem-estar financeiro. Adotar uma abordagem proativa para ajudar os clientes a gerir seus empréstimos de forma eficaz. Fornecer orientação e aconselhamento financeiro para garantir créditos bem-sucedidos e pontuais. Identificar clientes dentro deste grupo que podem beneficiar de opções de refinanciamento, consolidação de dívidas ou extensões de empréstimos.

- **Grupo CAV**: Oferecer uma experiência de cliente personalizada e de alto valor para que estes se sintam valorizados e apreciados. É possível, por exemplo, indicar gestores de contas ou representantes do atendimento ao cliente dedicados para atender as suas necessidades

específicas. Oferecer benefícios, recompensas e privilégios exclusivos, como acesso prioritário a novos produtos ou serviços, eventos exclusivos ou recomendações de produtos personalizadas. Identificar oportunidades relevantes de venda cruzada e *upselling* com base no histórico de transações e preferências. Deve-se priviligiar a construção de relacionamentos de longo prazo com este grupo por meio de comunicação regular, coleção de *feedback* e interações personalizadas.

- **Grupo CP**: Desenvolver campanhas de *marketing* focadas em apresentar novos produtos e serviços, destacando os benefícios e recursos que estejam alinhados com as necessidades e preferências, por exemplo dentro da faixa etária. Algumas sugestões são um crédito pessoal aos mais jovens ou contas poupanças, um crédito automóvel ou habitacional ao grupo com idade média, ou um plano poupança de reforma. Implementar um programa de fidelidade que recompense transações frequentes e ofereça ofertas e descontos exclusivos. Criar conteúdo informativo, como posts de *blog*, vídeos ou *webinars*, que eduquem este grupo sobre educação financeira, opções de investimento e gestão inteligente do dinheiro.

Capítulo 6

Conclusões e trabalho futuro

Os clientes são o motor de qualquer negócio, desempenhando um papel fundamental no sucesso ou fracasso de uma empresa. Sem clientes, qualquer negócio deixaria de existir. Consequentemente, a fidelização e retenção dos mesmos são das estratégias mais prioritárias para o sucesso de uma empresa a longo prazo. O uso inteligente de dados tem desempenhado um papel crucial para entender os clientes e as suas necessidades e preferências. Por sua vez, compreender o comportamento do cliente e atender às suas expectativas é essencial para criar uma base sólida de consumidores fiéis satisfeitos, que irão impulsionar o crescimento da empresa, para novamente inovar e os atender de melhor forma, propagando este ciclo no tempo. A área de *Customer Intelligence* é responsável por estudar os dados dos clientes de forma a tomar decisões mais informadas e planear estratégias eficazes. Para um banco, é importante perceber que clientes possuem maior risco e quais são os mais valiosos. Academicamente existem poucos trabalhos compreensivos de *Customer Intelligence* na indústria bancária disponíveis. Com isto, de forma a ajudar a colmatar esta limitação, neste trabalho, foi explorada uma nova metodologia numa perspetiva de *Knowledge Discovery in Databases* com o objetivo de analisar os dados dos clientes bancários e extrair informações interessantes para um administrador bancário ou responsável de *marketing*.

O foco principal foi o estudo do comportamento em relação às transações dos clientes, com base na análise RFM, e deteção de grupos semellhantes no conjunto de dados. Assim foi possível identificar perfis e treinar um modelo que consiga classificar um cliente no seu grupo. Recorrendo ao método *k-means*, identificaram-se 3 grupos: um grupo de baixo valor sem atividade de empréstimo, de alto valor sem atividade de empréstimo e de valor diversificado com atividade de emoréstimo. Este "valor" é definido com base nos KPIs, RFM e VVC, bastante interessantes para um banco. Em seguida, foi também estudado os padrões nas transações para estudar o nível de risco e perceber a relação do cliente com o seu dinheiro. Além disso, criou-se um *dashboard* para um indíviduo poder explorar livremente estes grupos, consoante a necessidade.

Esta metodologia permitiu encontrar grupos de interesse para o banco com base na análise RFM, muito utilizada para efetuar a segmentação de clientes na literatura. Adicionalmente, utilizámos estas métricas em conjunto com outras variáveis, como as dos empréstimos, numa perspetiva de ciência de dados para nos permitir segmentar os clientes de uma forma mais detalhada e profunda. Ainda mais, com a análise detalhada de padrões de cada grupo específico foram cri-

ados perfis, que resumem as características mais recorrentes dentro de cada grupo, e analisados os padrões sequenciais com o objetivo de estudar melhor os comportamentos das transações, indo além da análise RFM. Assim, foi possível conhecer com maior detalhe estes grupos e clientes e caracterizá-los de acordo com o seu comportamento, baseado em dados. O *dashboard* é o elemento que finaliza a metodologia e permite disponibilizar o trabalho desenvolvido numa análise clara, objetiva e flexível, usualmente deixado de parte em trabalhos académicos.

Os dados utilizados para desenvolver esta metodologia foram *open-source* com apenas algumas variáveis, não correspondendo à realidade atual (dado serem da década de 90s) e os serviços e produtos disponibilizados serem já bastante diferentes e digitais. Assim, esta foram duas limitações do trabalho: um conjunto de dados simplificado que não corresponde à realidade da informação disponível nos sistemas informáticos dos bancos mais atuais e também os produtos e serviços serem hoje em dia mais diversificados e complexos, como a interação do utilizador com diferentes canais, como o telemóvel e a internet, bem como a multitude de outros serviços que o banco disponibiliza, como seguros, cartões e carteiras virtuais, assistentes virtuais, serviços externos subscritos, categorias de transações etc. Ainda assim, esta metodologia é flexível e permite integrar, por exemplo, outras variáveis no agrupamento, respetivas a diferentes produtos bancários, bem como analisar por exemplo as variáveis RFM e padrões sequenciais nos diferentes canais do banco. Com mais informação do cliente será possível aprofundar ainda mais as análises efetuadas com esta metodologia.

Para finalizar, abordamos o trabalho futuro. Primeiramente, poder-se-iam criar e testar novas variáveis relevantes que reflitam a evolução temporal do cliente, como, por exemplo, a variação percentual semanal/mensal das variáveis RFM. Seguidamente, seria interessante replicar esta metodologia num conjunto de dados *streaming* atual e com mais informações dos clientes como outros produtos bancários ou a sua interação com os diferentes canais bancários. Dado que a INM, a empresa parceira deste trabalho, tem como principal foco de atuação o país de Angola, poder-se-ia estudar os clientes angolanos. Os dados utilizados neste trabalho foram provenientes da República Checa devido à falta de fornecimento em tempo útil dos dados relativos a Angola por parte da INM, daí a metodologia desenvolvida ter uma perspetiva generalista relativa. Um problema interessante de explorar específico a Angola é a dificuldade da adesão ao *Internet Banking* (IB). A transição para o digital em Angola tem sido rápida e prevê-se uma evolução do mercado de pagamentos eletrónicos exponencial, apesar da preferência hoje em dia ser transações em dinheiro. Existem estudos que comprovam a recetividade da população angolana ao IB, no entanto só 27% da população tem acesso à Internet (98), sendo que, em 2020, 6% dos luandeses (99) utilizam este serviço. Como tal, replicar a metodologia com o fim de criar perfis de adoção e utilização do serviço IB pode levar a novas conclusões. Neste sentido, foi desenvolvido um questionário (ver anexo A.4), o qual não foi possível realizar em tempo útil, que permitiria avaliar os fatores determinantes da adoção do IB em Angola. Este ajudaria a identificar perfis de adoção que, em conjunto com a análise das informações bancárias dos clientes, permitem criar estratégias eficazes e personalizadas, com o objetivo final de impulsionar a transição para o digital da banca neste país.

Abreviaturas

AA Aprendizagem Automática. x, xviii, 1, 3, 5, 6, 8, 10, 19, 23, 31, 32, 35, 57–61, 63, 64, 67, 69–71, 74

API *Application Programming Interface*. xviii, 4, 71, 74–76

BDCV *Bayes Search Cross Validation*. 64–66

CAV Clientes de alto valor. xvii, 47, 52–54, 56, 57, 66, 69, 70, 76

CD Ciência de Dados. 1, 3, 69, 74

CE Clientes com empréstimos. xvii, 47, 52–54, 56, 57, 67–70, 76

CI *Customer Intelligence*. 2, 40, 79

CP Clientes com potencial. xvii, 47, 52–57, 66, 69, 70, 77

DBSCAN *Density-Based Spatial Clustering of Applications with Noise*. xiv, 12, 48–50, 69

GDCV *Grid Search Cross Validation*. 64, 65

GFP Gestor de Finanças Pessoais. 2

GNB *Gaussian Naive Bayes*. xviii, 58, 59

HCA *Hierarchical Clustering Aglomerative*. xiii, 11, 12, 42, 44

HSBC Corporação Bancária de Hong Kong e Xangai. 3

IA Inteligência Artificial. 3–6, 8–10, 18, 19

IB *Internet Banking*. 80

INM *Innovation Makers*. 1, 80

KPIs *Key Performance Indicators*. 40, 79

PD Prospeção de Dados. 3, 10, 14, 17, 35, 54

RFC *Random Forest Classifier*. xviii, 58–60, 65, 67, 70, 75

RFM *Recency-Frequency-Monetary*. ix, 18, 36, 37, 39, 40, 59, 62, 69, 71–73, 75, 76, 79, 80

RL Regressão Logística. xviii, 58, 59

SHAP *Shapley Additive Explanations*. 67, 68, 70

TPA Terminal de Pagamento Automático. 8, 54

VVC Valor Vitalício do Cliente. 40, 41, 69, 79

XGB *XGBoostClassifier*. xviii, 17, 58–60

Bibliografia

[1] "dbdiagram.io - database relationship diagrams design tool." https://dbdiagram.io/home. (Accessed on 09/21/2023).

[2] A. Amini, T. Wah, and H. Saboohi, "On density-based data streams clustering algorithms: A survey," *Journal of Computer Science and Technology*, vol. 29, pp. 116–141, 01 2014.

[3] T. Parr and P. Grover, "Dtreeviz: Decision tree visualization," *GitHub repository*, 2020.

[4] "Map of the czech republic — czech republic regions — rough guides — rough guides." https://www.roughguides.com/maps/europe/czech-republic/. (Consultado em 08/31/2023).

[5] V. Kostin and L. Halounová, "An analysis of spatial structure of urban regional networks using gis," *Acta Polytechnica*, vol. 59, pp. 35–41, 02 2019.

[6] "Xgboost documentation — xgboost 1.7.6 documentation." https://xgboost.readthedocs.io/en/stable/. (Consultado em 09/11/2023).

[7] "sklearn.ensemble.randomforestclassifier — scikit-learn 1.3.0 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html. (Consultado em 09/11/2023).

[8] "sklearn.naive_bayes.gaussiannb — scikit-learn 1.3.0 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html. (Consultado em 09/11/2023).

[9] "sklearn.linear_model.logisticregression — scikit-learn 1.3.0 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html. (Consultado em 09/11/2023).

[10] N. Arora, D. Ensslen, L. Fiedler, W. W. Liu, K. Robinson, E. Stein, and G. Schüler, "The value of getting personalization right—or wrong—is multiplying," 11 2021.

[11] L. Cao, "Data science: A comprehensive overview," *ACM Comput. Surv.*, vol. 50, jun 2017.

[12] "Inm website." https://www.inm.pt/, 2021. (Consultado em 12/26/2022).

[13] "Personal financial management (pfm)." https://www.investopedia.com/personal-financial-management-pfm-5181311, 2022. (Consultado em 12/26/2022).

[14] "What is pfm, and what is it good for? — tink blog." https://tink.com/blog/open-banking/what-is-pfm/, 2020. (Consultado em 12/26/2022).

[15] J. Lochy, "Pfm, bfm, financial butler, financial cockpit... - will the cumbersome administrative tasks on your financials finally be taken over?." https://www.linkedin.com/pulse/pfm-bfm-financial-butler-cockpit-cumbersome-tasks-your-joris-lochy/, 2019. (Consultado em 12/26/2022).

[16] "Hsbc is using ai to personalize its rewards program - tearsheet." https://tearsheet.co/artificial-intelligence/hsbc-is-using-ai-to-personalize-its-rewards-program/. (Consultado em 08/23/2023).

[17] "Bank of england - prudential regulation authority rulebook." https://www.prarulebook.co.uk/rulebook/Glossary/FullDefinition/52212/26-12-2022, 2020. (Consultado em 12/26/2022).

[18] "Europe's migration to digital services during covid-19 — mckinsey." https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/europes-digital-migration-during-covid-19-getting-past-the-broad-trends-and-averages, 2020. (Consultado em 12/26/2022).

[19] J. Ferreira, L. Fiedler, C. Giovine, L. Herring, and M. Kansal, "Getting personal: How banks can win with consumers," 7 2022.

[20] "Customers in the spotlight: How fintech is reshaping banking." https://www.pwc.com/gx/en/industries/financial-services/publications/fintech-is-reshaping-banking.html. (Consultado em 08/23/2023).

[21] "The challenge of innovation in financial services — financier worldwide." https://www.financierworldwide.com/the-challenge-of-innovation-in-financial-services. (Consultado em 08/23/2023).

[22] A. Adarkar, S. Cantú, K. Dallerup, V. Giudici, E. Lucchinetti, and Z. Orlando, "Reshaping retail banks: Enhancing banking for the next digital age," 10 2022.

[23] "The future of banking: A $20 trillion opportunity — mckinsey." https://www.mckinsey.com/industries/financial-services/our-insights/the-future-of-banks-a-20-trillion-dollar-breakup-opportunity. (Consultado em 08/23/2023).

[24] K. C. I. Libby Wells, Nell McPherson, "Brick-and-mortar banks vs. online banks: Pros and cons — bankrate." https://www.bankrate.com/banking/savings/online-vs-brick-and-mortar-banks/. (Accessed on 09/21/2023).

[25] "Digital transformation in banking: Why it's time." https://www.cascade.app/blog/digital-transformation-banking, 2022. (Consultado em 12/26/2022).

[26] "Rbs uses analytics to make customer service more than just a slogan." https://www.forbes.com/sites/tomgroenfeldt/2018/05/03/rbs-uses-analytics-to-make-customer-service-more-than-just-a-slogan/?sh=4110d5172108, 2018. (Consultado em 12/26/2022).

[27] "The wonderful big data strategy at royal bank of scotland." https://www.forbes.com/sites/bernardmarr/2016/04/13/the-wonderful-big-data-strategy-at-royal-bank-of-scotland/?sh=3f02212f788b, 2016. (Consultado em 12/26/2022).

[28] C. Essvale, *Business Knowledge for IT in Global Retail Banking: The Complete Handbook for IT Professionals*. Bizle professional series, Essvale Corporation, 2011.

[29] R. I. Harinder Hari and B. Sampat, "Customer brand engagement through chatbots on bank websites– examining the antecedents and consequences," *International Journal of Human–Computer Interaction*, vol. 38, no. 13, pp. 1212–1227, 2022.

[30] A. Černá, E. Tošovská, and P. Cetkovský, "Chapter 16 - economic transformation and the environment," in *The Czech Republic and Economic Transition in Eastern Europe* (J. SVEJNAR, ed.), pp. 377–394, Boston: Academic Press, 1995.

[31] S. Barisitz and O. Nationalbank, "Banking transformation 1980-2006 in central and eastern europe–from communism to capitalism," *South-Eastern Europe Journal of Economics*, vol. 2, pp. 161–180, 01 2009.

[32] J. Tošovský, *22 Restructuring the Banking Sector: The Case of the Czech Republic*, p. ch056. USA: International Monetary Fund.

[33] J. McCarthy, "What is artificial intelligence?," 2007.

[34] K. Didur, "Machine learning in finance: Why, what & how — by konstantin didur — towards data science." https://towardsdatascience.com/machine-learning-in-finance-why-what-how-d524a2357b56, 2018. (Consultado em 12/26/2022).

[35] H. Hassani, X. Huang, and E. Silva, "Digitalisation and big data mining in banking," *Big Data and Cognitive Computing*, vol. 2, no. 3, p. 18, 2018.

[36] "The future of payments in africa." https://www.mckinsey.com/industries/financial-services/our-insights/the-future-of-payments-in-africa, 2022. (Consultado em 12/26/2022).

[37] M. Biallas and F. O'Neill, "Artificial intelligence innovation in financial services," 2020.

[38] L. Ryll, M. E. Barton, B. Z. Zhang, R. J. McWaters, E. Schizas, R. Hao, K. Bear, M. Preziuso, E. Seger, R. Wardrop, *et al.*, "Transforming paradigms: A global ai in financial services survey," 2020.

[39] E. Digalaki, "The impact of artificial intelligence in the banking sector & how ai is being used in 2020," *online]. Datum pristupa dokumentu*, vol. 27, no. 6, p. 2020, 2019.

[40] J. M. T. d. Neves, "The impact of artificial intelligence in banking," 2022.

[41] "Artificial intelligence applications in financial services." https://dataconomy.com/2022/11/artificial-intelligence-applications-in-financial-services/. (Consultado em 12/26/2022).

[42] M. A. Scridon *et al.*, "Understanding customers-profiling and segmentation," *Management & Marketing-Craiova*, no. 1, pp. 175–184, 2008.

[43] P. A. Ionut, "Evolution of customers'segmentation techniques in retail banking.," *Annals of'Constantin Brancusi'University of Targu-Jiu. Economy Series*, 2017.

[44] V. Mihova and V. Pavlov, "A customer segmentation approach in commercial banks," in *AIP conference proceedings*, vol. 2025, p. 030003, AIP Publishing LLC, 2018.

[45] Y. S. Patel, D. Agrawal, and L. S. Josyula, "The rfm-based ubiquitous framework for secure and efficient banking," in *2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH)*, pp. 283–288, IEEE, 2016.

[46] F. Abdi and S. Abolmakarem, "Customer behavior mining framework (cbmf) using clustering and classification techniques," *Journal of Industrial Engineering International*, vol. 15, no. 1, pp. 1–18, 2019.

[47] M. Aryuni, E. D. Madyatmadja, and E. Miranda, "Customer segmentation in xyz bank using k-means and k-medoids clustering," in *2018 International Conference on Information Management and Technology (ICIMTech)*, pp. 412–416, IEEE, 2018.

[48] S. Ren, Q. Sun, and Y. Shi, "Customer segmentation of bank based on data warehouse and data mining," in *2010 2nd IEEE International Conference on Information Management and Engineering*, pp. 349–353, IEEE, 2010.

[49] E. A. E. Dawood, E. Elfakhrany, and F. A. Maghraby, "Improve profiling bank customer's behavior using machine learning," *IEEE Access*, vol. 7, pp. 109320–109327, 2019.

[50] M. A. Farajian and S. Mohammadi, "Mining the banking customer behavior using clustering and association rules methods," *International Journal of Industrial Engineering & Production Research*, vol. 21, no. 4, pp. 239–245, 2010.

[51] M. K. Sharahi and M. Aligholi, "Classify the data of bank customers using data mining and clustering techniques (case study: Sepah bank branches tehran)," 2015.

[52] M. R. Kumar, J. Venkatesh, and A. Rahman, "Data mining and machine learning in retail business: developing efficiencies for better customer retention," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–13, 2021.

[53] J. Watada and K. Yamashiro, "A data mining approach to consumer behavior," in *First International Conference on Innovative Computing, Information and Control-Volume I (ICICIC'06)*, vol. 2, pp. 652–655, IEEE, 2006.

[54] S. Moradi and F. Mokhatab Rafiei, "A dynamic credit risk assessment model with data mining techniques: evidence from iranian banks," *Financial Innovation*, vol. 5, no. 1, pp. 1–27, 2019.

[55] P. Nixon and E. Gilbert, "Unsupervised machine learning to reveal south african risk behaviour archetypes in the domain of discretionary investment decisions," *Journal of Behavioral and Experimental Finance*, p. 100757, 2022.

[56] V. L. Miguéis, A. S. Camanho, and J. F. e Cunha, "Customer data mining for lifestyle segmentation," *Expert Systems with Applications*, vol. 39, no. 10, pp. 9359–9366, 2012.

[57] J. Yang, J. Zhao, F. Wen, and Z. Dong, "A model of customizing electricity retail prices based on load profile clustering analysis," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3374–3386, 2018.

[58] F. Wang, K. Li, N. Duić, Z. Mi, B.-M. Hodge, M. Shafie-khah, and J. P. Catalão, "Association rule mining based quantitative analysis approach of household characteristics impacts on residential electricity consumption patterns," *Energy conversion and management*, vol. 171, pp. 839–854, 2018.

[59] V. Sevinç *et al.*, "A classification of the banks in turkey with bayesian cluster analysis based on mixture models," *Eurasian Econometrics, Statistics and Emprical Economics Journal*, vol. 2, no. 2, pp. 16–24, 2015.

[60] M. Hassan and M. Tabasum, "Customer profiling and segmentation in retail banks using data mining techniques," *International journal of advanced research in computer science*, vol. 9, no. 4, pp. 24–29, 2018.

[61] M. Benassi, S. Garofalo, F. Ambrosini, R. P. Sant'Angelo, R. Raggini, G. De Paoli, C. Ravani, S. Giovagnoli, M. Orsoni, and G. Piraccini, "Using two-step cluster analysis and latent class cluster analysis to classify the cognitive heterogeneity of cross-diagnostic psychiatric inpatients," *Frontiers in Psychology*, vol. 11, p. 1085, 2020.

[62] F. Ecer, "Comparing the bank failure prediction performance of neural networks and support vector machines: The turkish case," *Economic research-Ekonomska istraživanja*, vol. 26, no. 3, pp. 81–98, 2013.

[63] M. Abduh, Z. Dahari, and M. A. Omar, "Bank customer classification in indonesia: Logistic regression vis-à-vis artificial neural networks," *World Applied Sciences Journal*, vol. 18, no. 7, pp. 933–938, 2012.

[64] X. T. Pham and T. H. Ho, "Using boosting algorithms to predict bank failure: An untold story," *International Review of Economics & Finance*, vol. 76, pp. 40–54, 2021.

[65] C. Apté and S. Weiss, "Data mining with decision trees and decision rules," *Future generation computer systems*, vol. 13, no. 2-3, pp. 197–210, 1997.

[66] "Home page of pkdd discovery challenge." https://sorry.vse.cz/~berka/challenge/PAST/. (Consultado em 08/26/2023).

[67] "Bank customer segmentation (1m+ transactions) — kaggle." https://www.kaggle.com/datasets/shivamb/bank-customer-segmentation. (Consultado em 08/26/2023).

[68] "Credit card fraud detection — kaggle." https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud. (Consultado em 08/26/2023).

[69] "Bank marketing data - dataset by data-society — data.world." https://data.world/data-society/bank-marketing-data. (Consultado em 08/26/2023).

[70] "Atm banking - dataset by dcopendata — data.world." https://data.world/dcopendata/atm-banking. (Consultado em 08/26/2023).

[71] I. Bilbao and J. Bilbao, "Overfitting problem and the over-training in the era of data: Particularly for artificial neural networks," in *2017 eighth international conference on intelligent computing and information systems (ICICIS)*, pp. 173–177, IEEE, 2017.

[72] L. A. Bakker, "Know your customers with rfm. in this blog post we'll show you how to… — by leif arne bakker — towards data science." https://towardsdatascience.com/know-your-customers-with-rfm-9f88f09433bc. (Accessed on 09/21/2023).

[73] A. Jain, "What is customer lifetime value (clv), and how to calculate it." https://www.gartner.com/en/digital-markets/insights/what-is-customer-lifetime-value. (Consultado em 09/01/2023).

[74] "What is customer lifetime value (clv) & how to calculate it." https://delighted.com/blog/customer-lifetime-value-formula. (Consultado em 09/01/2023).

[75] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[76] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974.

[77] P. Brus, "Clustering: How to find hyperparameters using inertia — by patrick brus — towards data science." https://towardsdatascience.com/clustering-how-to-find-hyperparameters-using-inertia-b0343c6fe819. (Consultado em 09/05/2023).

[78] J. Sander, M. Ester, H. Kriegel, and et al., "Density-based clustering in spatial databases: The algorithm gdbscan and its applications," *Data Mining and Knowledge Discovery*, vol. 2, pp. 169–194, 1998.

[79] N. Rahmah and I. S. Sitanggang, "Determination of optimal epsilon (eps) value on dbscan algorithm to clustering data on peatland hotspots in sumatra," *IOP Conference Series: Earth and Environmental Science*, vol. 31, p. 012012, jan 2016.

[80] W. A. Fuller, *Sampling statistics*. John Wiley & Sons, 2011.

[81] "Redes neurais multiclasse: Softmax — machine learning — google for developers." https://developers.google.com/machine-learning/crash-course/multi-class-neural-networks/softmax?hl=pt-br. (Consultado em 09/11/2023).

[82] "sklearn.metrics.log_loss — scikit-learn 1.3.0 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.metrics.log_loss.html. (Consultado em 09/11/2023).

[83] "sklearn.model_selection.train_test_split — scikit-learn 1.3.0 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html. (Consultado em 09/11/2023).

[84] M. B. Kursa and W. R. Rudnicki, "Feature selection with the boruta package," *Journal of Statistical Software*, vol. 36, no. 11, p. 1–13, 2010.

[85] "Github - scikit-learn-contrib/boruta_py: Python implementations of the boruta all-relevant feature selection method.." https://github.com/scikit-learn-contrib/boruta_py. (Consultado em 09/11/2023).

[86] "sklearn.linear_model.lasso — scikit-learn 1.3.0 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html. (Consultado em 09/11/2023).

[87] "sklearn.ensemble.extratreesclassifier — scikit-learn 1.3.0 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html. (Consultado em 09/11/2023).

[88] "sklearn.feature_selection.selectkbest — scikit-learn 1.3.0 documentation." https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectKBest.html. (Consultado em 09/11/2023).

[89] "skopt.bayessearchcv — scikit-optimize 0.8.1 documentation." https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html. (Consultado em 09/12/2023).

[90] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: an overview," 2020.

[91] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, "Multiclass skin cancer classification using efficientnets – a first step towards preventing skin cancer," *Neuroscience Informatics*, vol. 2, no. 4, p. 100034, 2022.

[92] W. E. Marcílio and D. M. Eler, "From explanations to feature selection: assessing shap values as feature selection mechanism," in *2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 340–347, Ieee, 2020.

[93] B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable artificial intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model," *Complexity*, vol. 2021, pp. 1–11, 2021.

[94] R. Gentleman and V. J. Carey, "Unsupervised machine learning," in *Bioconductor case studies*, pp. 137–157, Springer, 2008.

[95] "O que é o power bi? definição e funcionalidades — microsoft power bi." https://powerbi.microsoft.com/pt-pt/what-is-power-bi/. (Consultado em 12/26/2022).

[96] J. Richardson, R. Sallam, K. Schlegel, A. Kronz, and J. Sun, "Magic quadrant for analytics and business intelligence platforms," *Gartner ID G00386610*, 2020.

[97] "Fastapi." https://fastapi.tiangolo.com/. (Consultado em 09/13/2023).

[98] E. Cristóvão, "Jornal de angola - notícias - número de utilizadores da internet cresce no país." https://www.jornaldeangola.ao/ao/noticias/numero-de-utilizadores-da-internet-cresce-no-pais/, 2 2023. (Accessed on 09/13/2023).

[99] "Apenas 6% dos luandenses utilizam internet banking." https://www.expansao.co.ao/expansao-mercados/interior/apenas-6-dos-luandenses-utilizam-internet-banking-96415.html. (Accessed on 09/13/2023).

Apêndice A

Anexos

A.1 Dicionário de dados anexado

account.csv

Account

COLUMN NAME	TYPE	DESCRIPTION
# account_id	integer	identification of the account
# district_id	integer	location of the branch
T frequency	string	frequency of issuance of statements: "POPLATEK MESICNE" stands for monthly issuance "POPLATEK TYDNE" stands for weekly issuance "POPLATEK PO OBRATU" stands for issuance after transaction
# date	integer	date of creating of the account: in the form YYMMDD

card.csv

Credit Card

COLUMN NAME	TYPE	DESCRIPTION
# card_id	integer	record identifier
# disp_id	integer	disposition to an account
T type	string	type of card: possible values are "junior", "classic", "gold"
T issued	string	issue date: in the form YYMMDD

client.csv

Client

COLUMN NAME	TYPE	DESCRIPTION
# client_id	integer	client identifier
# birth_number	integer	birthday and sex: the number is in the form YYMMDD for men, the number is in the form YYMM+50DD for women, where YYMMDD is the date of birth
# district_id	integer	address of the client

disp.csv

Disposition

COLUMN NAME	TYPE	DESCRIPTION
# disp_id	integer	record identifier
# client_id	integer	identification of a client
# account_id	integer	identification of an account
T type	string	type of disposition (owner/user): only owner can issue permanent orders and ask for a loan

district.csv

Demographic

COLUMN NAME	TYPE	DESCRIPTION
# a1	integer	district code
T a2	string	district name
T a3	string	region
# a4	integer	no. of inhabitants
# a5	integer	no. of municipalities with inhabitants < 499
# a6	integer	no. of municipalities with inhabitants 500-1999
# a7	integer	no. of municipalities with inhabitants 2000-9999
# a8	integer	no. of municipalities with inhabitants >10000
# a9	integer	no. of cities
# a10	decimal	ratio of urban inhabitants
# a11	integer	average salary
# a12	decimal	unemployment rate '95
# a13	decimal	unemployment rate '96
# a14	integer	no. of entrepreneurs per 1000 inhabitants
# a15	integer	no. of committed crimes '95
# a16	integer	no. of committed crimes '96

loan.csv

Loan

COLUMN NAME	TYPE	DESCRIPTION
# loan_id	integer	record identifier
# account_id	integer	identification of the account
# date	integer	date when the loan was granted: in the form YYMMDD
# amount	integer	amount of money
# duration	integer	duration of the loan
# payments	decimal	monthly payments
T status	string	status of paying off the loan: 'A' stands for contract finished, no problems, 'B' stands for contract finished, loan not payed, 'C' stands for running contract, OK so far, 'D' stands for running contract, client in debt

order.csv

Permanent Order (Debits Only)

COLUMN NAME	TYPE	DESCRIPTION
# order_id	integer	record identifier
# account_id	integer	account, the order is issued for
T bank_to	string	bank of the recipient: each bank has unique two-letter code
# account_to	integer	account of the recipient
# amount	decimal	debited amount
T k_symbol	string	characterization of the payment: "POJISTNE" stands for insurrance payment "SIPO" stands for household payment "LEASING" stands for leasing "UVER" stands for loan payment

trans.csv

Transaction

COLUMN NAME	TYPE	DESCRIPTION
# trans_id	integer	record identifier
# account_id	integer	account the transaction deals with
# date	integer	date of transaction: in the form YYMMDD
T type	string	+/- transaction: "PRIJEM" stands for credit "VYDAJ" stands for withdrawal
T operation	string	mode of transaction: "VYBER KARTOU" credit card withdrawal "VKLAD" credit in cash "PREVOD Z UCTU" collection from another bank "VYBER" withdrawal in cash "PREVOD NA UCET" remittance to another bank
# amount	decimal	amount of money
# balance	decimal	balance after transaction
T k_symbol	string	characterization of the transaction: "POJISTNE" insurance payment "SLUZBY" payment for statement "UROK" interest credited "SANKC. UROK" sanction interest if negative balance "SIPO" household "DUCHOD" old-age pension "UVER" loan payment
T bank	string	bank of the partner: each bank has unique two-letter code
# account	integer	account of the partner

A.2 Mapas da República Checa

Figura A.1: Mapa com as cidades da República Checa (4).

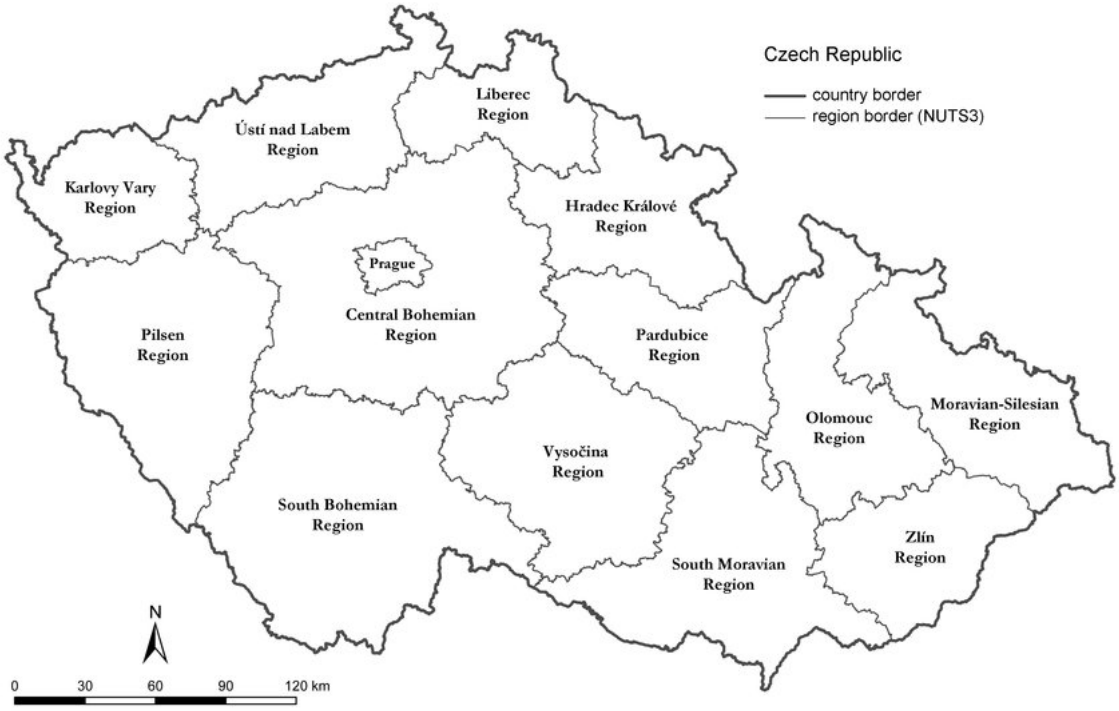

Figura A.2: Mapa com as regiões da República Checa (5).

A.3 Distribuição das variáveis por grupo

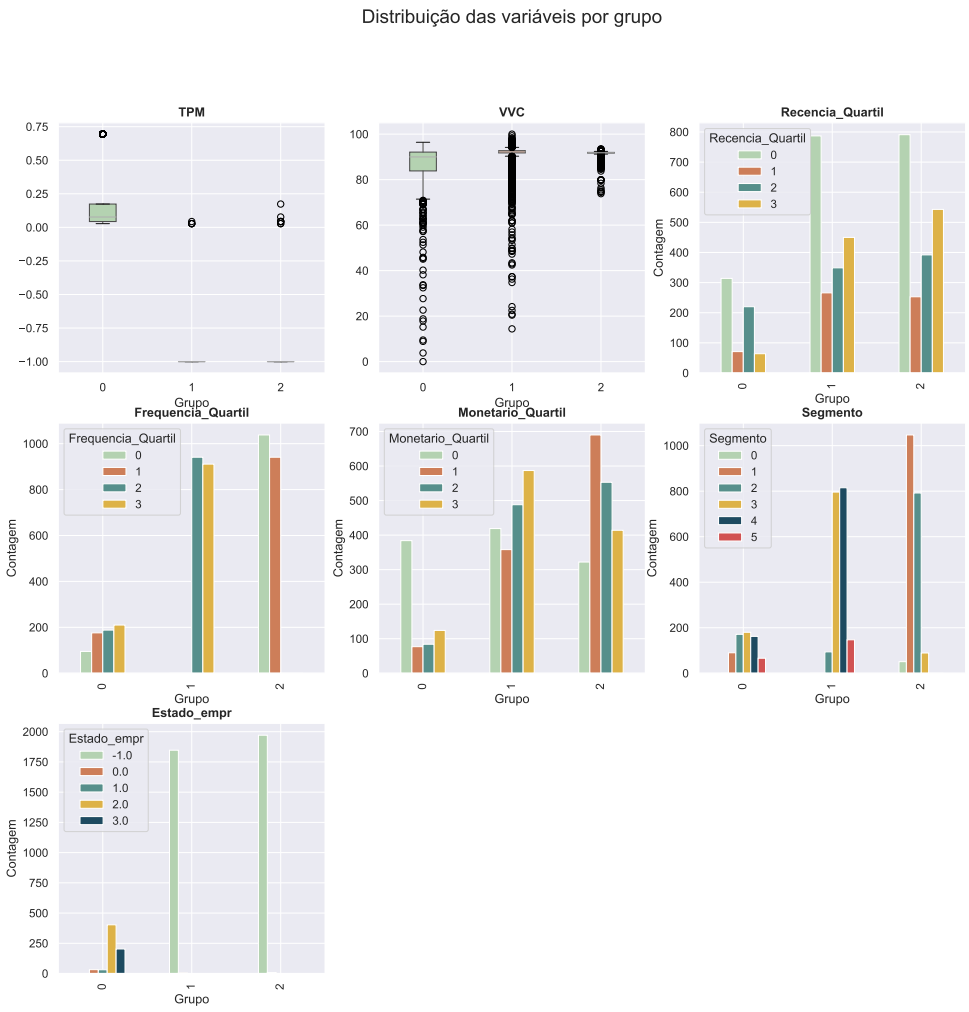

Figura A.3: Distribuição das variáveis utilizadas no algoritmo de segmentação, por grupo.

A.4 Questionário adoção do *internet banking*

Adoção do serviço de Internet Banking

O presente questionário é realizado no âmbito de uma dissertação do Mestrado em Ciência de Dados, na Universidade de Lisboa, em Portugal.

O objetivo deste estudo é perceber os fatores que influenciam a adoção do serviço de Internet Banking e utilização por parte dos clientes bancários luandeses. Além disso, pretende-se perceber para os clientes que não usam Internet Banking, quais as barreiras e desafios existentes. Por fim, pretende-se estudar qual a recetividade em relação ao serviço de um Gestor de Finanças Pessoais digital.

As respostas recolhidas serão totalmente anónimas e confidenciais, sendo usadas exclusivamente para fins analíticos. Por favor, responda a todas as perguntas para validar a sua participação.

Obrigado pela disponibilidade e colaboração.

* Indica uma pergunta obrigatória

1. Termos e Condições *

Marcar apenas uma oval.

- ◯ Concordo que qualquer informação fornecida neste inquérito pode ser utilizada para os fins acima mencionados.

Internet Banking

Internet Banking refere-se à plataforma disponibilizada no website do banco que permite um cliente efetuar um conjunto de operações bancárias à distância sem necessidade de deslocação física ao ATM ou agência bancária.

2. Atualmente efetua operações através do serviço de Internet Banking? *

Marcar apenas uma oval.

- ◯ Não. *Avançar para a pergunta 43*
- ◯ Sim. *Avançar para a pergunta 4*

3. Se respondeu que não, indique qual a principal razão...

Marcar apenas uma oval.

- ◯ Não tenho acesso à internet.
- ◯ Não confio no serviço de Internet Banking.
- ◯ Prefiro deslocar-me à agência bancária/ATM.
- ◯ Já usei mas o serviço Internet Banking não satisfez as minhas necessidades.
- ◯ Já usei mas o serviço de Internet Banking foi difícil de utilizar.
- ◯ É mais conveniente para mim recorrer a outras ** alternativas (por exemplo, a aplicação do telemóvel). **
- ◯ Outra: _____

Perfil de Utilizador do Internet Banking

4. Com que frequência utiliza o serviço de Internet Banking? *

Marcar apenas uma oval.

- ◯ Todos os dias.
- ◯ Duas ou três vezes por semana.
- ◯ Duas ou três vezes por mês.
- ◯ Uma vez por mês ou menos.

5. Nos últimos 12 meses, com que frequência utilizou os serviços de Internet Banking? *

Marcar apenas uma oval por linha.

	Nunca.	Uma vez por mês ou menos.	Duas ou três vezes por mês.	Duas ou três vezes por semana.	Todos os dias.
Consulta de saldo e movimentos.	◯	◯	◯	◯	◯
Carregamento e recarga da conta telefónica.	◯	◯	◯	◯	◯
Pagamentos de serviços.	◯	◯	◯	◯	◯
Transferências nacionais ou internacionais.	◯	◯	◯	◯	◯
Pedido de levantamento sem cartão.	◯	◯	◯	◯	◯

6. Indique o grau de satisfação geral para com os serviços de Internet Banking fornecidos pelo seu banco. **(1 = Pouco satisfeito, 5 = Muito sastifeito)** *

Marcar apenas uma oval.

	Pouco Satisfeito					Muito Satisfeito
	1	2	3	4	5	

7. Costuma deslocar-se ao balcão da sua agência bancária? *

Marcar apenas uma oval.

- Sim.
- Não.

Fatores determinantes da adoção do Internet Banking.

Indique quanto concorda com as seguintes afirmações, de uma escala de 1 a 5, onde **1 = Discordo Totalmente** e **5 = Concordo Totalmente**.

8. Uso de forma hábil o serviço de Internet Banking. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯
2 ◯
3 ◯
4 ◯
5 ◯

Concordo Totalmente

9. Acho que usar o serviço de Internet Banking é fácil. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯
2 ◯
3 ◯
4 ◯
5 ◯

Concordo Totalmente

10. Para mim, é fácil aprender a usar o serviço de Internet Banking. *

Marcar apenas uma oval.

Discordo Totalmente

1
2
3
4
5

Concordo Totalmente

11. A minha interação com o serviço de Internet Banking é clara e compreensível. *

Marcar apenas uma oval.

Discordo Totalmente

1
2
3
4
5

Concordo Totalmente

12. Sinto que o serviço de Internet Banking é útil. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Concordo Totalmente

13. O serviço de Internet Banking permite-me fazer pagamentos de forma mais eficiente. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Concordo Totalmente

14. O serviço de Internet Banking permite-me fazer pagamentos de forma mais conveniente. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯
2 ◯
3 ◯
4 ◯
5 ◯

Concordo Totalmente

15. O serviço de Internet Banking permite-me fazer pagamentos de forma mais rápida. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯
2 ◯
3 ◯
4 ◯
5 ◯

Concordo Totalmente

16. Considero que o serviço de Internet Banking é uma boa ideia. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯
2 ◯
3 ◯
4 ◯
5 ◯

Concordo Totalmente

17. Recomendaria o uso do serviço de Internet Banking a outras pessoas. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯
2 ◯
3 ◯
4 ◯
5 ◯

Concordo Totalmente

18. A minha atitude em relação ao serviço de Internet Banking é positiva. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯
2 ◯
3 ◯
4 ◯
5 ◯

Concordo Totalmente

19. O serviço de Internet Banking encaixa no meu estilo de vida. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯
2 ◯
3 ◯
4 ◯
5 ◯

Concordo Totalmente

20. O serviço de Internet Banking tem preços razoáveis. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Concordo Totalmente

21. O preço do serviço de Internet Banking é aceitável face às funcionalidades deste serviço. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Concordo Totalmente

22. Ao preço atual, o serviço de Internet Banking proporciona um bom valor ao utilizador. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Concordo Totalmente

23. Ao preço atual, o serviço de Internet Banking proporciona um bom valor ao utilizador. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Concordo Totalmente

24. O uso do serviço de Internet Banking tornou-se um hábito para mim. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Concordo Totalmente

25. Utilizo com muita frequência o serviço de Internet Banking. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Concordo Totalmente

26. Preciso de usar, muitas vezes, o serviço de Internet Banking. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Concordo Totalmente

27. Usar o serviço de Internet Banking tornou-se natural para mim. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Concordo Totalmente

28. Tenho os recursos necessários para usar o serviço de Internet Banking. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯
2 ◯
3 ◯
4 ◯
5 ◯

Concordo Totalmente

29. Tenho o conhecimento necessário para usar o serviço de Internet Banking *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯
2 ◯
3 ◯
4 ◯
5 ◯

Concordo Totalmente

30. Se tiver dificuldade a usar o serviço de Internet Banking, existem profissionais que me podem ajudar. *

Marcar apenas uma oval.

	Discordo Totalmente
1	◯
2	◯
3	◯
4	◯
5	◯
	Concordo Totalmente

31. O serviço de Internet Banking é bem administrado e atual. *

Marcar apenas uma oval.

	Discordo Totalmente
1	
2	
3	
4	
5	
	Concordo Totalmente

32. Todas as operações bancárias que necessito estão disponibilizadas no serviço de Internet Banking. *

Marcar apenas uma oval.

	Discordo Totalmente
1	◯
2	◯
3	◯
4	◯
5	◯
	Concordo Totalmente

33. Na generalidade, o serviço de Internet Banking satisfaz as minhas expectativas. *

Marcar apenas uma oval.

Discordo Totalmente

1
2
3
4
5

Concordo Totalmente

34. Eu acredito que o serviço de Internet Banking é confiável. *

Marcar apenas uma oval.

Discordo Totalmente

1
2
3
4
5

Concordo Totalmente

35. Acredito que o serviço de Internet Banking considera a relação com os seus clientes como prioridade máxima. *

Marcar apenas uma oval.

	Discordo Totalmente					Concordo Totalmente
	1	2	3	4	5	

36. O serviço de Internet Banking é previsível. *

Marcar apenas uma oval.

	Discordo Totalmente
1	
2	
3	
4	
5	
	Concordo Totalmente

37. Acho que serviço de Internet Banking tem mecanismos para garantir a transmissão segura de informações pessoais dos seus utilizadores. *

Marcar apenas uma oval.

	Discordo Totalmente
1	
2	
3	
4	
5	
	Concordo Totalmente

38. Sinto-me seguro para realizar transações quando uso o serviço de Internet Banking. *

Marcar apenas uma oval.

	Discordo Totalmente
1	
2	
3	
4	
5	
	Concordo Totalmente

39. O serviço de Internet Banking é um serviço seguro, através do qual posso enviar informações confidenciais. *

Marcar apenas uma oval.

	Discordo Totalmente
1	◯
2	◯
3	◯
4	◯
5	◯
	Concordo Totalmente

40. Pretendo usar o Internet Banking no futuro. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Concordo Totalmente

41. Pretendo usar continuamente o serviço de Internet Banking. *

Marcar apenas uma oval.

Discordo Totalmente

1 ◯

2 ◯

3 ◯

4 ◯

5 ◯

Concordo Totalmente

42. Recomendo o Internet Banking aos meus amigos e familiares. *

Marcar apenas uma oval.

	Discordo Totalmente
1	
2	
3	
4	
5	
	Concordo Totalmente

Avançar para a pergunta 47

Não utilizador do serviço de Internet Banking

43. Com que frequência se desloca ao balcão do seu banco?

Marcar apenas uma oval.

- Todos os dias.
- Duas a três vezes por semana.
- Duas a três vezes por mês.
- Uma vez por mês ou menos.
- Nunca.

44. Quais as principais razões que o/a levam a deslocar-se ao balcão? *

Marcar tudo o que for aplicável.

- [ ] Não ter acesso à internet.
- [ ] Não conhecer meios alternativos.
- [ ] Não saber usar os meios alternativos.
- [ ] Tenho mais confiança na interação com a pessoa ao balcão.
- [ ] As alternativas não satisfazem o que preciso.
- [ ] Porque as filas para os ATMs são grandes.
- [ ] Outra: _____

45. Quais as principais ações que efetua quando vai ao balcão? *

Marcar tudo o que for aplicável.

- [ ] Consulta de saldo e movimentos.
- [ ] Consulta de estado de poupanças, investimentos e seguros.
- [ ] Pedido de créditos e seguros.
- [ ] Consulta de créditos e seguros.
- [ ] Pagamento de serviços.
- [ ] Transferências nacionais ou internacionais.
- [ ] Solicitação de cartões de crédito ou débito.
- [ ] Administração e gestão da conta.
- [ ] Pedido de cheques e cadernetas.
- [ ] Outra: _____

46. Quais os principais motivos que o levariam a adotar o serviço de Internet Banking? *

Marcar tudo o que for aplicável.

- [ ] Acesso à Internet.
- [ ] Campanhas realizadas pelo banco sobre a segurança dos serviços alternativos.
- [ ] Mais informações sobre os possíveis meios alternativos.
- [ ] Melhorar os meios alternativos já existentes (como a usabilidade e rapidez).
- [ ] Outra: _____

Gestor de Finanças Pessoais

O **Gestor de Finanças Pessoais** é uma aplicação que permite ao utilizador gerir as suas finanças pessoais numa única plataforma digital, integrada no serviço de Internet Banking.

47. Classifique as seguintes funcionalidades do Gestor de Finanças Pessoais de acordo com as suas necessidades e preferências. *

Marcar apenas uma oval por linha.

	Muito Inútil.	Inútil.	Indiferente.	Útil.	Muito Útil.
Visualizar através de gráficos as minhas despesas e rendimentos, por categoria.	◯	◯	◯	◯	◯
Gerir várias contas bancárias numa só aplicação.	◯	◯	◯	◯	◯
Definir os meus objetivos de poupança/orçamento.	◯	◯	◯	◯	◯
Receber conselhos financeiros personalizados com base nos meus gastos e hábitos.	◯	◯	◯	◯	◯
Receber previsões de despesas do futuro.	◯	◯	◯	◯	◯
Ter acesso a um calendário financeiro que me permite visualizar os meus movimentos bancários ao longo do mês.	◯	◯	◯	◯	◯

48. De acordo com as funcionalidades acima descritas, com que frequência usaria o Gestor de Finanças Pessoais? **(1 = Nunca, 5 = Todos os dias)** *

Marcar apenas uma oval.

	Nunca					Todos os dias
	1	2	3	4	5	

49. Se não utilizasse frequentemente o Gestor de Finanças Pessoais, quais seriam as principais razões?

Marcar tudo o que for aplicável.

- [ ] A aplicação não me daria valor.
- [ ] Não confio na aplicação.
- [ ] Não satisfaz as minhas necessidades.
- [ ] Não tenho acesso fácil à Internet.
- [ ] Não compreendo o que é um Gestor de Finanças Pessoais.
- [ ] Outra: _____

Características demográficas.

50. Género. *

Marcar apenas uma oval.

- ◯ Feminino.
- ◯ Masculino.
- ◯ Prefiro não dizer.

51. Idade. *

Marcar apenas uma oval.

- ◯ 18-25 anos.
- ◯ 26-35 anos.
- ◯ 36-50 anos.
- ◯ Mais de 50 anos.

52. Habilitações literárias. *

Marcar apenas uma oval.

- ◯ Ensino primário.
- ◯ Ensino secundário.
- ◯ Bacherel/Licenciado.
- ◯ Mestrado/Doutoramento.

53. Dados profissionais. *

Marcar apenas uma oval.

- ◯ Estudante.
- ◯ Trabalhador estudante.
- ◯ Desempregado(a).
- ◯ Trabalhador por conta de outrém.
- ◯ Trabalhador por conta própria.
- ◯ Reformado(a).

54. Qual o seu rendimento líquido mensal em Kz? *

Marcar apenas uma oval.

- ◯ Até 100.000,00.
- ◯ De 100.001,00 até 300.000,00.
- ◯ De 300.001,00 até 500.000,00.
- ◯ De 500.001,00 até 1.000.000,00.
- ◯ Acima de 1.000.000,00.
- ◯ Prefiro não dizer.

55. Com que frequência utiliza os seguintes dispositivos tecnológicos. *

Marcar apenas uma oval por linha.

	Nunca.	Uma vez por mês ou menos.	Duas a três vezes por mês.	Duas a três vezes por semana.	Todos os dias.
Telemóvel.	◯	◯	◯	◯	◯
Computador.	◯	◯	◯	◯	◯
Tablet.	◯	◯	◯	◯	◯