|NST|TUTO
UN|VERS|TÁR|O
DE L|SBOA

# Modelo para retenção dos clientes e análise de afinidade no Ramo Automóvel – Aplicação no Mercado Segurador Português

João Manuel Matos Dinis

Mestrado em Sistemas Integrados de Apoio à Decisão

Orientador:
Doutor João Carlos Amaro Ferreira, Professor Auxiliar com Agregação
ISCTE – Instituto Universitário de Lisboa (ISCTE-IUL)

Coorientador:
Doutor Luís Carlos Barruncho dos Santos Gonçalves, Professor Auxiliar Convidado
ISCTE – Instituto Universitário de Lisboa (ISCTE-IUL)

outubro, 2021

TECNOLOGIAS
E ARQUITETURA

Departamento de Ciências e Tecnologias da Informação

# Modelo para retenção dos clientes e análise de afinidade no Ramo Automóvel – Aplicação no Mercado Segurador Português

João Manuel Matos Dinis

Mestrado em Sistemas Integrados de Apoio à Decisão

Orientador:
Doutor João Carlos Amaro Ferreira, Professor Auxiliar com Agregação
ISCTE – Instituto Universitário de Lisboa (ISCTE-IUL)

Coorientador:
Doutor Luís Carlos Barruncho dos Santos Gonçalves, Professor Auxiliar Convidado
ISCTE – Instituto Universitário de Lisboa (ISCTE-IUL)

outubro, 2021

# Agradecimentos

Gostaria de agradecer aos orientadores João Carlos Ferreira e Luís Carlos Gonçalves, por todos os comentários, por todo o apoio e dedicação que me foram transmitindo ao longo desta dissertação assim como todo o tempo dispensado ao longo deste ano.

Um agradecimento especial para os avaliadores Soraia Bornett e Luís Silva pela disponibilidade demonstrada e que contribuíram com os seus comentários e avaliações de forma a obter os melhores resultados possíveis, assim como à companhia seguradora que permitiu efetuar esta dissertação.

Por fim, gostaria de agradecer a todos os colegas, amigos e familiares por todo o apoio demonstrado ao longo deste último ano.

# Resumo

Desde a idade média que o ser humano procura garantir uma segurança que o cubra a si mesmo e aos seus bens materiais. Para esse efeito surgiu o mercado segurador, onde o risco passa a ser assegurado por uma companhia seguradora, tranquilizando o dono do bem material, deixando o mesmo focar-se noutros temas.

O mercado segurador português tem sofrido um crescimento constante ao longo dos anos, tornando-se cada vez mais num mercado competitivo entre companhias seguradoras.

Uma vez que este mercado se torna bastante competitivo, as companhias tentam manter ou aumentar a sua quota parte do mercado. As companhias seguradoras possuem cada vez mais interesse em possuir bons modelos de retenção de clientes.

Para tal, nesta dissertação foram desenvolvidos dois artefactos, sendo um artefacto orientado a uma retenção de clientes através de uma abordagem preventiva; e um segundo artefacto orientado a uma retenção de clientes através de uma abordagem mais reativa.

Na abordagem preventiva efetuou-se uma análise de afinidade entre produtos, permitindo assim detetar padrões de contratações entre os clientes.

Na abordagem reativa, foi realizado um modelo de *customer churn*, onde se efetuou uma análise ao ramo automóvel, com o objetivo de conseguir identificar os principais motivos que levam os clientes a abandonar a companhia.

Como resultados da abordagem preventiva, obteve-se um quadro final com as regras de afinidade entre produtos para os diversos conjuntos de clientes detetados. Na abordagem reativa, foi obtido uma análise das variáveis mais significativas em relação ao abandono dos clientes do ramo automóvel.



**Palavras Chave:** Mercado Segurador, Risco, *Big Data*, Ciência de Dados, CRISP-DM, Algoritmos, Perfil de Clientes, Retenção de clientes.

# Abstract

Ever since the Middle Ages, the human being has felt the need to be secure, along with his material goods. The insurance market came to support that need, where the risk for his possessions is now insured by an insurance company, providing comfort to the personal owner and letting him focus on other issues.

The Portuguese insurance market has been growing consistently over the years, ending up as a highly competitive market between insurance companies.

Given how competitive the situation has become, each insurance company tries to retain or increase their own share of the market. Insurance companies are growing interest in having good business models for customer retention.

In the interest of that, this dissertation contains two artifacts, first one being more focused on retaining customers through a preventive approach, whereas the second one is more oriented towards a more reactive approach.

In the preventive approach, an analysis was performed on the affinity between products, allowing the detection of contractual patterns with the customers.

In the reactive approach, it was created a customer churn model where an analysis was performed on the motor branch, with the goal of identifying the main reasons for customers abandoning the company.

In the results of the preventive approach, a final board was obtained with the rules for the affinity between products regarding the multiple groups of customers. In the reactive approach, an analysis was obtained with the most significant variables related to the abandonment of the customers in the motor branch.



**Key words**: Insurance Market, Risk, Big Data, Data Science, CRISP-DM, Algorithms, Customer Profile, Customer Churn

# Índice

# Índice de Figuras

# Índice de Tabelas

# Glossário

ASF - Autoridade de Supervisão de Seguros e Fundos de Pensões

CRISP-DM - Cross Industry Standard Process for Data Mining

CRM – Customer Relationship Management

MBA – Market Basket Analysis

NIF – Número de Identificação Fiscal

NPC – Número Pessoa Coletiva

PRISMA - Preferred Reporting Items for Systematic Reviews and Meta-Analyses

SVM – Support Vector Machine

# Capítulo 1 – Introdução

Neste capítulo irá ser dado um breve enquadramento ao tema e aos objetivos desta dissertação. De seguida, irá ser detalhado a estrutura e organização da dissertação. Por fim, é referido a contribuição que se pretende fornecer à companhia seguradora.

## *1.1 – Motivação e Enquadramento do Tema*

A importância dada a modelos de retenção de clientes tem subido ao longo dos últimos anos, uma vez que o mercado se torna mais competitivo entre si e os clientes tornam-se mais suscetíveis a trocar de companhia, como por exemplo, com um prémio inferior ou com melhores benefícios contratuais.

A função destes modelos de retenção prende-se com o facto de conseguir prever com antecedência, quais são os clientes que irão deixar a companhia. Esta informação acaba por ser bastante valiosa para a companhia uma vez que conseguirão abordar o cliente e oferecer alguns benefícios de modo que o cliente permaneça na companhia.

Para tal, irá também ser desenvolvido uma segmentação de clientes de modo a conseguirmos entender padrões de produtos que clientes semelhantes tenham, permitindo conseguir oferecer benefícios personalizados aos clientes que estejam em vias de abandonar a companhia.

## *1.2 – Objetivos*

Como foi descrito no ponto acima, o principal objetivo desta dissertação está na retenção de clientes que detenham atualmente apólices na companhia. Com este objetivo definido foi identificado duas abordagens distintas a serem criadas:

1 – Uma abordagem preventiva, que irá permitir cross-selling de produtos, reforçando a ligação do cliente à companhia;

2 – Uma abordagem reativa, que irá permitir identificar os clientes que estejam em vias de anular as suas apólices, e por sua vez de abandonar a companhia.

Com estas abordagens, surge assim as *main research questions*:

1- É possível detetar padrões de contratação de apólices, com base na caminhada do cliente na companhia?
2- É possível detetar clientes que estejam em vias de abandonar a companhia?

As duas abordagens referidas são independentes entre si, mas com o mesmo objetivo: reter os clientes, sendo que através de uma abordagem preventiva é pretendido ir fidelizando o cliente, oferecendo produtos personalizados com base no seu histórico de contratações, evitando que o mesmo pretenda vir a abandonar a companhia; e uma abordagem mais reativa que pretende evitar perder os clientes que já se encontrem a considerar abandonar a companhia.

## *1.3 – Estrutura e Organização da Dissertação*

Esta dissertação encontra-se organizada através de seis capítulos, sendo eles os seguintes:

- **Capítulo 2 (Estado de Arte)**, contêm uma análise da literatura sobre os tópicos relacionados com os temas desta dissertação, nomeadamente, mercado segurador, risco, *Market Basket Analysis* e *Customer Churn*;
- **Capítulo 3 (Análise de afinidade de produtos)**, é descrito todos os passos efetuados relacionados com o modelo de afinidade de produtos, incluindo uma visualização de dados e uma modelação de dados de modo a detetar regras de afinidade de produtos;
- **Capítulo 4 (*Customer Churn*)**, é descrito o modelo de *Customer Churn* criado, descrevendo as diversas etapas da metodologia usada (CRISP-DM) e os resultados obtidos;
- **Capítulo 5 (Demonstração e Avaliação)** reflete as demonstrações efetuadas junto dos avaliadores e são apresentadas as avaliações atribuídas pelos mesmos a cada uma das iterações efetuadas;
- **Capítulo 6 (Conclusões e Trabalho Futuro)**, engloba uma análise comparativa entre os objetivos definidos no primeiro capítulo com os resultados obtidos nos capítulos anteriores e enumera-se futuros trabalhos possíveis de se realizar de modo a enriquecer o trabalho já efetuado e descritos nesta dissertação.

## *1.4 – Contribuição para a companhia seguradora*

Com a elaboração destes modelos pretende-se fornecer à companhia duas novas ferramentas que ajudem a detetar previamente clientes que estejam em vias de anular as suas apólices, e que auxilie a promover ações orientadas de modo a evitar a perda dos mesmos.

Com estas ferramentas será possível obter um novo conjunto de regras, permitindo dar origem a ações com maior nível de detalhe. Estas ações podem incluir ofertas personalizadas, uma vez que será possível assinalar produtos que sejam atraentes a cada cliente, de modo que a companhia demonstre atenção especial com o cliente e que o mesmo continue afiliado à companhia.

## *1.5 – Metodologia de Avaliação*

De modo a podermos avaliar os modelos que foram sendo obtidos ao longo de todo o processo de análise e de desenvolvimento, foi adotada a metodologia *Design Science Research Methodology*, também designada como DSRM.

O principal objetivo desta metodologia é de gerar artefactos que consigam adicionar valor aos utilizadores finais [1].

Na Figura 1 é possível observar todas as etapas que engloba um projeto com esta metodologia. Dos quatro pontos de entrada descritos na Figura 1, esta dissertação tem como ponto de entrada o ponto de *Objetive-Centered Solution,* uma vez que o problema já estava diagnosticado antes do início desta dissertação. Tendo a dissertação arrancado com a definição de objetivos e com a escolha dos artefactos que mais acrescessem valor à companhia.

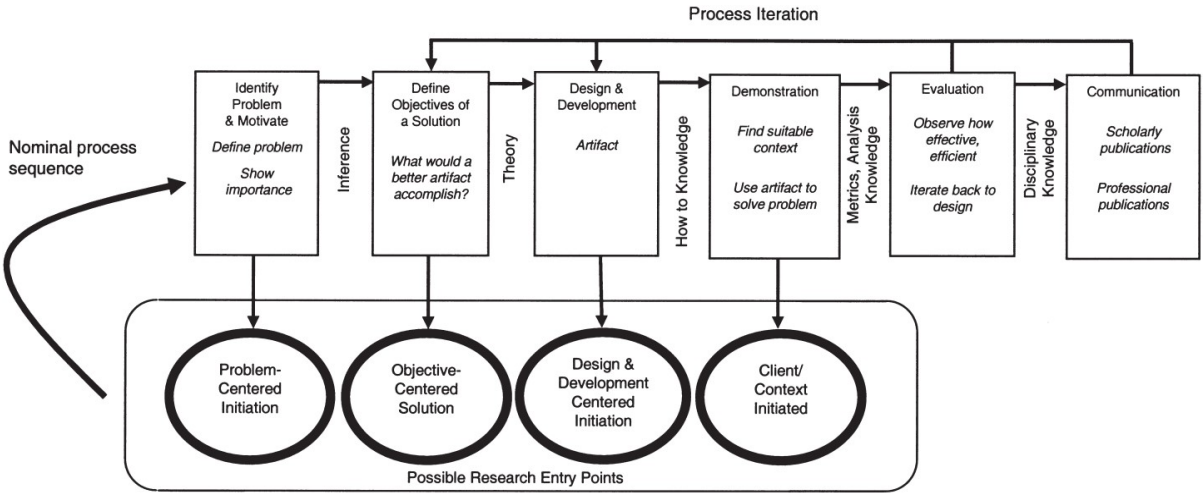



**Figura 1 – Etapas da metodologia DSRM (Fonte: Peffers, K et al [1])**

Esta dissertação está estruturada em quatro iterações de avaliação, junto de especialistas que irão analisar os artefactos obtidos, comentar possíveis melhoramentos assim como responder a um pequeno inquérito sobre o mesmo.

A primeira iteração de avaliação (maio de 2021) irá servir para avaliar o artefacto de abordagem preventiva, *Market Basket Analysis*, sendo posteriormente efetuada uma segunda iteração sobre a mesma abordagem (junho de 2021) tendo por base todos os melhoramentos indicados pelos avaliadores.

A terceira iteração (julho de 2021) irá incidir sobre o segundo artefacto que diz respeito à abordagem reativa, *Customer Churn*, sendo para tal demonstrado todas as análises e modelos obtidos com este modelo. A quarta iteração (agosto de 2021) irá ser usada para avaliar o estado final do artefacto desta abordagem reativa juntamente com todos os comentários obtidos na terceira iteração.

O inquérito fornecido aos avaliadores no final de cada iteração, será composto pelos fatores de avaliação que foram selecionados numa reunião inicial. Para a análise destes critérios foi tido em conta os critérios referidos por Prat et al [3].

Tendo em conta esses critérios de avaliação definiram-se os seguintes critérios de avaliação:

Ambos os artefactos terão como objetivo, a eficácia de modo a obtermos confiança nos resultados obtidos.

No que diz respeito às avaliações da abordagem preventiva, foram definidos os seguintes critérios: Performance, uma vez que pode ser necessário efetuar análises num curto espaço de tempo e Capacidade de aprendizagem, uma vez que o mercado está em movimento constante assim como os seus clientes, sendo necessário que o artefacto criado consiga aprender essas alterações.

Os critérios relativos à abordagem reativa, foram os seguintes: compreensibilidade de forma a providenciar informação coerente ao utilizador; fácil de usar; que providencie informação que permita à companhia reter clientes; e que contenha nível de detalhe de modo a oferecer à companhia regras detalhadas sobre os diferentes tipos de clientes que existem.

Estes critérios estão identificados na Figura 2, sendo que a verde estão destacados os critérios relativo a ambos os artefactos; a azul estão mencionados os critérios relativos à abordagem preventiva e a amarelo estão mencionados os artefactos relativos à abordagem reativa.

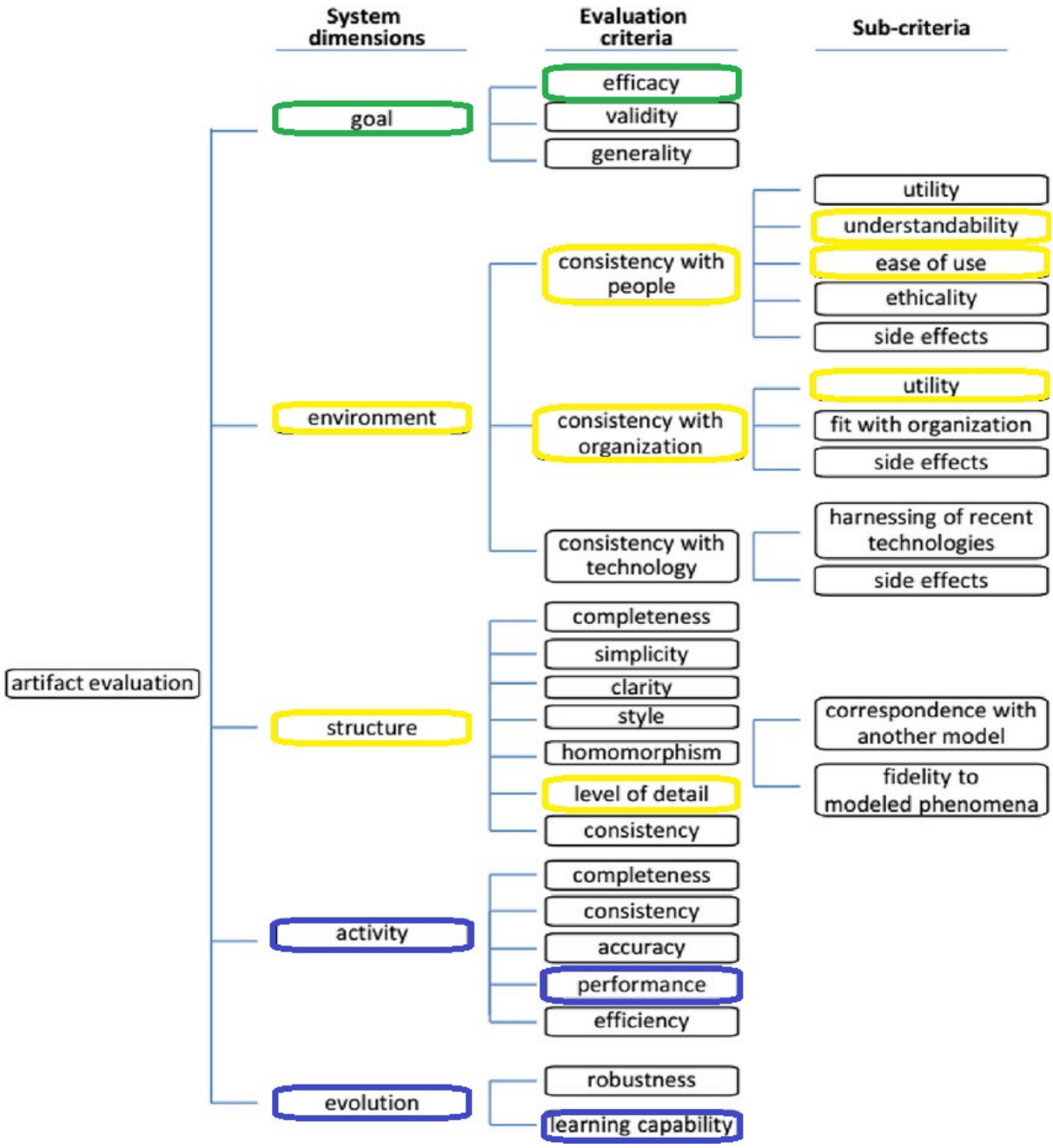



**Figura 2 – Critérios de avaliação definidos por Prat et al. (Fonte:Prat et al [3])**

No que se refere à avaliação de cada um destes tópicos acima mencionados, foi adotado o seguinte grau de avaliação:

**A** – Entre 100% e 75% de acordo com o objetivo esperado;

**B** – Entre 75% e 50% de acordo com o objetivo esperado;

**C** – Entre 50% e 25% de acordo com o objetivo esperado;

**D** – Entre 25% e 0% de acordo com o objetivo esperado.

# Capítulo 2 – Estado de arte

Este capítulo é descrito o estado de arte dos temas abordados nesta dissertação. Primeiro irá ser dado uma visão do mercado segurador e de seguida iremos abordar o conceito de risco. De seguida, irá ser descrito uma análise sobre os temas de *Market Basket Analysis* e de *Customer Churn*.

## *2.1 - Mercado Segurador*

### 2.1.1. – Definição

O mercado segurador funciona assente sobre apólices. Uma apólice, contém associado um ou mais riscos – que poderá ser um objeto ou uma pessoa segurada – e que caso algum incidente ocorra sobre o risco irá ser designado como um sinistro. Uma apólice tem como objetivo prevenir consequências negativas, caso se concretize um sinistro sobre o risco assegurado [4].
Consoante as características do risco, irá ser associado um prémio ao mesmo. Um prémio da apólice corresponde ao valor que o tomador – pessoa responsável pela apólice – irá ter que remunerar pela apólice. As características do risco poderão resultar em agravamentos (aumento de prémio) ou em bónus (diminuição de prémio).
No caso desta dissertação – focado no ramo Automóvel - exemplos desses agravamentos estão frequentemente relacionados com o historial do condutor (sinistralidade recente, anos com carta de condução, ...), zona em que conduz frequentemente, dados relacionados com o risco assegurado (potência do motor, peso do veículo, ...) [5].

### 2.1.2. – Origem

Como Larrarnendi refere [6], o ser humano sempre sentiu necessidade de segurança contra todos os perigos que o rodeiam ou que o poderão vir a rodear. Através desse receio acabou por surgir o primeiro conceito de seguro. Este conceito surgiu numa altura em que o ser humano transportava bens através das atividades comerciais que começavam a surgir, e de modo a que esses comerciantes começassem a sentir-se seguros desses transportes começaram a recorrer a "seguros" de que protegiam as suas cargas e que reembolsavam caso a mesma se perdesse ou danificasse durante a viagem.
Entre o século XVII e século XVIII, com o desenvolver dos estudos científicos, dos avanços matemáticos e do campo das estatísticas, foi possível criar grandes avanços nas funções de atuariado, o que permitiu ao mercado segurador desenvolver-se, tendo sido esta fase designada "the birth of modern insurance" [7].

### 2.1.3. – Relevo Nacional

A nível nacional, observando os resultados divulgados pela ASF, demonstrado na Figura 3, conseguimos visualizar que o valor dos prémios cobrados pelas companhias tem subido ao longo dos últimos anos.

Apesar de no ano de 2019 observarmos um decréscimo de 5,8% face a 2018, conseguimos verificar que entre 2016 e 2019 o mercado segurador em Portugal cresceu cerca de 12,2%.

Conseguimos ainda observar que os ramos mais significativos no mercado nacional, por ordem decrescente, são o ramo de Seguro de Vida, ramo de Acidentes e Doença e o ramo Automóvel.

**Produção* de seguro direto em Portugal por ramos - Mercado**

u.m: milhares de euros

| | 2016 | 2017 | 2018 | 2019** | D 19/18 |
|---|---|---|---|---|---|
| **Vida** | **6 677 410** | **7 088 577** | **8 122 717** | **6 991 849** | **-13,9%** |
| Seguro de Vida | 4 991 079 | 4 900 162 | 6 354 702 | 5 283 115 | -16,9% |
| Seguros Ligados a Fundos de Investimento | 1 686 331 | 2 187 101 | 1 767 235 | 1 704 184 | -3,6% |
| Operações de Capitalização | 0 | 1 313 | 780 | 4 550 | 483,3% |
| **Não Vida** | **4 194 198** | **4 493 706** | **4 825 262** | **5 209 209** | **8,0%** |
| Acidentes e Doença | 1 482 196 | 1 631 712 | 1 789 327 | 1 962 335 | 9,7% |
| Acidentes de trabalho | 623 952 | 705 189 | 800 638 | 895 066 | 11,8% |
| Doença | 693 770 | 751 466 | 807 135 | 877 385 | 8,7% |
| Acidentes (outros) | 164 474 | 175 058 | 181 554 | 189 885 | 4,6% |
| Incêndio e Outros Danos | 778 658 | 804 454 | 847 743 | 905 816 | 6,9% |
| Automóvel | 1 522 149 | 1 610 396 | 1 719 425 | 1 839 031 | 7,0% |
| Marítimo e Transportes | 24 633 | 25 826 | 25 343 | 26 523 | 4,7% |
| Aéreo | 6 215 | 7 045 | 7 195 | 8 784 | 22,1% |
| Mercadorias Transportadas | 21 558 | 21 296 | 20 807 | 20 576 | -1,1% |
| Responsabilidade Civil Geral | 116 284 | 127 071 | 131 289 | 143 023 | 8,9% |
| Diversos | 242 504 | 265 905 | 284 135 | 303 122 | 6,7% |
| **TOTAL** | **10 871 608** | **11 582 282** | **12 947 979** | **12 201 058** | **-5,8%** |

* Inclui prémios brutos emitidos de contratos de seguro e receita processada de contratos de investimento e de prestação de serviço

** Valores provisórios

**Figura 3 - Evolução de prémios de seguro direto em Portugal (Fonte: ASF [2])**

Segundo a ASF [2], no que diz respeito à distribuição dos prémios no mercado nacional conseguimos observar que as cinco principais companhias existentes são: *Fidelidade* com 24,96% de quota de mercado, *Ocidental Vida* com uma quota de 10,42%, *Seguradoras Unidas* com 7,15%, *BPI Vida e Pensões* com 6,81% e *Santander Totta Vida* com 5,51%.

### 2.1.4. – Risco

Como já foi mencionado no ponto anterior, existem diferentes tipos de riscos assegurados.

Estes riscos poderão ser sobre pessoas ou sobre bens imóveis. No caso de pessoas seguradas, poderá ser sobre doenças ou sobre casos de morte ou invalidez. No caso de bens, poderá ser sobre qualquer objeto.

A nível do ramo automóvel, o objeto seguro refere-se a veículos motorizados, que podem conter diversas finalidades (fins particulares, para transporte de passageiros, para colecionadores, entre outros).

## *2.3 - Market Basket Analysis*

### 2.3.1. – Análise

Para o artefacto de *Market Basket Analysis*, foi efetuado um levantamento de artigos revelantes com utilização do método PRISMA, conforme descrito por Mohan D. et al [8], que está descrito na Figura 4. Efetuou-se consultas nos repositórios *Scopus* e *IEEE Explore* que contivessem as palavras *Market Basket Analysis*, publicados entre 2016 e 2020, referentes às categorias *Computer Science*, *Engineering* ou *Business, Management and Accounting* tendo obtido 192 e 69 documentos respetivamente.

Dando origem à seguinte consulta:

TITLE-ABS-KEY ( market AND basket AND analysis ) AND ( LIMIT-TO ( PUBYEAR , 2020 ) OR LIMIT-TO ( PUBYEAR , 2019 ) OR LIMIT-TO ( PUBYEAR , 2018 ) OR LIMIT-TO ( PUBYEAR , 2017 ) OR LIMIT-TO ( PUBYEAR , 2016 ) ) AND ( LIMIT-TO ( SUBJAREA , "COMP" ) OR LIMIT-TO ( SUBJAREA , "ENGI" ) OR LIMIT-TO ( SUBJAREA , "BUSI" ) )

De seguida, removeram-se os documentos em duplicado, passando a contar com 230 documentos.

Dos 230 documentos foi efetuada uma exclusão de 154 documentos através da análise do título dos documentos e pela análise às palavras-chaves que não continham as palavras-chave selecionadas nesta dissertação.

Dos 76 documentos restantes, efetuou-se a leitura dos *abstracts*, tendo resultado numa exclusão de 56 artigos uma vez que não tinham a mesma finalidade desta dissertação.

Após leitura integral dos restantes 20 artigos, obtiveram-se 4 artigos relevantes para o tema em análise.

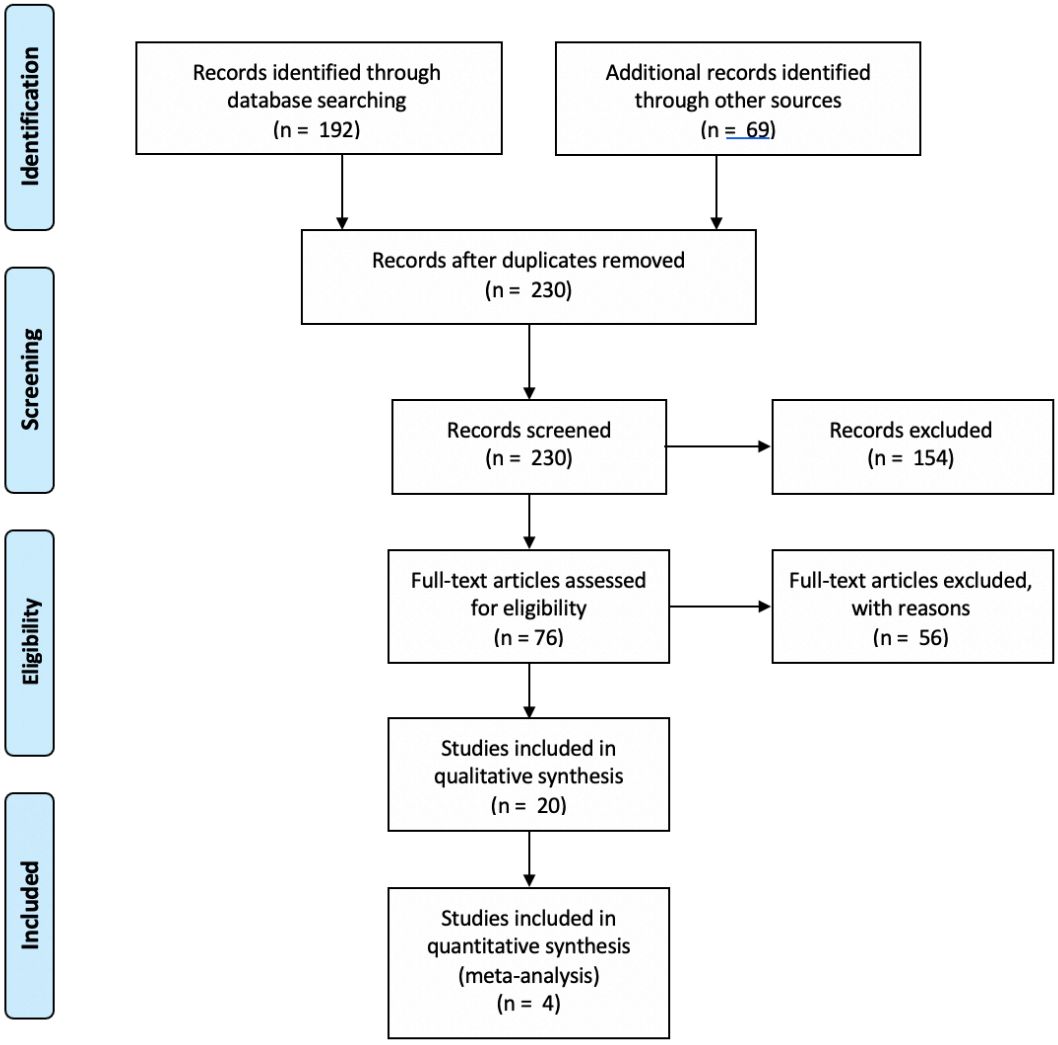



**Figura 4 – Esquema PRISMA efetuado para a pesquisa de documentos relacionados com MBA (adaptado de Mohan, D et al [8])**

Com recurso à ferramenta *VOSviewer* – ferramenta que deteta relações entre artigos – podemos visualizar na Figura 5 que este tema reúne três tópicos principais: tópicos de negócio (a verde), tópicos de informática (a vermelho) e tópicos referentes a clientes (a azul). Ao observar a Figura 5 visualizamos que existe uma grande relação entre algoritmos e regras de associação, isto demonstra que são conceitos frequentemente usados neste tipo de artigos, comprovando assim a grande interligação das diversas áreas fruto da grande investigação e crescente importância do tema de afinidade de produtos nos últimos anos.

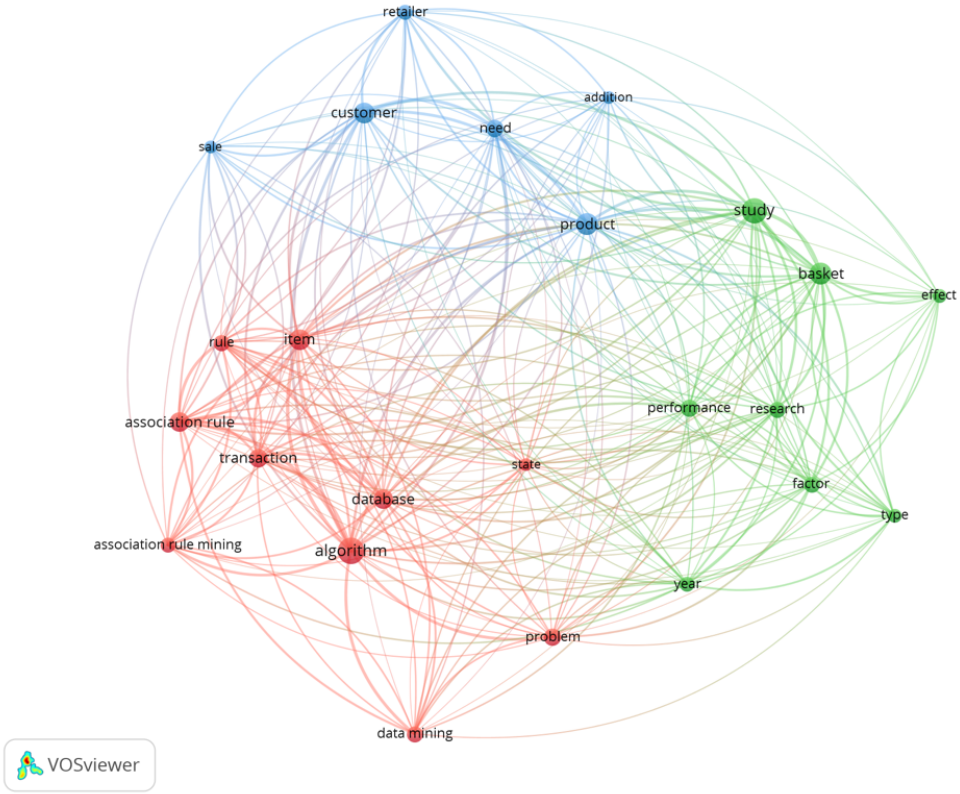



**Figura 5 – Gráfico de relação executado no VOSviewer[9] entre artigos referentes ao tema *Market Basket Analysis***

### 2.3.2. – Definição

*Market Basket Analysis* tem como objetivo agrupar diversos produtos que tendem a ser comprados em conjunto pelos clientes [10]. Para a análise e criação destes modelos é frequentemente usado uma metodologia CRISP-DM.

Estes modelos são frequentemente observáveis em hipermercados, uma vez que estes estabelecimentos tendem a colocar produtos que costumam ser comprados simultaneamente próximos entre si [11]. Tendo como objetivo de aumentarem as vendas de ambos os produtos. Outro exemplo onde é facilmente observável estes modelos são os sites de comércio *online*, no momento em que se adiciona um determinado produto ao carrinho de compras, e em que se apresenta um leque de produtos "Quem compra costuma comprar também...".

Estes modelos baseiam-se no algoritmo *Apriori*, este algoritmo recebe como input todas as compras efetuadas e como output, apresenta um conjunto de regras de associação com os seguintes campos: antecedentes, consequência, suporte e confiança [12].

Antecedentes correspondem aos produtos que o cliente já comprou, a consequência corresponde ao produto que o cliente poderá comprar em conjunto. Para cada regra está

associado um determinado grau de suporte (%) e um determinado grau de confiança (%) sobre a regra. Enquanto que o primeiro valor é referente à percentagem de compras em que os produtos antecedentes se comprovaram, o segundo valor é referente à percentagem de compras em que o produto antecedente e consequente se verificaram [10].

De modo a serem criadas regras especificas para cada cliente, as empresas tendem a criar diversos modelos consoante diversos atributos dos clientes (por exemplo, idade, vencimento, género, ...). Para estas análises de divisões é frequentemente usado modelos de divisões dos clientes, através da criação de clusters de clientes.

Com estes clusters, a empresa consegue criar regras específicas para cada tipo de cliente e assim apresentar produtos mais adequados a cada cliente.

### 2.3.3. – Aplicações no mercado

Como já foi referido, a análise de *Market Basket Analysis* é frequentemente usada em cadeias de supermercado, assim como está publicado no artigo de Trnka [13], onde o mesmo usa como conjunto de dados, todas as transações efetuadas durante o espaço temporal de um ano que ocorreram numa cadeia de supermercados específica.

No artigo de Trnka [13], o mesmo apresenta a Figura 6 onde demonstra em forma de tabela os resultados que obteve após a aplicação do modelo de CRISP-DM e do algoritmo de *Apriori*.

Na tabela da Figura 6 é apresentado um conjunto de regras obtidas através do algoritmo de *Apriori*. Cada linha equivale a uma regra encontrada pelo algoritmo.

A primeira regra apresentada indica que os clientes que adquiriram os produtos B, D e F (que aconteceu em 3,47% de todas as transações) têm 95% de confiança de adquirir o produto G.

| Consequência | Antecedente | Suporte % | Confiança % |
|---|---|---|---|
| Produto G | Produto B<br>Produto D<br>Produto F | 3,47 | 95,45 |
| Produto D | Produto B<br>Produto F<br>Produto G | 3,47 | 95,45 |
| Produto D | Produto A<br>Produto F<br>Produto G | 4,42 | 92,86 |
| Produto F | Produto C<br>Produto D<br>Produto G | 2,21 | 92,86 |
| Produto D | Produto E<br>Produto F<br>Produto G | 3,79 | 91,67 |
| Produto F | Produto B<br>Produto D<br>Produto G | 3,63 | 91,3 |
| Produto F | Produto D<br>Produto G | 16,56 | 88,57 |
| Produto G | Produto A<br>Produto D<br>Produto F | 4,73 | 86,67 |
| Produto F | Produto A<br>Produto D<br>Produto G | 4,73 | 86,67 |
| Produto G | Produto C<br>Produto D<br>Produto F | 2,37 | 86,67 |

**Figura 6 – Associações entre produtos (traduzido) (Fonte : Trnka [13] )**

O autor também apresenta os mesmos resultados em formato gráfico, Figura 7, onde é possível visualizar as regras de correlação dos produtos através das ligações entre os mesmos. No exemplo do autor, os produtos ligados entre retas apresentam correlação forte entre os produtos. Na Figura 7 é possível verificar a correlação dos produtos D, F e G e da correlação entre os produtos A e J.

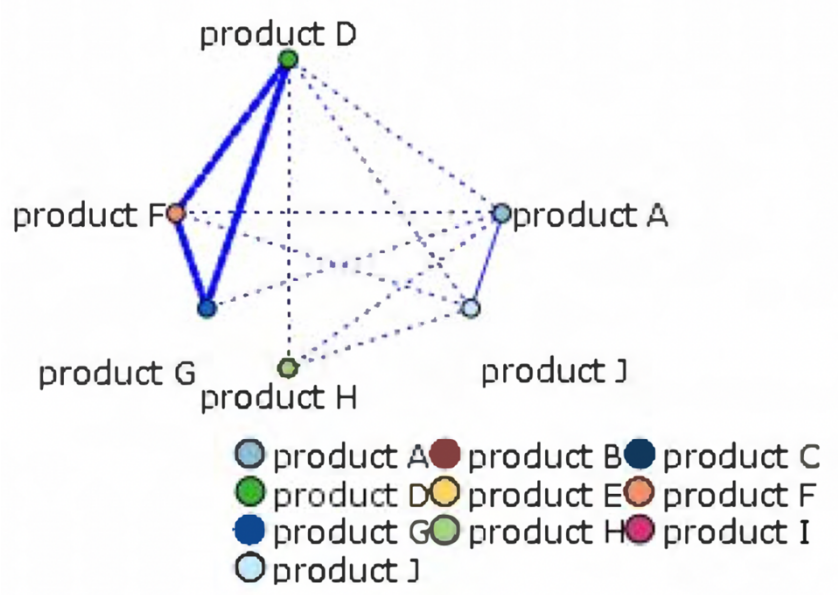



**Figura 7 – Regra de ligação entre produtos (Fonte : Trnka [13] )**

### 2.3.4. – Algoritmos

De forma a obtermos estas análises, é necessário recorrer a algoritmos não supervisionados, como o algoritmo *Apriori*, que é responsável por detetar as regras de associação entre produtos.

Uma vez que este algoritmo é treinado sem ter por base uma variável objetivo conhecida, estes são designados por algoritmos não supervisionados [12].

Para entender o funcionamento deste algoritmo é necessário mencionar os seguintes conceitos:

- *Transaction* – evento efetuado, no caso de supermercados, é visto como uma compra que contempla diversos produtos (itens);
- *Support (A)* – refere-se à percentagem de transações onde o item X foi adquirido;
- *Confidence (A → B)* – refere-se à percentagem de transações onde os itens X e A foram adquiridos, sobre o número total de transações onde o item X foi adquirido.

De modo de funcionamento deste algoritmo começa por determinar os itens frequentemente comprados e de seguida passa por gerar as regras de associação sobre cada item frequentemente comprado, calculando para cada possibilidade os seus valores de *confidence* e de *support* [12].

## *2.4 – Customer Churn*

### 2.4.1. – Análise

No que diz respeito aos modelos de Customer Churn, foi efetuado um levantamento de artigos revelantes com utilização do método PRISMA, conforme descrito por Mohan D. et al [8], que está descrito na Figura 8. Efetuou-se consultas nos repositórios *Scopus* e *IEEE Explore* que contivessem as palavras *Customer Churn Model*, publicados entre 2016 e 2020, referentes às categorias *Computer Science* tendo obtido 330 e 320 documentos respetivamente.

Dando origem à seguinte consulta:

TITLE-ABS-KEY ( customer AND churn AND model ) AND ( LIMIT-TO ( PUBYEAR , 2020 ) OR LIMIT-TO ( PUBYEAR , 2019 ) OR LIMIT-TO ( PUBYEAR , 2018 ) OR LIMIT-TO ( PUBYEAR , 2017 ) OR LIMIT-TO ( PUBYEAR , 2016 ) ) AND ( LIMIT-TO ( SUBJAREA , "COMP" ) )

De seguida, removeram-se os documentos em duplicado, reduzindo para 553 documentos. Dos 533 documentos foi efetuada uma exclusão de 464 documentos através da análise do título dos documentos e pela análise às palavras-chaves que não continham as palavras-chave selecionadas nesta dissertação.

Dos 89 documentos restantes, efetuou-se a leitura dos *abstracts*, tendo resultado numa exclusão de 67 artigos uma vez que não tinham a mesma finalidade desta dissertação.

Após leitura integral dos restantes 22 artigos, obtiveram-se 7 artigos relevantes para o tema em análise.

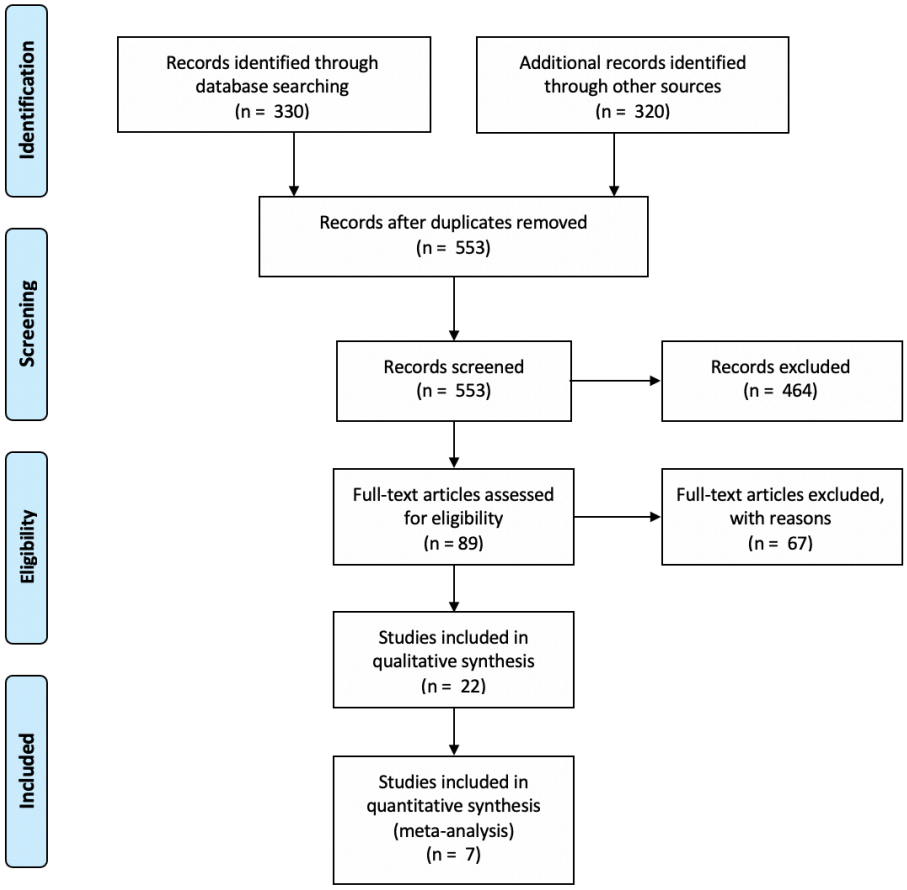



**Figura 8 - Esquema PRISMA efetuado para a pesquisa de documentos relacionados com Customer Churn Model (adaptado de Mohan, D et al [8])**

Recorrendo à ferramenta *VOSviewer* – ferramenta que deteta relações entre artigos – podemos visualizar na Figura 9 que o tema "Customer Churn" tem no centro o tema de retenção. Podemos verificar ainda que este tema reúne três tópicos principais: tópicos de negócio (a vermelho), tópicos de informática (a verde) e tópicos referentes a clientes (a azul). Esta visualização comprova que o tema de retenção de clientes está fortemente relacionada com estes três tópicos.

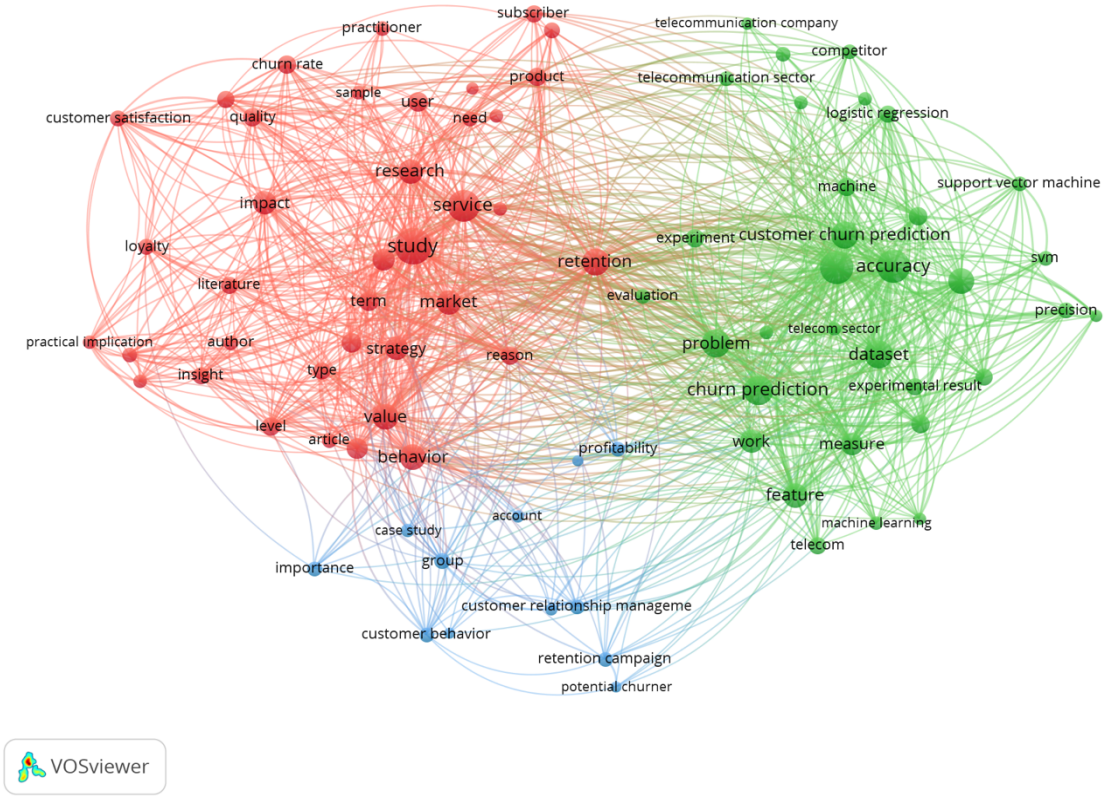



**Figura 9 – Gráfico de relação executado no VOSviewer[9] referente ao tema *Customer Churn***

### 2.4.2. – Definição

Os modelos de *Customer Churn* desempenham um papel fundamental no *Customer Relationship Management (CRM)* de uma companhia. O seu principal objetivo é conseguir aumentar a taxa de retenção dos seus clientes.

As companhias contêm registos de todos os movimentos (quer sejam emissões, renovações, alterações, anulações, entre outros) que aconteceram sobre as apólices dos seus clientes. Todos esses dados contêm muita informação, que através de modelos de *Customer Churn*, permitem à companhia identificar clientes que se encontram em vias de anular as suas apólices.

Estes modelos são frequentemente usados nos setores de telecomunicações, financeiro e segurador, isto deve-se ao facto de serem modelos de negócio onde o custo de reter um cliente é muito inferior ao de conseguir captar um novo cliente [14].

### 2.4.3 – Artigos já existentes

He et al, publicaram um artigo em que usam como conjunto de dados dados de uma companhia seguradora [15]. Neste artigo, segundo os autores, é possível verificar que existem

variáveis que estão fortemente relacionadas com a variável objetivo – apólice renovada. Estas variáveis encontram-se representadas na Figura 10.

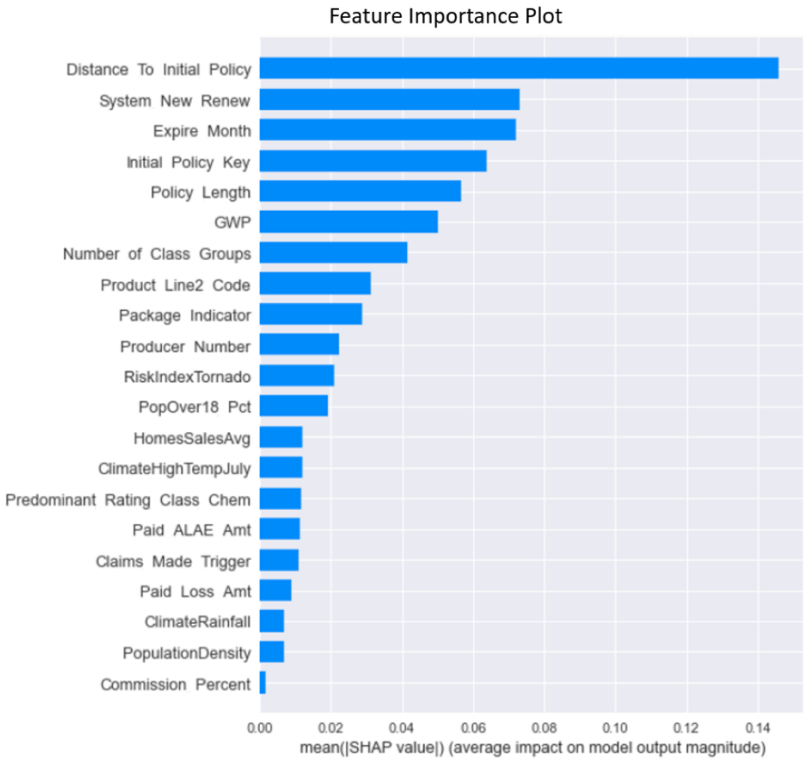



**Figura 10 – Importância das variáveis em relação à variável objetivo (Fonte: He et al [15])**

M. Spiteri e G. Azzopardi publicaram um artigo onde é igualmente aplicado um modelo de retenção de clientes no mercado segurador [16]. Neste artigo os autores apresentam quais as variáveis mais relacionadas com a variável objetivo, Figura 11.

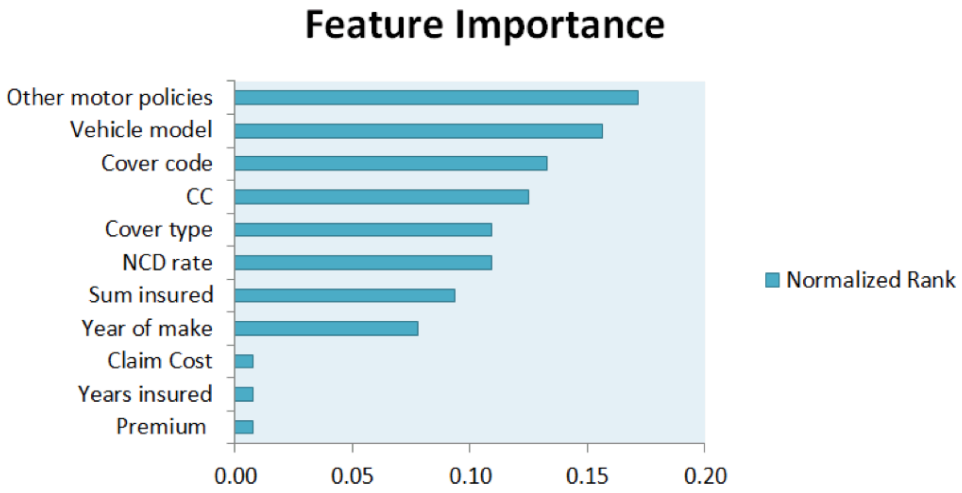



**Figura 11 – Importância das variáveis em relação à variável objetivo (Fonte: M. Spiteri e G. Azzopardi [16])**

Os autores deste artigo publicaram igualmente quais os resultados de acerto obtidos para os diversos algoritmos que utilizaram na criação do modelo de retenção de clientes.

| Algoritmo | Taxa de acerto - Todas as variáveis | Taxa de acerto – Variáveis Selecionadas |
|---|---|---|
| Random Forest | 89.90% | 88.33% |
| Naive Bayes | 88.33% | 87.75% |
| Decision Tree | 88.23% | 88.37% |
| SVM Radial | 81.36% | 49.63% |
| Logistic Regression | 75.23% | 70.65% |
| SVM Linear | 74.65% | 72.76% |

**Figura 12 – Valores de acerto obtidos com diversos algoritmos (traduzido) (Fonte: M. Spiteri e G. Azzopardi [16])**

Através da Figura 12 é ainda possível observar que os autores aplicaram os diversos algoritmos ao conjunto de dados com todas as variáveis – coluna Accuracy – All Variables – e também aplicaram os mesmos algoritmos a um conjunto de dados com apenas as variáveis com mais correlação com a variável objetivo – coluna Accuracy – Selected Variables.

É também possível de verificar que o algoritmo com melhor taxa de acerto foi com recurso de *Random Forest*, tendo obtido uma taxa de 89,90%.

Após a obtenção do melhor algoritmo a usar, os autores obtiveram a matriz de confusão para o modelo que obtiveram, que é demonstrado na Figura 13.

Através da matriz de confusão é possível verificar a quantidade de valores preditos corretamente e a quantidade de valores preditos incorretamente. No caso da Figura 13, o algoritmo de *Random Forest* previu corretamente 19539 (17837 + 1702) registos e previu com o valor incorreto 2194 (814 + 1380) registos.

| | **Renewed** | **Not Renewed** |
|---|---|---|
| **Renewed** | 17,837 | 814 |
| **Not Renewed** | 1,380 | 1,702 |

**Figura 13 – Matriz de Confusão para o algoritmo Random Forest (Fonte: M. Spiteri e G. Azzopardi [16])**

Na Tabela 1, é apresentado um resumo de vários artigos referentes ao tema, a área de atuação da empresa (telecomunicações, seguros, ...) e quais os algoritmos que foram usados.

| Artigo | Área | Algoritmos usados |
|---|---|---|
| Ullah, I et al.[17] | Telecomunicações | J48, Random Forest, Attribute Selected Classifier |
| Keramati, A. et al [18] | Banca | Decision Tree |
| Gordini, N et al [19] | E-commerce | Redes Neuronais, Support Vector Machine, Regressão Logística |
| Miloševic, M et al [20] | Jogos online | Regressão Logística, Naive Bayes, Decision Tree, Gradient Boosting, Random Forest |
| Ahmad, A.K. et al[21] | Telecomunicações | XGBOOST, GSM, Random Forest, Decision Tree |
| He, Y. et al [15] | Seguros | Regressão Logística, Random Forest, Support Vector Machine, Gradient Boosting, Redes Neuronais |

**Tabela 1 – Artigos sobre Customer Churn**

### 2.4.4 – Aprendizagem Supervisionada

Os modelos de *Customer Churn* são criados através de algoritmos de aprendizagem automática. A aprendizagem automática, *Machine Learning* em inglês, refere-se a modelos criados através de máquinas que aprendem com um conjunto inicial de dados.

Os modelos de *Machine Learning* podem ser supervisionados ou não supervisionados. Os modelos supervisionados, no momento de aprendizagem, usam uma variável objetivo conhecida. Nos modelos não supervisionados não se conhece qual é a variável objetivo.

Nos modelos de *Customer Churn* são usados modelos supervisionados, uma vez que sabemos qual é a variável objetivo, que se prende com o facto de o cliente ter, ou não, anulado a apólice.

# Capítulo 3 – Análise de afinidade de produtos

Neste capítulo irá ser abordado a criação das regras de afinidade entre produtos. Irá começar com uma explicação dos dados que foram extraídos e com que filtros foram aplicados. De seguida, irá ser demonstrado algumas análises que foram executadas, entre elas, normalizações de dados, análise de gráficos e a criação das zonas geográficas e de clusters. Por fim, irá ser demonstrado a criação das regras de afinidade dos produtos com recurso ao algoritmo *Apriori* e a visualização das regras graficamente.

Para a criação deste artefacto adotou-se a metodologia CRISP-DM. As criações destes modelos de aprendizagem automática costumam seguir uma metodologia CRISP-DM. Na Figura 14 é demonstrado o ciclo de fases que esta metodologia contém.

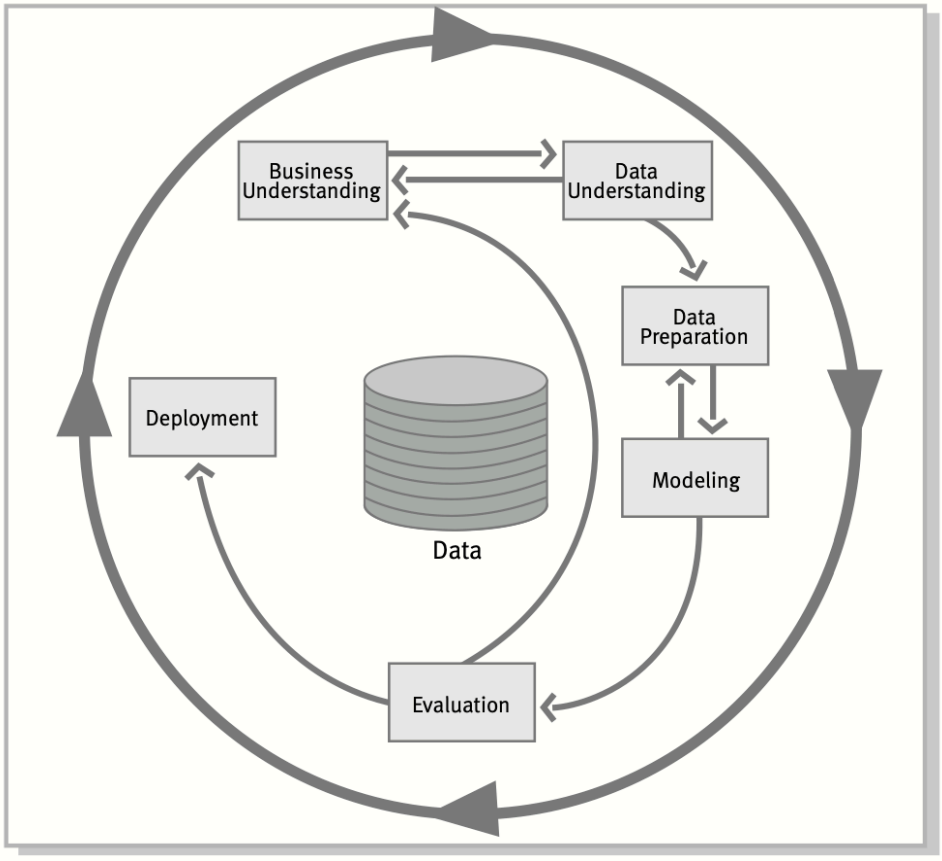



**Figura 14 - Etapas Metodologia CRISP-DM (Fonte: CRISP-DM [22])**

De forma a obtermos um bom modelo é fundamental cumprir todas as etapas representadas na Figura 14. Começando pela fase de ***Business Understanding*** (compreender o negócio) e que está muito relacionado com a fase de ***Data Understanding*** (consiste em compreender os dados e analisá-los, relacionando-os com os conceitos de negócio que decorrem da fase anterior).

Após a etapa de análise, surge a etapa de ***Data Preparation***, esta etapa compreende a fase de tratamento e correção dos dados. Exemplos desta etapa são normalizações de dados, correcção ou tratamento de valores (e.g., tratamento de *outliers*) ou criação de intervalos, entre outros.

Com os dados já tratados, segue-se a fase de ***Modeling***, onde se aplica o conjunto de dados a diversos algoritmos e com possíveis *tunnings* (alterações de parâmetros).

Após a obtenção de um bom modelo com o conjunto de treino. É aplicado o mesmo modelo a um conjunto de teste para se efetuar a fase de ***Evaluation***, onde se poderá analisar se o modelo desenvolvido cumpre os requisitos ou se será necessário rever todas as etapas anteriores.

Caso o modelo desenvolvido cumpra os requisitos, é chegado à última etapa da metodologia CRISP-DM, ***Deployment***, onde o modelo passa a ser usado com a sua finalidade final.

## *3.1* – Compreensão do Negócio

Para efetuar a análise entre produtos, começou-se por verificar junto dos responsáveis da companhia seguradora, os dados disponíveis e selecionar as variáveis que iriam ser úteis no modelo de afinidade de produtos - MBA.

Ficou acordado analisar a afinidade de produtos respeitante aos tomadores que efetuaram novas apólices nos dois anos anteriores (2018 e 2019). Esta decisão de espaço temporal foi baseada no fator de se tratar de anos completos, para não incluir dados de anos parciais e para que incluísse uma boa amostragem de tomadores (clientes) ativos.

Foi definido analisar apenas clientes com NIF (a título individual), ou seja, excluir apólices em que o tomador seja um NPC (Número de Pessoa Coletiva), uma vez que o cerne desta dissertação se foca em reter os clientes individuais.

Foi também decidido efetuar a encriptação de todos os dados sensíveis, tais como o id do cliente. Para tal foi aplicado sobre estes dados um algoritmo que impossibilita a reversão dos dados.

Definido o espaço temporal e os critérios de clientes, foi necessário definir os dados a extrair.

Para cada um dos tomadores que se encontram dentro dos critérios definidos, foi necessário extrair as suas apólices, novas assim como as apólices que o tomador adquiriu no passado, de modo a obter toda a sua caminhada dentro da companhia.

Existem dois tipos de dados que são necessários para efetuar a análise de afinidade: dados referentes à transação efetuada (neste caso, apólices emitidas) e dados referentes ao cliente que efetuou a transação (data de nascimento, sexo, distrito, *etc*).

Em relação aos dados referentes à transação, decidiu-se utilizar os seguintes dados: ramo da transação efetuada, a modalidade escolhida e a data da emissão da apólice.

No que se refere aos dados do cliente selecionou-se o id do cliente (encriptado), a data de nascimento, o estado civil, o sexo e o distrito.

Como métrica, foi usado uma variável numérica Quantidade – sendo a mesma preenchida com o valor de apólices de determinado ramo, que um determinado cliente adquiriu.

## *3.2* – Compreensão dos Dados

Como passo seguinte, efetuou-se uma análise sobre a localização destas variáveis pelas diversas tabelas do sistema transacional com a finalidade de obter o *dataset* necessário. Para tal, efetuou-se a análise que gerou o seguinte diagrama de tabelas demonstrado na Figura 15.

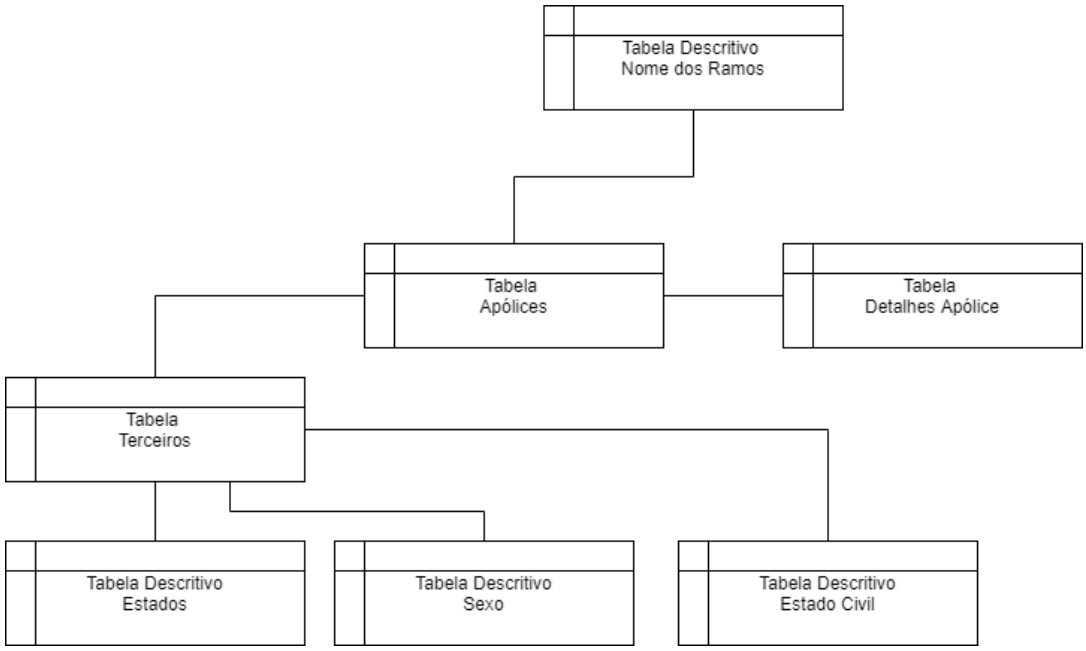



**Figura 15 - Diagrama de tabelas**

Foi verificado que existiam valores vazios para a coluna Estado Civil, após questionado à equipa responsável foi decidido corrigir o valor para "S" – Solteiro, uma vez que este valor é assumido por defeito – de modo a evitar possíveis conflitos com os clientes.

Após realização da *query* SQL, que liga todas as variáveis, avançou-se para a extração da mesma e análise dos dados, conforme demonstrado de seguida.

Numa primeira análise ao dataset do modelo de afinidade de produtos constatou-se que o mesmo contém 65532 registos, com 10 colunas.

Começou-se por verificar que não existiam valores vazios assim como valores inválidos uma vez que o tratamento dos mesmos foi efetuado durante a extração dos dados na *query* SQL.

De modo a efetuar análises gráficas, importou-se este conjunto de dados na ferramenta *Tableau*.

Na Figura 16 é possível visualizar a distribuição de apólices emitidas por cada distrito existente. Através da Figura 16 é possível visualizar que não se encontra uma distribuição equivalente por cada distrito. Havendo uma maior concentração no distrito de Lisboa.

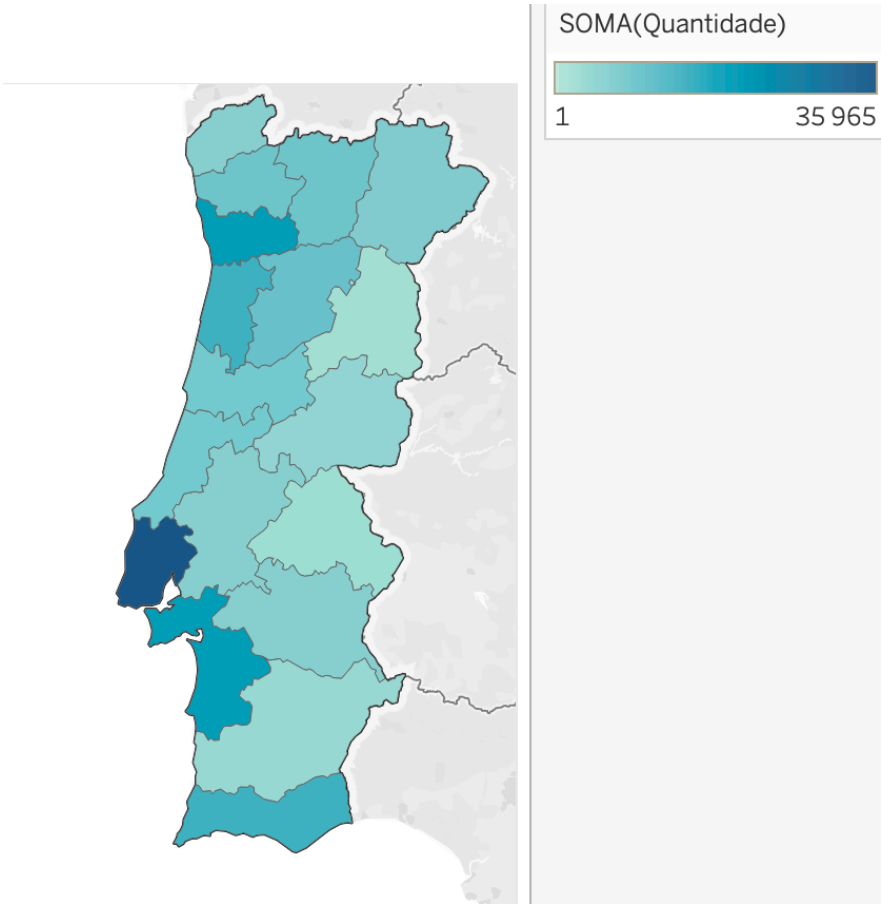



**Figura 16 - Distribuição de clientes por distrito de residência (com recurso à ferramenta *Tableau*)**

Assim como foi demonstrado na Figura 3, na Figura 17 é demonstrado a distribuição das apólices pelos diversos ramos.

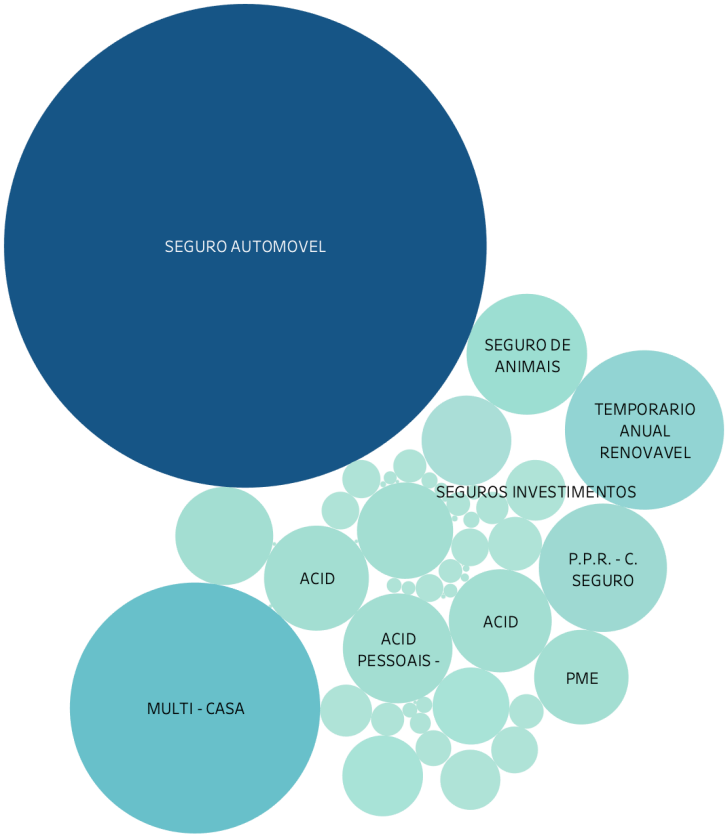



**Figura 17 - Distribuição de apólices por ramo contratado (com recurso à ferramenta *Tableau*)**

Na Figura 17, cada círculo representa um ramo existente na companhia. O tamanho do círculo é correspondente ao número de apólices existentes para esse ramo. Conseguimos verificar que existem dois ramos que se destacam em relação aos restantes. Verificamos igualmente que existem muitos ramos com quantidades muito reduzidas de apólices.

Na Figura 18 é apresentado um gráfico onde indica a distribuição das apólices por mês da emissão das apólices. Após a análise da figura, verifica-se que existem diversos picos ao longo do ano, não sendo possível retirar nenhuma conclusão sobre a linearidade das emissão das apólices ao longo do ano.

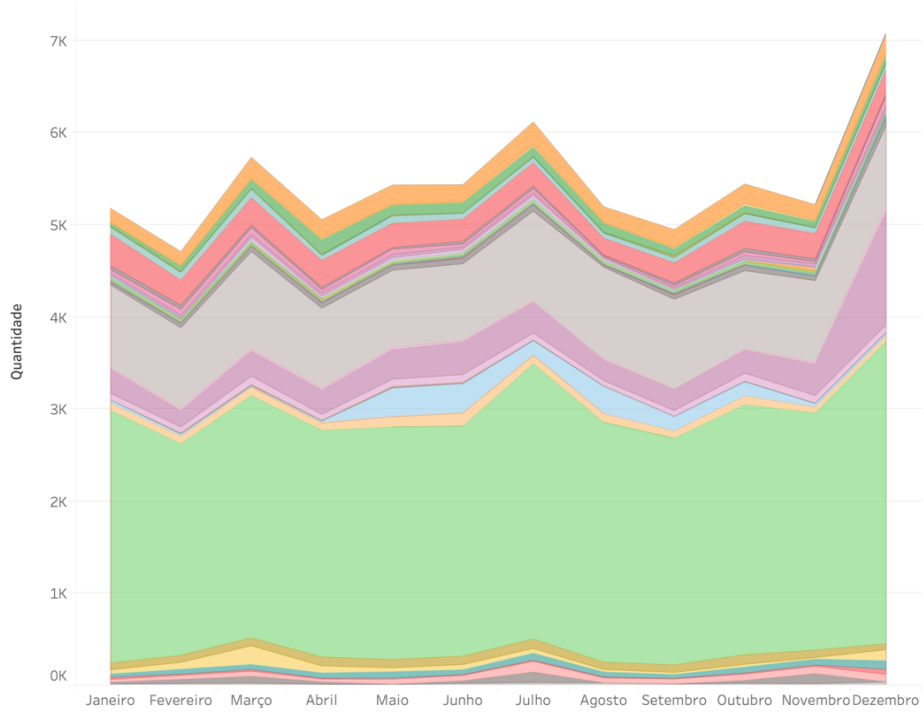



**Figura 18 - Distribuição de apólices por mês de criação (com recurso à ferramenta *Tableau*)**

## *3.3* – Preparação dos Dados

Tendo analisado os dados anteriores, foi decidido efetuar-se algumas normalizações de modo a balancear algumas variáveis com a finalidade de reduzir o número de valores possíveis existentes.

Verificou-se que existiam vários ramos que pertenciam ao mesmo ramo principal, assim normalizaram-se esses valores agrupando os vários ramos ao ramo principal. Na Figura 19 é possível verificar que existem diversos ramos que são pertencentes ao ramo de Acidentes Pessoais. Deste modo foi possível agrupar diversos ramos que estão relacionados entre si e reduzindo o número de produtos existentes.

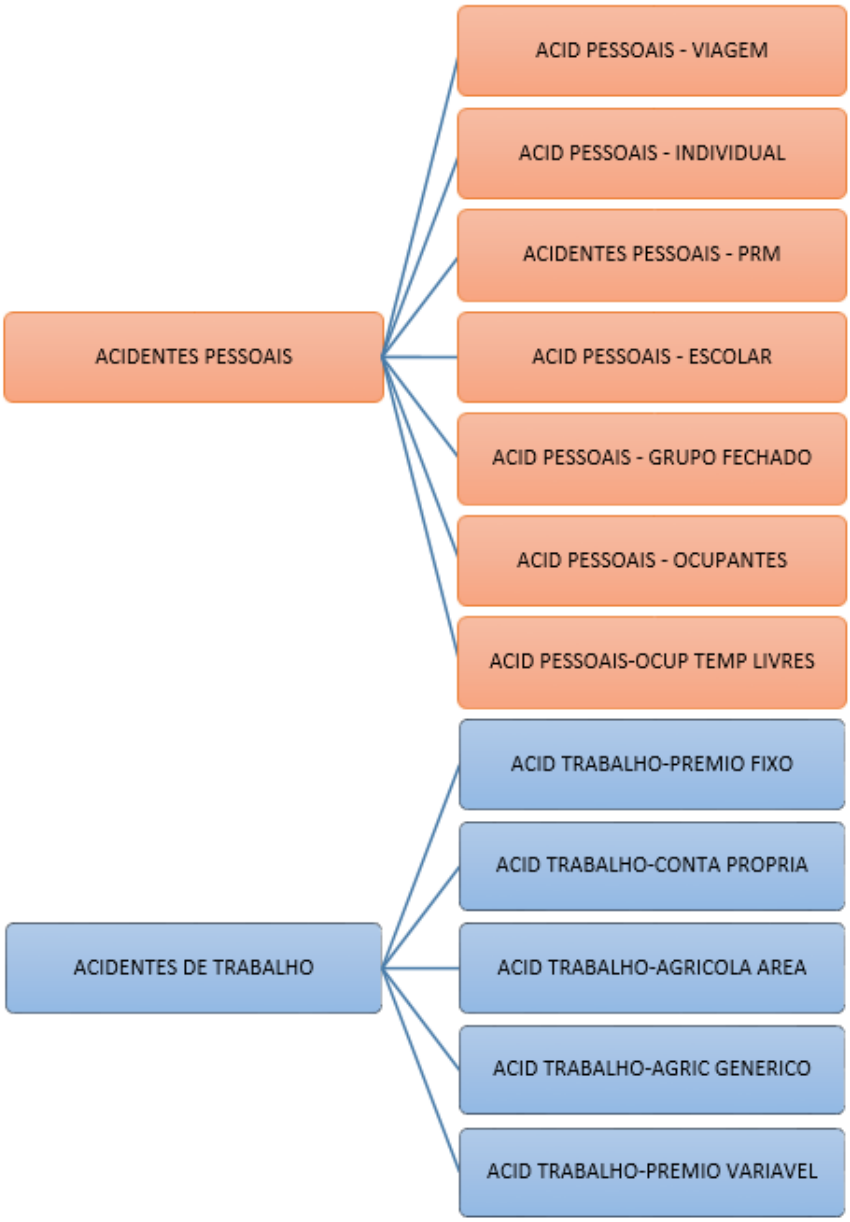
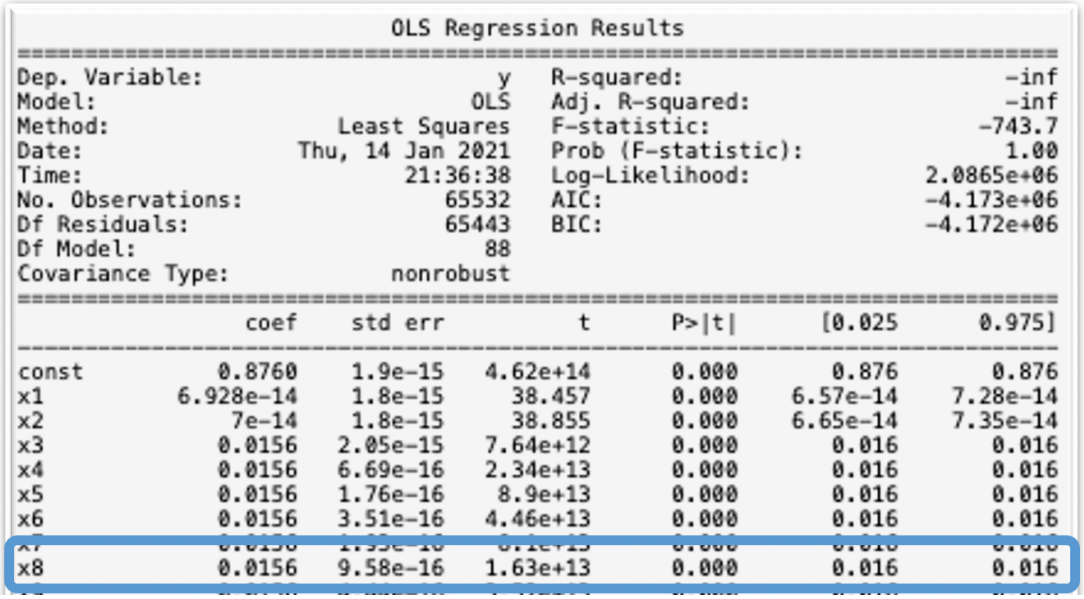



**Figura 19 - Exemplo de normalizações efetuadas ao Nome dos Ramos**

De seguida, foi validado que dado a desproporção de apólices pelos diversos distritos se justificava agrupar por zonas geográficas, com o objetivo de termos grupos mais equitativos. Para tal foi efetuado uma regressão linear tendo como variável objetivo a quantidade de apólices por cada zona para verificar a viabilidade desta opção.



OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | y | R-squared: | -inf |
| Model: | OLS | Adj. R-squared: | -inf |
| Method: | Least Squares | F-statistic: | -743.7 |
| Date: | Thu, 14 Jan 2021 | Prob (F-statistic): | 1.00 |
| Time: | 21:36:38 | Log-Likelihood: | 2.0865e+06 |
| No. Observations: | 65532 | AIC: | -4.173e+06 |
| Df Residuals: | 65443 | BIC: | -4.172e+06 |
| Df Model: | 88 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 0.8760 | 1.9e-15 | 4.62e+14 | 0.000 | 0.876 | 0.876 |
| x1 | 6.928e-14 | 1.8e-15 | 38.457 | 0.000 | 6.57e-14 | 7.28e-14 |
| x2 | 7e-14 | 1.8e-15 | 38.855 | 0.000 | 6.65e-14 | 7.35e-14 |
| x3 | 0.0156 | 2.05e-15 | 7.64e+12 | 0.000 | 0.016 | 0.016 |
| x4 | 0.0156 | 6.69e-16 | 2.34e+13 | 0.000 | 0.016 | 0.016 |
| x5 | 0.0156 | 1.76e-16 | 8.9e+13 | 0.000 | 0.016 | 0.016 |
| x6 | 0.0156 | 3.51e-16 | 4.46e+13 | 0.000 | 0.016 | 0.016 |
| x8 | 0.0156 | 9.58e-16 | 1.63e+13 | 0.000 | 0.016 | 0.016 |

**Figura 20 - Regressão linear com base na variável Zona**

Uma vez que o valor de *p-value*, destacado na Figura 20, foi abaixo do intervalo de 0.05% podemos rejeitar a hipótese nula e verificamos que a zona é uma variável que contribui para a quantidade de apólices existentes. Assim o dataset passou a ter 7 zonas distintas:

- Norte (constituído por Porto, Viana do Castelo, Braga, Vila Real e Bragança);
- Centro (constituído por Aveiro, Coimbra, Viseu, Guarda e Castelo Branco);
- Grande Lisboa (constituído por Leiria, Lisboa e Santarém);
- Alentejo (constituído por Portalegre, Évora, Setúbal e Beja);
- Sul (constituído por Faro);
- Ilhas (constituído por Madeira e Açores);
- Estrangeiro (constituído pelos restantes valores).

Foi ainda necessário efetuar uma conversão *One Hot Encoding* – onde cada valor existente da variável é transformado numa nova coluna com valor 1 ou 0 (existindo no registo ou não existindo no registo, respetivamente) – para cada variável com mais de dois valores categóricos possíveis. Um exemplo desta conversão é a variável Estado Civil, que deu origem às seguintes colunas "Estado_Solteiro", "Estado_Casado", "Estado_Divorciado", "Estado_Uniao_Facto", "Estado_Viuvo".

Para a variável de início da apólice, foi criado uma nova coluna com o mês do início da apólice e uma variável com o ano de início da apólice.

## *3.4* – Modelação

Após o tratamento efetuado ao dataset, foram obtidos sete conjuntos distintos – onde cada conjunto correspondia a uma das zonas geográficas definidas.

Para cada zona geográfica, foi aplicado um algoritmo de *clustering*, *AgglomerativeClustering*, para conseguir analisar os diferentes tipos de clientes que existem de modo a efetuar regras de afinidade de produtos mais específicas. Uma vez que o algoritmo de Cluster atribui apenas o cluster correspondente a cada registo foi necessário aplicar o algoritmo de Árvore de Decisão para obter as regras de divisão para cada cluster. Na Figura 21 é possível obter a árvore de decisão para a zona geográfica do Norte.

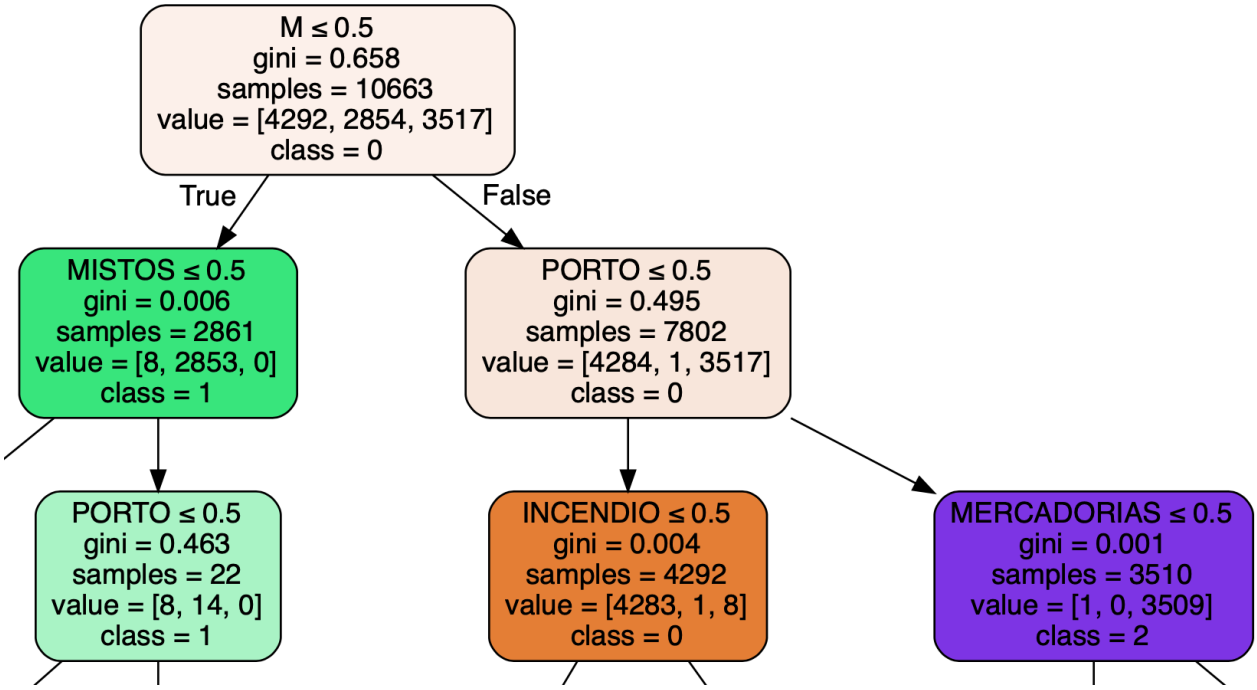



**Figura 21 - Árvore de Decisão com divisão de clusters para a zona Norte**

Através da árvore de decisão, para a zona Norte é possível visualizar as suas regras de divisão (através da Classe correspondente):

- Cluster 0 Norte – Clientes do Sexo Masculino residentes nos distritos Viana do Castelo, Braga, Vila Real e Bragança;
- Cluster 1 Norte – Clientes do Sexo Feminino;
- Cluster 2 Norte – Clientes do Sexo Masculino e residentes no distrito do Porto;

## *3.5* – Avaliação

Com base no *dataset* gerado, o mesmo foi dividido em dois subconjuntos: um sub-conjunto de treino com 70% dos registos e um sub-conjunto de teste com 30% dos registos, divididos aleatoriamente.

A razão desta divisão serve para o próprio algoritmo conseguir elaborar um modelo com base no sub-conjunto de treino e avaliar o modelo obtido sobre um sub-conjunto de dados distinto, o sub-conjunto de teste.

Após esta divisão, avançou-se com a aplicação do algoritmo Apriori para cada cluster identificado no passo anterior.

Para cada cluster, após a aplicação do algoritmo *Apriori,* é identificado as associações de produtos existentes para esse cluster, assim como o seu nível de suporte e o nível de confiança de cada regra associada.

Foi definido junto da equipa responsável definir um valor de nível de suporte mínimo de 2%.

Na Figura 22 é demonstrado as regras obtidas pelo algoritmo *Apriori* para o Cluster 0 da zona Norte.

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|---|
| 21 | (MULTI - CASA, RESPONSABILIDADE CIVIL GERAL) | (SEGURO AUTOMOVEL) | 0.022634 | 0.784636 | 0.021948 | 0.969697 | 1.235855 |
| 9 | (MULTI - CASA, ACIDENTES DE TRABALHO) | (SEGURO AUTOMOVEL) | 0.041838 | 0.784636 | 0.039095 | 0.934426 | 1.190903 |
| 14 | (MULTI - CASA, ACIDENTES PESSOAIS) | (SEGURO AUTOMOVEL) | 0.037723 | 0.784636 | 0.032922 | 0.872727 | 1.112270 |
| 22 | (SEGURO AUTOMOVEL, RESPONSABILIDADE CIVIL GERAL) | (MULTI - CASA) | 0.038409 | 0.336077 | 0.021948 | 0.571429 | 1.700292 |
| 10 | (SEGURO AUTOMOVEL, ACIDENTES DE TRABALHO) | (MULTI - CASA) | 0.078189 | 0.336077 | 0.039095 | 0.500000 | 1.487755 |
| 16 | (ACIDENTES PESSOAIS, SEGURO AUTOMOVEL) | (MULTI - CASA) | 0.077503 | 0.336077 | 0.032922 | 0.424779 | 1.263934 |

**Figura 22 - Regras de Afinidade de Produtos para o Cluster 0 Norte**

Assim como foi demonstrado na Figura 7, para cada cluster elaborado foi igualmente criado as regras de afinidade de produtos em formato gráfico, como é possível visualizar na Figura 23.

Na Figura 23, cada ponto vermelho representa uma regra de afinidade definida, cada ponto preto representa um ramo da companhia.

De modo a verificar uma regra de afinidade é necessário verificar quais as setas que estão na direção ramo → regra (representa um ramo antecedente) e verificar quais as setas que estão na direção regra → ramo (representa um ramo consequente). Cada regra é definida por uma cor específica.

Deste modo, é possível verificar de um modo mais eficiente quais as regras que se aplicam tendo por base quais os ramos que o cliente já possui.

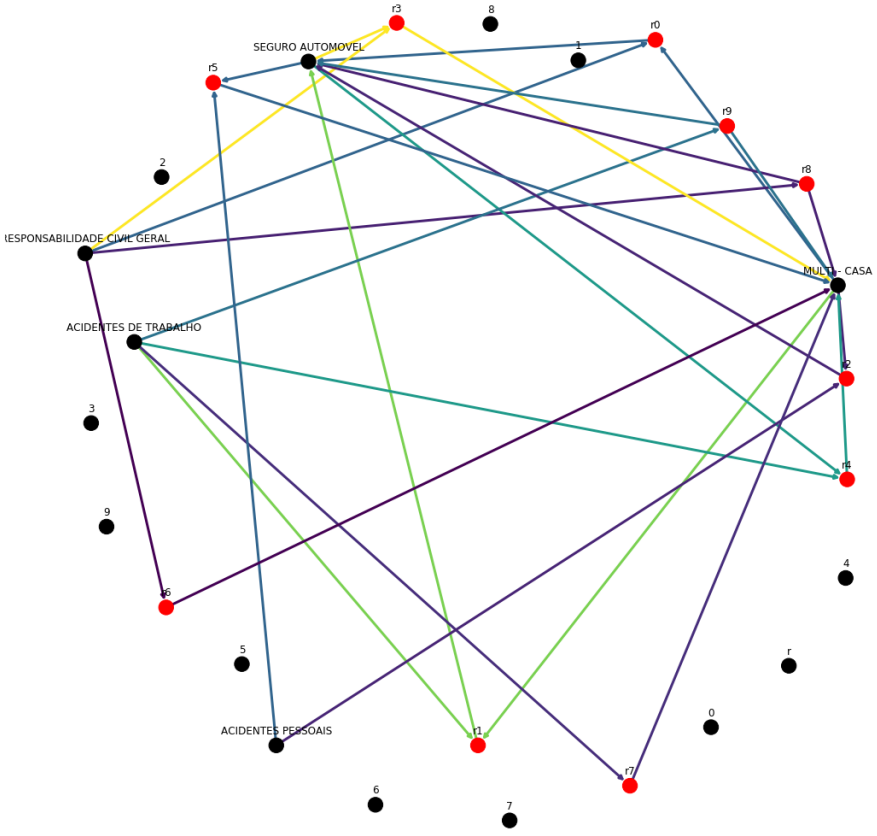



**Figura 23 - Regras de Afinidade entre Produtos para o Cluster 0 Norte em formato gráfico**

No total foram efetuados 21 clusters e um conjunto de regras de afinidade para cada um dos clusters definidos.

Todos estes clusters e suas respetivas regras foram compiladas num manual, em formato PDF, e transmitidas à equipa responsável pela criação de campanhas específicas com o objetivo de efetuar campanhas específicas de cross-selling.

No Capítulo 5, irá ser feita a demonstração destas regras obtidas, juntamente com as avaliações obtidas junto dos avaliadores.

# Capítulo 4 – *Customer Churn Model*

Neste capítulo irá ser detalhado a criação do modelo de retenção de clientes. Irá começar com uma explicação dos dados que foram extraídos e com que filtros foram aplicados. De seguida, irá ser demonstrado algumas análises que foram executadas, entre elas, normalizações de dados. Por fim, irá ser demonstrado a criação do modelo e os seus resultados.

Para a criação deste segundo artefacto adotou-se a mesma metodologia que no capítulo anterior, ou seja, a metodologia CRISP-DM.

## *4.1* – Compreensão do Negócio

Para efetuar o modelo de *Churn*, aplicado ao ramo automóvel, começou-se por reunir com a equipa responsável de Desenvolvimento de Produtos, por forma a entender o negócio em si, ou seja, a filosofia de como os dados se interligam, e numa fase seguinte analisou-se todos os dados que existem relacionados às apólices deste ramo, de modo a entender o que poderá ser possível de ser extraído para o *dataset*.

Destas primeiras reuniões, foi demonstrado como se emitia apólices de automóvel no sistema, como se consultava determinados dados de cada apólice, inclusive o significado de cada variável. Com toda esta informação começou a ser possível entender toda a interligação entre os dados em si.

Das reuniões iniciais ficou ainda decidido que iria ser usado como limite temporal todas as apólices vigentes no ano de 2019 e todas as apólices anuladas no ano de 2019. No que se refere a clientes iremos apenas atuar nos clientes a nível individual, excluindo assim os clientes empresariais. Foi igualmente decidido, direcionar esta ferramenta apenas para os clientes particulares, resultando assim nas seguintes modalidades:

- Responsabilidade Civil (Através de agente);
- Responsabilidade Civil + (Através de agente);
- Danos Próprios (Através de agente);
- Responsabilidade Civil (Contratado Online);
- Responsabilidade Civil + (Contratado Online);
- Danos Próprios (Contratado Online).

Excluindo-se assim, todas as modalidades destinadas a clientes empresariais ou campanhas para os mesmos.

Verificou-se ainda que existem diversos motivos de anulação, desse modo, efetuou-se uma análise dos motivos existentes, uma vez que muitos dos motivos de anulação eram por iniciativa da companhia. Uma vez que o foco do modelo incide sobre detetar as anulações por parte do cliente, não seria benéfico incluir anulações por iniciativa da companhia no nosso modelo.

| # ANULAÇÃO | MOTIVO ANULAÇÃO |
|---|---|
| 26 | ANUL - TOM.N.ACEIT.COND.D.INEX |
| 27 | ANUL - TOM.N.ACEIT.COND.AG.RIS |
| 32 | ANUL - INICIATIVA CIA DEC.INEX |
| 33 | ANUL - INICIATIVA CIA AG.RIS |
| 12 | ANULACAO - POR FURTO |
| 13 | ANULACAO - FALSAS DECLARACOES |
| 14 | ANULACAO - NOVA LEI |
| 31 | ANULACAO - POR PERDA TOTAL |
| 4 | ANUL - INICIATIVA DA COMPANHIA |
| 43 | APOLICES EXPIRADAS |
| 5 | ANUL - FALECIMENTO DO TOMADOR |
| 6 | ANULACAO - POR SUBSTITUICAO |
| 7 | ANULACAO - VENDA DO BEM SEGURO |
| 8 | ANULACAO - A PEDIDO DO TOMADOR |
| 9 | ANUL - ERRO EMISSAO COMPANHIA |

**Figura 24 - Motivos de Anulação após seleção**

Na Figura 24, são demonstrados todos os motivos de anulação possíveis para uma apólice do ramo automóvel. Após análise, optou-se por usar no modelo de *churn*, as anulações destacadas a verde, descartando assim as anulações que eram por iniciativa da companhia – destacadas a vermelho.

Com esta filtragem, evitou-se a inserção de dados que iriam desvirtuar o foco deste modelo.

## *4.2* – Compreensão dos Dados

No que se refere à etapa de *Data Understanding*, começou-se por segmentar a análise de variáveis pelas seguintes áreas:

- Variáveis genéricas relacionadas com a apólice;
- Variáveis relacionadas com o tomador da apólice;
- Variáveis relacionadas com o agente da apólice;
- Variáveis relacionadas com o histórico de sinistros da apólice;
- Variáveis relacionadas com o histórico de recibos da apólice;
- Variáveis relacionadas com o veículo assegurado;

- Variáveis relacionadas com o condutor do veículo.

Após definir as diferentes áreas que estão relacionadas com a apólice, avançou-se com o levantamento de todas as variáveis relacionadas com cada uma destas áreas. Deste modo, permitiu-se focar em cada sessão uma área em específico.

No que se refere a **Variáveis genéricas relacionadas com a apólice**, foram definidas como possíveis variáveis: o estado da apólice (vigente ou anulada); o número da apólice; a data de início da apólice; o último tipo de suplemento que foi efetuado (renovação, alteração, anulação); o número de suplementos efetuados; o número de riscos associados; se tem alguma apólice de grupo associada; o fracionamento da apólice; o número de renovações.

Em relação às **Variáveis relacionadas com o tomador da apólice**, foram definidas como possíveis variáveis: o código do cliente (encriptado); se é cliente físico ou empresarial; a nacionalidade; a localidade de residência; o país de residência; a data de nascimento; a data de carta de condução; a profissão; o seu estado civil; o sexo.

Em relação às **Variáveis relacionadas com o agente da apólice**, foram definidas como possíveis variáveis: o código do agente; o gestor responsável; o nível organizacional associado ao agente.

Em relação às **Variáveis relacionadas com o histórico de sinistros da apólice**, foram definidas como possíveis variáveis: número de sinistros associados; número de sinistros associados em que está definido como culpado.

Em relação às **Variáveis relacionadas com o histórico de recibos da apólice**, foram definidas como possíveis variáveis: quantidade de prémios pagos pelo tomador.

Em relação às **Variáveis relacionadas com o veículo assegurado**, foram definidas como possíveis variáveis: marca; modelo; se contém danos no veículo; o tipo de carroceria; o tipo de combustível; a cilindrada; o peso; a potência; o valor do veiculo; a modalidade; se é importado; país onde vai conduzir; a cor; o estado onde vai conduzir; a província onde vai conduzir; o uso do veículo; a data de primeira matrícula; a data da inspeção; número de quilómetros; a categoria do veículo; a assistência em viagem contratada; ano em que o veículo pertenceu a este condutor.

Em relação às **Variáveis relacionadas com o condutor do veículo**, foram definidas como possíveis variáveis: a data de nascimento do condutor; a data de carta de condução; o sexo.

Após estas coletânea de variáveis possíveis foi necessário analisar onde seria possível obter estes dados, vindos de uma base de dados normalizada, de modo a obter um *dataset* de uma apólice por registo. Para tal, efetuou-se uma análise e um posterior diagrama de ligações entre

tabelas, demonstrado na Figura 25, que seriam necessárias de cruzar entre si, por forma a obtermos os dados necessários.

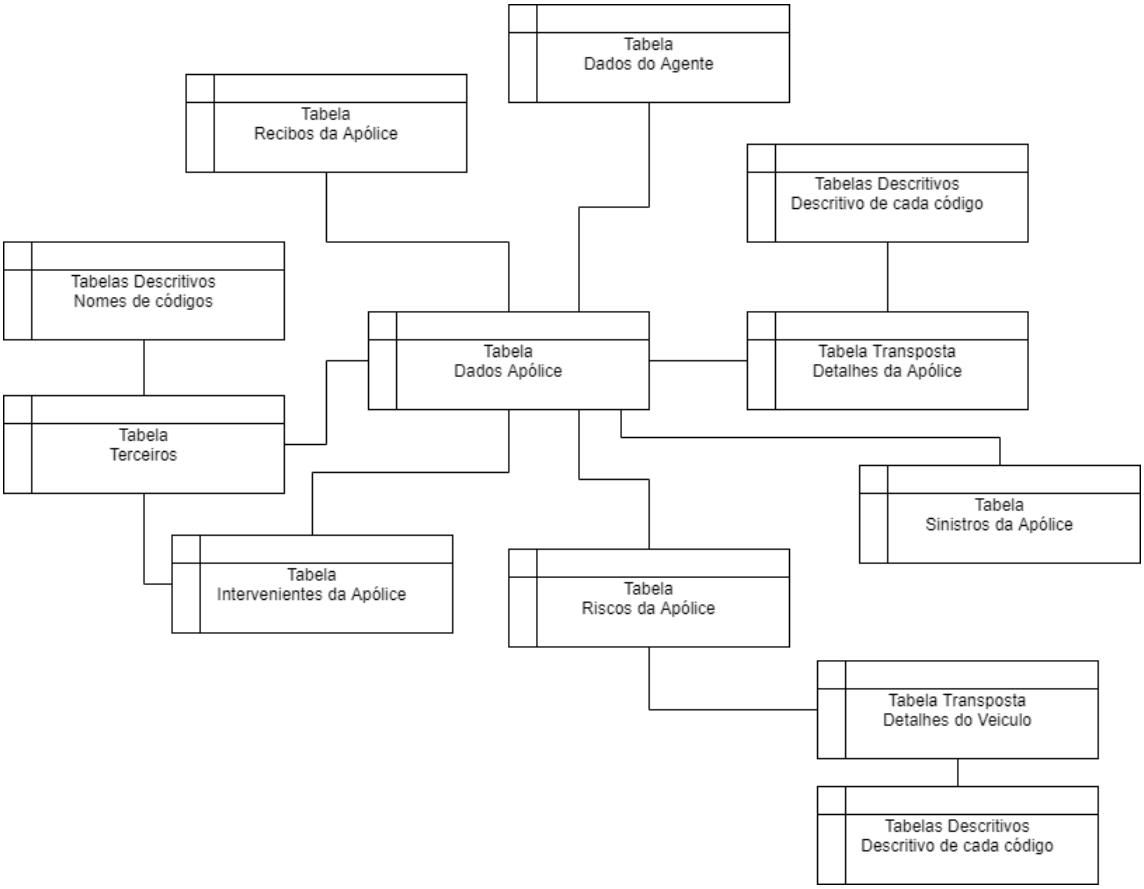



**Figura 25 - Diagrama de tabelas para elaboração do *dataset***

Como é possível verificar na Figura 25, tratou-se de uma *query* complexa, e que para obter um custo menor, analisou-se e tirou-se vantagem de todos os índices associados às tabelas por forma a obter uma consulta eficaz e com o mínimo de custos para a base de dados transacional.

Com base no conjunto de dados extraídos, foram efetuadas diversas análises gráficas.

A primeira análise foi tendo como base a quantidade de apólices vigentes e apólices anuladas no ano de 2019, sendo que se constatou um grande desequilíbrio entre os dois valores, demonstrado na Figura 26. Para tal, foi planeado efetuar técnicas de *oversampling* e *undersampling* de modo a equilibrar os valores.

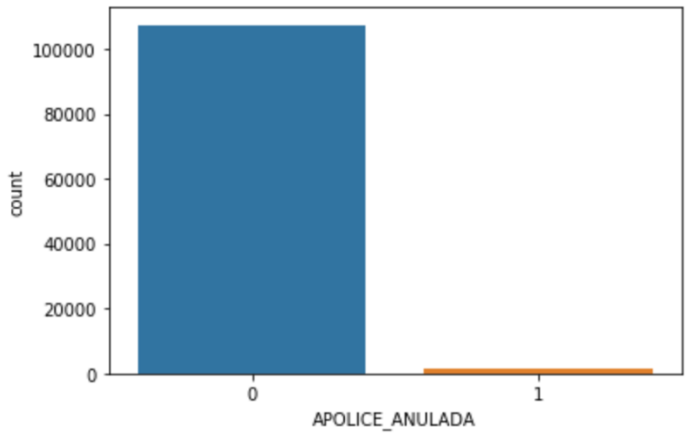



**Figura 26 - Quantidade apólices vigentes e apólices anuladas**

De seguida, foi analisado a variável Sexo do Tomador, e verificou-se que ambos os sexos possuíam taxas de anulação semelhantes como é possível visualizar na Figura 27.

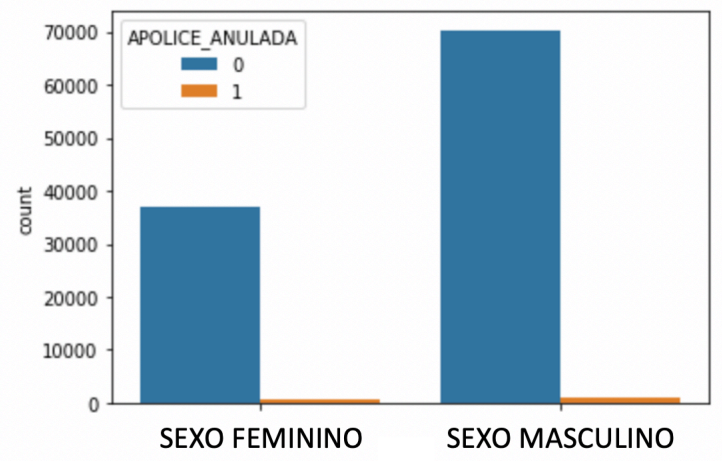



**Figura 27 - Distribuição de apólices vigentes e apólices anuladas distinguidas por sexo do tomador**

Na Figura 28, conseguimos verificar que a maioria das apólices anuladas apenas tinham essa apólice contratada na companhia. Verificamos ainda que há uma proporcionalidade direta entre o número de apólices contratadas e a anulação de apólices.

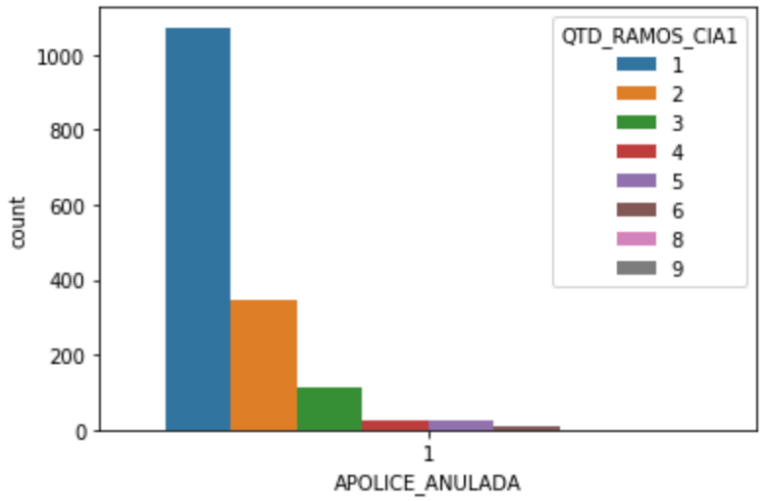



**Figura 28 - Quantidades de apólices nos diversos ramos contratadas pelo tomador**

No que diz respeito às modalidades contratadas para cada apólice verificou-se que as apólices com modalidades com menos coberturas contratadas (Light e Terceiros) são as apólices que contêm um maior número de anulações. É ainda possível verificar pela Figura 29 que as quantidades de apólices anuladas em balcões e agentes (modalidades Light, Plus e Prestige) são superiores às apólices anuladas através dos canais digitais (modalidades Terceiros, Terceiros Mais e Danos Próprios).

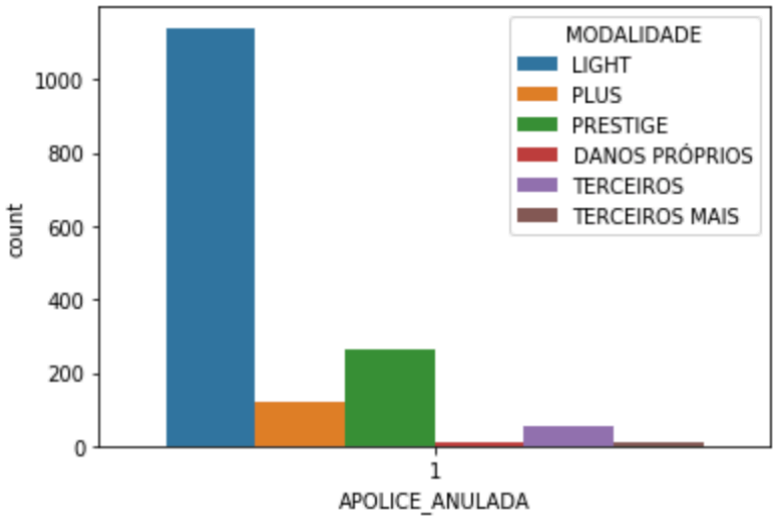



**Figura 29 - Distribuição das apólices anuladas por modalidade selecionada**

Analisando a forma de pagamento por parte dos clientes que anularam as suas apólices podemos verificar pela Figura 30 que quanto mais cotas (recibos) o cliente paga, menor a sua pretensão para anular a sua apólice.

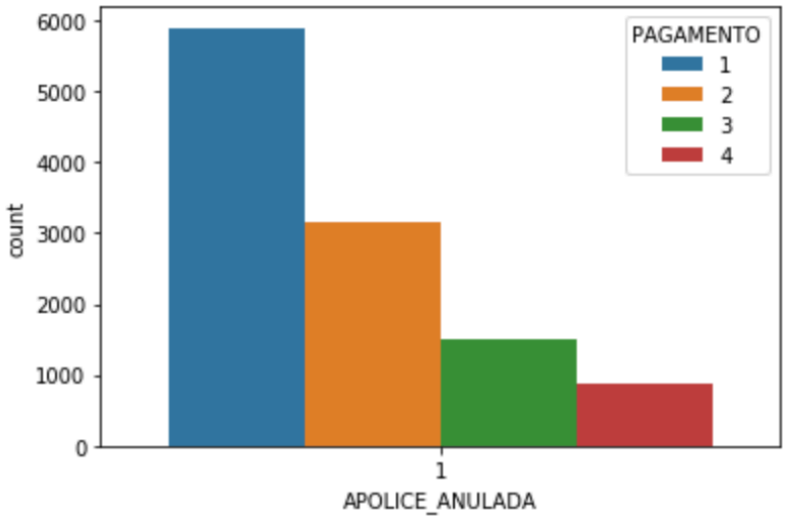



**Figura 30 - Distribuição das apólices anuladas por fracionamento selecionado**

Numa segunda iteração, tendo em conta os valores obtidos na Figura 26, optou-se por alterar o conjunto de dados a analisar, passando a focar num conjunto de apólices emitidas em 2015 e 2016 que no final do ano de 2020 apresentavam a seguinte distribuição de estados: 80% vigente e 20% já anuladas – como é possível observar pela Figura 31. Com esta distribuição, é previsível que se irá obter melhores resultados através dos algoritmos.

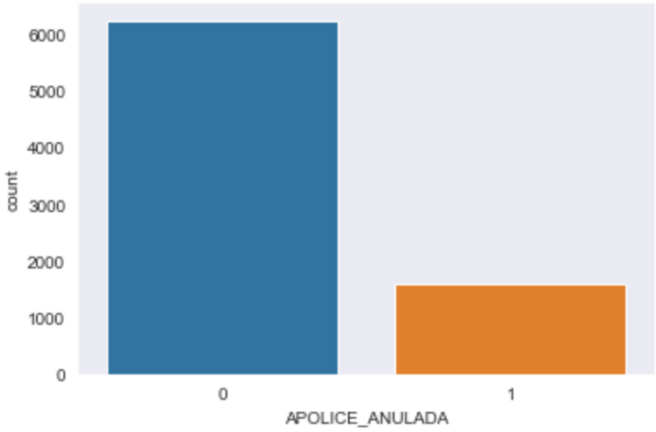



**Figura 31 - Distribuição das apólices anuladas por fracionamento selecionado**

## *4.3* – Preparação dos Dados

Tendo efeito as primeiras análises gráficas e estatísticas aos dados reunidos, efetuaram-se diversas operações de normalização de variáveis aos dados gerados, assim como a remoção de registos com valores considerados *outliers*.

Efetuou-se ainda várias operações de *one hot encoding* a diversas variáveis de modo a transformar as mesmas em colunas binárias (1 para sim, 0 para não). As operações de on*e hot*

*encoding* consistem em transformar uma variável categórica em várias colunas *dummy*, sendo que cada coluna *dummy* irá conter 1 ou 0.

Um exemplo desta transformação, foi a variável TipoCombustivel, sendo que após a operação de *one hot encoding* deu origem às variáveis Combustivel_Gasolina, Combustivel_Gasoleo, Combustivel_Eletrico, todas estas com valores 1 ou 0. Assim como foi efetuado à variável TipoCombustível, foi ainda efetuado a outras variáveis como a ModalidadeApolice, à MarcaVeiculo, ao TipoAssistencia, ao Fraccionamento e ao TipoCobrança e à CategoriaVeiculo.

Efetuou-se ainda o agrupamento em classes de diversas variáveis, por exemplo, a variável DataCartaCondução passou para DecadaCartaCondução, sendo que o valor 02/10/2014 passou para 2010, deste modo reduziu-se bastante o leque possível de valores. Esta operação foi igualmente efetuada para a variável do DataNascimento e AnoVeiculo.

Numa segunda iteração, retirou-se algumas variáveis criadas através da técnica one hot encoding, uma vez que se concluiu que as mesmas não contribuíam para o modelo e apenas acrescentavam ruído ao mesmo. Um exemplo foi a variável MarcaVeículo que gerou mais de 30 colunas e que consistia em demasiado ruído no conjunto de dados. Retiraram-se igualmente do conjunto de dados todas as variáveis que não continham correlação com a variável objetivo. Após estas remoções, efetuou-se uma análise às variáveis que estavam mais correlacionadas com a variável objetivo, sendo as mesmas apresentadas na Figura 32.

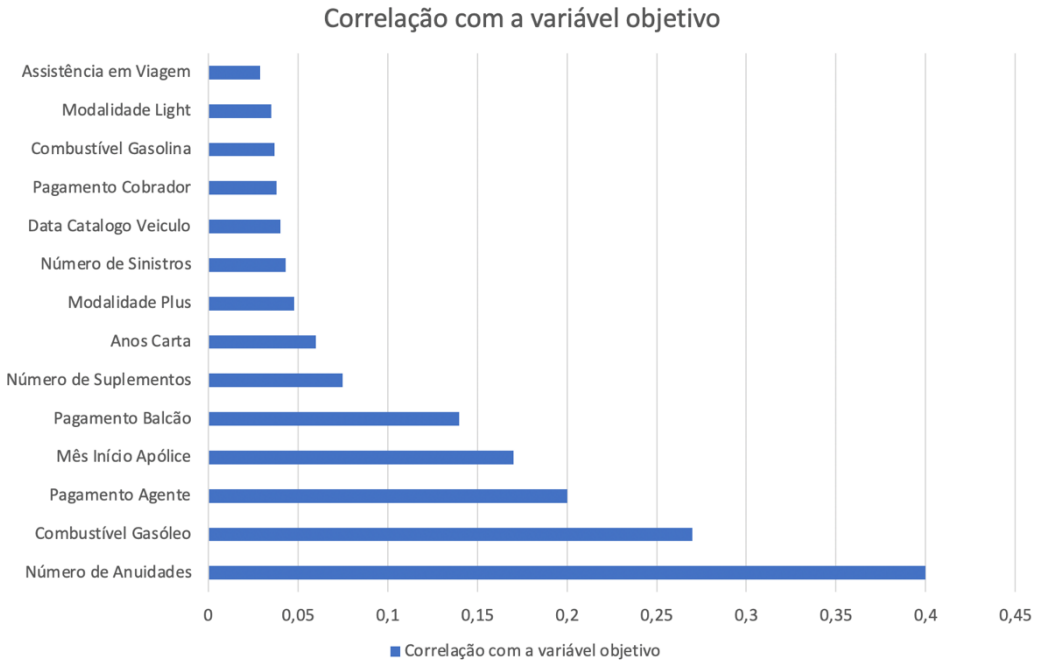



**Figura 32 – Correlação com a variável objetivo**

## *4.4* – Modelação

Antes de submeter o conjunto de dados aos algoritmos foi decidido fazer um balanceamento da variável objetivo. Assim como já foi mencionado na Figura 26, existia um grande desnível no que se refere à varíavel objetivo podendo ser um factor que influenciasse negativamente os resultados que fossemos obter. Assim, avançou-se com diversas técnicas de *oversampling* e de *undersampling*, até se obter um universo com 60% de apólices vigentes e 40% de apólices anuladas.

Na segunda iteração, uma vez que estávamos a lidar com uma variável objetivo mais equilibrada, com uma proporção de 80% / 20% decidiu-se não usar técnicas de *oversampling* e *undersampling* de modo a estas técnicas não virem influenciar as taxas de acerto.

## *4.5* – Avaliação

Na etapa de avaliação dos algoritmos procedeu-se à divisão aleatória do conjunto de dados original, numa proporção de 70% para o conjunto de treino e 30% para o conjunto de testes.

Com base nos artigos analisados na secção 2.4, foram selecionados os algoritmos Regressão Logística, KNN, SVM, Naive Bayes, Decision Tree e Random Forest.

Contudo, após aplicar os diversos algoritmos verificou-se que dado a diferença tão elevada entre o número de apólices anuladas e apólices vigentes que mesmo aplicando as técnicas que servem para equilibrar as duas classes, os valores de algoritmos continuavam a ser bastante desvirtuados – obtendo valores na ordem dos 99% de acerto como é possível observar na Figura 33, uma vez que indicava que todo o universo de apólices seria renovado.

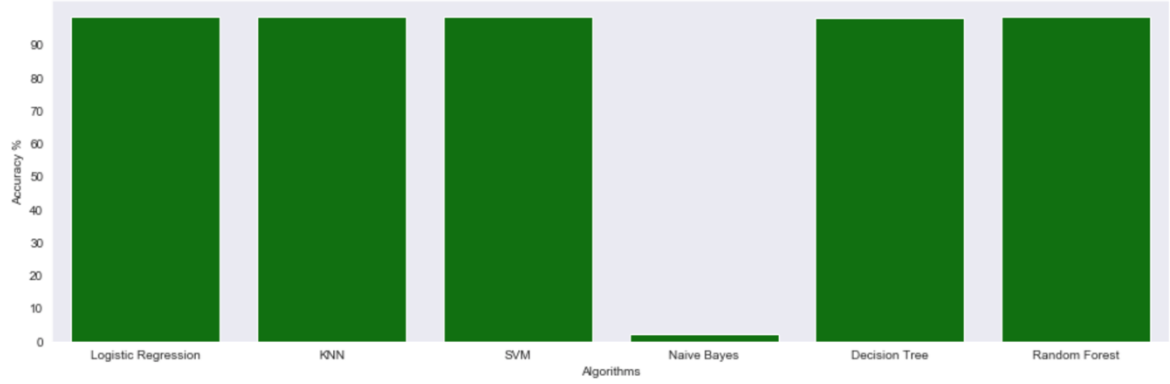



**Figura 33– Taxas de acerto obtidas pelos algoritmos na 1ª iteração**

Após analisar verificou-se que ao efetuar *oversampling* aos dados, toda a classe de apólices anuladas iria-se ser constantemente clonada, permitindo aos algoritmos detetar esses registos como apólices que iriam ser anuladas.

Na segunda iteração, voltou-se a dividir o conjunto de dados, num conjunto de treino e num conjunto de teste, separando numa proporção de 70%/30%.

Após submeter este novo conjunto aos algoritmos obtiveram-se resultados ligeiramente abaixo, comparativamente à primeira iteração, estando nesta iteração com taxas de acerto entre 94% e 98% - como é possível visualizar na Figura 34.

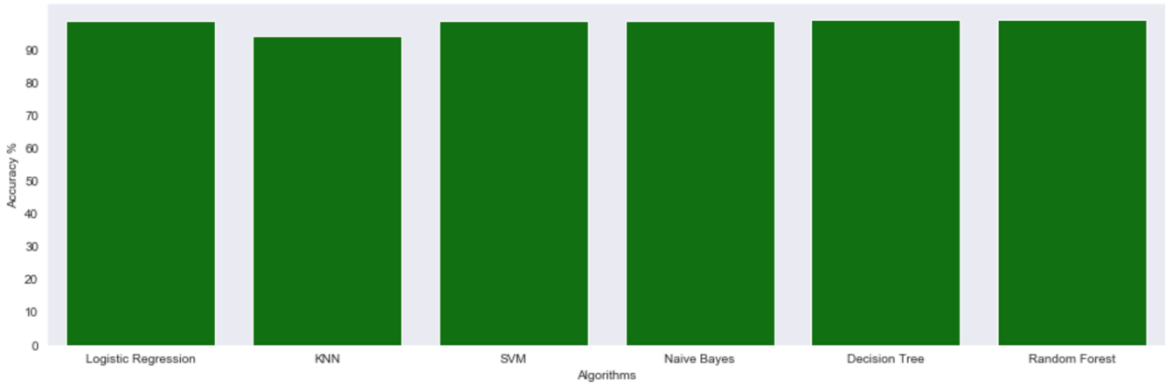



**Figura 34– Taxas de acerto obtidas pelos algoritmos na 2ª iteração**

# Capítulo 5 – Demonstração e Avaliação

O capítulo 5 descreve todas as etapas de demonstração e de avaliação dos modelos criados, assim como de todos os comentários efetuados pelos avaliadores. Os avaliadores selecionados para esta demonstração estão relacionados com a área que estuda e analisa estes dados em contexto real.

A avaliadora 1, Soraia Bornett, coordena departamentos da área de *market pricing*, enquanto o avaliador 2, Luís Silva, pertence à área de tecnologias. Desta forma teremos avaliadores de áreas distintas, o que irá beneficiar na avaliação destes artefactos.

As iterações relativas ao *Market Basket Analysis* foram efetuadas nos meses de maio e junho. As iterações relativas ao *Customer Churn* foram efetuadas nos meses de julho e agosto. Em cada uma das iterações foi efetuado uma demonstração, servido um rápido questionário e recolhido sugestões para melhorias.

## *5.1* – Análise de afinidade de produtos

### 5.1.1 – Demonstração

Para os efeitos de demonstração, efetuou-se a listagem de todas as regras obtidas junto do conjunto de dados retirado entre 2018 e 2019.

Para cada regra encontrada pelo modelo, efetuou-se uma validação com os dados de novas apólices em 2020. Com esta demonstração foi possível comprovar a fiabilidade do modelo construído.

### 5.1.2 – Primeira Iteração DSRM

Nesta primeira iteração, ao demonstrar as regras que foram obtidas verificou-se que a maioria das regras apresentavam baixos valores de suporte e de confiança.

Após análise das regras obtidas, os avaliadores responderam a um curto questionário, que é apresentado na Tabela 2.

| Critério | Objetivo | Avaliador 1<br>Soraia Bornett | Avaliador 2<br>Luís Silva |
|---|---|---|---|
| Eficácia | Aprender com o passado para conseguir reter clientes no futuro | C | B |
| Consistência com as pessoas: compreensibilidade | Providenciar informação coerente | B | B |
| Consistência com as pessoas: fácil de usar | Providenciar informação que seja fácil de usar | C | C |
| Consistência com a organização: utilitário | Providenciar informação que permita à companhia reter clientes | B | B |
| Nível de detalhe | Oferecer à companhia regras detalhadas sobre os diferentes tipos de clientes | B | B |
| Performance | Permitir ignorar regras com baixos níveis de confiança (ruído) | C | C |
| Capacidade de aprendizagem | Permitir que o modelo aprenda automaticamente ao longo do tempo | A | A |

**Tabela 2 – Resultados da 1ª iteração DSRM do MBA**

Foi sugerido pelos avaliadores efetuar uma normalização no que diz respeito aos ramos. Muitos ramos na companhia encontravam-se desagregados, como é possível verificar na Figura 15.

Foi ainda sugerido efetuar as visualizações em formato gráfico para cada um dos clusters que foram obtidos, assim iria ser fornecido uma visualização mais rápida e mais fácil de usar (identificando os ramos antecedentes, através das setas, era possível visualizar os ramos consequentes).

### 5.1.3 – Segunda Iteração DSRM

Nesta segunda iteração, após efetuar a normalização sugerida, obtiveram-se melhores valores de suporte e de confiança, obtendo assim melhores avaliações por parte dos especialistas. Efetuou-se uma tabela com as regras resultantes, presente no Anexo A. Foi igualmente produzido visualizações das regras em formato gráfico, como está demonstrado na Figura 19.

| **Critério** | **Objetivo** | **Avaliador 1 Soraia Bornett** | **Avaliador 2 Luís Silva** |
|---|---|---|---|
| Eficácia | Aprender com o passado para conseguir reter clientes no futuro | B | B |
| Consistência com as pessoas: compreensibilidade | Providenciar informação coerente | B | B |
| Consistência com as pessoas: fácil de usar | Providenciar informação que seja fácil de usar | A | B |
| Consistência com a organização: utilitário | Providenciar informação que permita à companhia reter clientes | B | B |
| Nível de detalhe | Oferecer à companhia regras detalhadas sobre os diferentes tipos de clientes | B | B |
| Performance | Permitir ignorar regras com baixos níveis de confiança (ruído) | A | A |
| Capacidade de aprendizagem | Permitir que o modelo aprenda automaticamente ao longo do tempo | A | A |

**Tabela 3 – Resultados da 2ª iteração DSRM do MBA**

## *5.2 – Customer Churn*

### 5.2.1 – Demonstração

Para os efeitos de demonstração, do segundo artefacto construído, submeteu-se o modelo construído ao conjunto de testes, tendo sido analisado e debatido as análises gráficas sobre as variáveis, as variáveis mais significativas em relação à variável objetivo e as taxas de acerto obtido pelos diversos algoritmos usados junto dos avaliadores.

### 5.2.2 – Primeira Iteração DSRM

Nesta primeira demonstração foram usados como conjunto de dados as apólices vigentes no ano de 2019 e as apólices anuladas em 2019. As avaliações menos positivas desta primeira iteração foram derivadas dos valores obtidos pelos modelos. Os modelos apresentados apresentavam taxas de acerto de 99%, uma vez que todos os modelos davam, para todos

Após análise e debate sobre os resultados obtidos, os avaliadores responderam a um curto questionário, que é apresentado na Tabela 4.

| Critério | Objetivo | Avaliador 1<br>Soraia Bornett | Avaliador 2<br>Luís Silva |
|---|---|---|---|
| Eficácia | Aprender com o passado para conseguir reter clientes no futuro | D | D |
| Consistência com as pessoas: compreensibilidade | Providenciar informação coerente | D | C |
| Consistência com as pessoas: fácil de usar | Providenciar informação que seja fácil de usar | C | C |
| Consistência com a organização: utilitário | Providenciar informação que permita à companhia reter clientes | C | C |
| Nível de detalhe | Oferecer à companhia regras detalhadas sobre os diferentes tipos de clientes | C | D |
| Performance | Permitir ignorar regras com baixos níveis de confiança (ruído) | D | D |
| Capacidade de aprendizagem | Permitir que o modelo aprenda automaticamente ao longo do tempo | C | C |

**Tabela 4 – Resultados da 1ª iteração DSRM do Customer Churn**

Foi sugerido pelos avaliadores reduzir o número de variáveis presentes no conjunto de dados uma vez que através das operações de *one hot encoding* foi gerada demasiadas variáveis e que as mesmas iriam provocar demasiada desagregação de valores, como no caso da Marca do Veículo, onde cada marca resultou numa coluna adicional.

Outro comentário seria alterar o filtro da busca da informação passando a analisar as apólices geradas de um determinado ano (permitindo obter um melhor equilíbrio no que diz respeito à variável objetivo).

Outra sugestão dos avaliadores seria tentar obter a lista das variáveis mais relacionadas com a variável objetivo de modo que se conseguisse obter a lista de variáveis a que seja necessário estar com atenção no futuro.

### 5.2.3 – Segunda Iteração DSRM

Nesta segunda avaliação, já com um conjunto de dados em que a variável objetivo se encontra mais bem equilibrada, obteve-se avaliações melhores.

Para tal, contribui o facto de os modelos terem obtido taxas de acerto mais realistas e ser possível apresentar quais as variáveis que estão mais correlacionadas com a variável objetivo (Figura 32), como é possível observar pela Tabela 5.

| **Critério** | **Objetivo** | **Avaliador 1 Soraia Bornett** | **Avaliador 2 Luís Silva** |
|---|---|---|---|
| Eficácia | Aprender com o passado para conseguir reter clientes no futuro | C | C |
| Consistência com as pessoas: compreensibilidade | Providenciar informação coerente | C | C |
| Consistência com as pessoas: fácil de usar | Providenciar informação que seja fácil de usar | C | C |
| Consistência com a organização: utilitário | Providenciar informação que permita à companhia reter clientes | B | B |
| Nível de detalhe | Oferecer à companhia regras detalhadas sobre os diferentes tipos de clientes | B | B |
| Performance | Permitir ignorar regras com baixos níveis de confiança (ruído) | C | C |
| Capacidade de aprendizagem | Permitir que o modelo aprenda automaticamente ao longo do tempo | B | C |

**Tabela 5 – Resultados da 2ª iteração DSRM do Customer Churn**

# Capítulo 6 – Conclusões e Trabalho Futuro

## *6.1 – Conclusões*

O objetivo principal desta dissertação foi relacionado com a retenção dos clientes. Para tal foram desenvolvidas duas abordagens, que embora diferentes, contribuíam para o mesmo fim. Sendo uma abordagem mais preventiva – Análise de afinidade dos produtos, e uma abordagem mais reativa – Modelo de *Customer Churn*.
Tendo estas abordagens em mente, surgiram assim as duas questões-chaves que esta dissertação decidiu responder:

1- É possível detetar padrões de contratação de apólices, com base na caminhada do cliente na companhia?

2- É possível detetar clientes que estejam em vias de abandonar a companhia?

No que diz respeito à primeira questão-chave, é possível afirmar que é possível detetar os padrões dos clientes. Tendo como base o seu histórico de contratações e da sua zona é possível identificar quais os produtos que um determinado cliente poderá estar apto a adquirir. Para tal, poderá efetuar-se campanhas de cross-selling junto de determinados clientes por forma a fidelizar os mesmos.
No que diz respeito à segunda questão-chave, foi possível detetar quais as variáveis que contêm mais influência no abandono dos clientes para o ramo automóvel – o ramo analisado neste artefacto. Deste modo a área responsável irá conseguir focar-se nestas variáveis para possíveis estudos e campanhas com o objetivo de reter os clientes do ramo automóvel.
Para a companhia seguradora, foram disponibilizadas duas novas ferramentas com a finalidade de reter os clientes. Uma das ferramentas permite detetar afinidades de produtos para os seus clientes, enquanto a segunda ferramenta deteta as principais variáveis que indicam um possível abandono do cliente à companhia. Ambas irão contribuir para as ações de retenção dos clientes, tornando-se assim como duas ferramentas bastantes úteis num mercado cada vez mais competitivo.
As limitações do trabalho desenvolvido foram a necessidade de entender a estrutura do negócio, a extração dos dados, e a disponibilidade de tempo.

## *6.2 – Trabalho Futuro*

A nível de Trabalho Futuro, na última iteração foram mencionadas algumas melhorias que são possíveis de serem implementadas aos modelos já desenvolvidos e que poderão vir a ajudar a melhorar os resultados.

No que respeita ao modelo de *Market Basket Analysis*, foi sugerido passar a incluir igualmente uma avaliação por setores (agregação de ramos por diferentes áreas – sector de diversos, setor de acidentes, setor automóvel, etc).

No que respeita ao modelo de *Customer Churn*, foi sugerido incluir os dados oriundos do *Call Center*, que atualmente são armazenados em texto livre. Por forma a adicionar estas variáveis no dataset, seria também necessário efetuar alguns processos de tratamento de *text mining* sobre os textos efetuados pelos operadores da área de *Call Center*. Com a inclusão dos dados de *Call Center* prevê-se que os modelos de *Customer Churn* melhorem a sua taxa de acerto.

# Bibliografia


[1] "A Design Science Research Methodology for Information Systems Research: Journal of Management Information Systems: Vol 24, No 3." https://www.tandfonline.com/doi/abs/10.2753/MIS0742-1222240302 (accessed Aug. 16, 2021).

[2] Autoridade de Supervisão de Seguros e Fundos de Pensões, "Produção de Seguro Direto – 2019." Jan. 2020. [Online]. Available: https://www.asf.com.pt/ISP/Estatisticas/seguros/estatisticas_anuais/historico/PremiosSeguroDireto2019.pdf

[3] N. Prat, I. Comyn-Wattiau, and J. Akoka, "Artifact Evaluation in Information Systems Design-Science Research - a Holistic View," 2014.

[4] Manuel Guedes-Vieira, *Introdução aos seguros*. Vida Económica - Editorial, SA, 2012.

[5] "Motor liability insurance | Calculate price and bonuses." https://www.op.fi/private-customers/insurance/vehicle-insurance/motor-liability-insurance (accessed May 09, 2021).

[6] I. H. de Larramendi and J. Castelo, "Manual Basico de Seguros." Editorial MAPFRE, SA. [Online]. Available: http://www.larramendi.es/es/catalogo_imagenes/grupo.do?path=1024102

[7] N. V. Haueter, "A History of Insurance." Swiss Re Corporate History, 2017. [Online]. Available: https://www.swissre.com/dam/jcr:638f00a0-71b9-4d8e-a960-dddaf9ba57cb/150_history_of_insurance.pdf

[8] "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement | The BMJ." https://www.bmj.com/content/339/bmj.b2535 (accessed Aug. 24, 2021).

[9] "VOSviewer - Visualizing scientific landscapes." https://www.vosviewer.com/ (accessed Aug. 03, 2021).

[10] R. Agrawal, T. Imieliński, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," *SIGMOD Rec.*, vol. 22, no. 2, pp. 207–216, Jun. 1993, doi: 10.1145/170036.170072.

[11] I. Cil, "Consumption universes based supermarket layout through association rule mining and multidimensional scaling," *Expert Systems with Applications*, vol. 39, no. 10, pp. 8611–8625, 2012, doi: https://doi.org/10.1016/j.eswa.2012.01.192.

[12] Y. Liu, "Study on Application of Apriori Algorithm in Data Mining," in *2010 Second International Conference on Computer Modeling and Simulation*, 2010, vol. 3, pp. 111–114. doi: 10.1109/ICCMS.2010.398.

[13] A. Trnka, "Market Basket Analysis with Data Mining methods," in *2010 International Conference on Networking and Information Technology*, Jun. 2010, pp. 446–450. doi: 10.1109/ICNIT.2010.5508476.

[14] J. Wertz, "Don't Spend 5 Times More Attracting New Customers, Nurture The Existing Ones", [Online]. Available: https://www.forbes.com/sites/jiawertz/2018/09/12/dont-spend-5-times-more-attracting-new-customers-nurture-the-existing-ones/?sh=1822251f5a8e

[15] Y. He, Y. Xiong, and Y. Tsai, "Machine Learning Based Approaches to Predict Customer Churn for an Insurance Company," in *2020 Systems and Information Engineering Design Symposium (SIEDS)*, Apr. 2020, pp. 1–6. doi: 10.1109/SIEDS49339.2020.9106691.

[16] M. Spiteri and G. Azzopardi, “Customer Churn Prediction for a Motor Insurance Company,” in *2018 Thirteenth International Conference on Digital Information Management (ICDIM)*, Sep. 2018, pp. 173–178. doi: 10.1109/ICDIM.2018.8847066.
[17] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, “A Churn Prediction Model Using Random Forest: Analysis of Machine Learning Techniques for Churn Prediction and Factor Identification in Telecom Sector,” *IEEE Access*, vol. 7, pp. 60134–60149, 2019, doi: 10.1109/ACCESS.2019.2914999.
[18] A. Keramati, H. Ghaneei, and S. M. Mirmohammadi, “Developing a prediction model for customer churn from electronic banking services using data mining,” *Financial Innovation*, vol. 2, no. 1, 2016, doi: 10.1186/s40854-016-0029-6.
[19] N. Gordini and V. Veglio, “Customers churn prediction and marketing retention strategies. An application of support vector machines based on the AUC parameter-selection technique in B2B e-commerce industry,” *Ind. Mark. Manage.*, vol. 62, pp. 100–107, 2017, doi: 10.1016/j.indmarman.2016.08.003.
[20] M. Milošević, N. Živić, and I. Andjelković, “Early churn prediction with personalized targeting in mobile social games,” *Expert Sys Appl*, vol. 83, pp. 326–332, 2017, doi: 10.1016/j.eswa.2017.04.056.
[21] A. K. Ahmad, A. Jafar, and K. Aljoumaa, “Customer churn prediction in telecom using machine learning in big data platform,” *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0191-6.
[22] Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler), “CRISP-DM 1.0 - Step-by-step data mining guide.” 1999. [Online]. Available: https://the-modeling-agency.com/crisp-dm.pdf

# Anexos

**ANEXO A – Associação entre Produtos obtidos**

| Cluster | Antecedente | Consequente | Confiança % |
|---|---|---|---|
| | Zona Norte | | |
| 0 | [PRODUTO A] + [PRODUTO B] | [PRODUTO E] | 97% |
| 0 | [PRODUTO A] + [PRODUTO C] | [PRODUTO E] | 93% |
| 0 | [PRODUTO A] + [PRODUTO D] | [PRODUTO E] | 87% |
| 1 | [PRODUTO A] + [PRODUTO D] | [PRODUTO E] | 73% |
| 2 | [PRODUTO A] + [PRODUTO D] | [PRODUTO E] | 71% |
| 2 | [PRODUTO A] + [PRODUTO C] | [PRODUTO E] | 71% |
| | Zona Centro | | |
| 0 | [PRODUTO A] + [PRODUTO C] | [PRODUTO E] | 83% |
| 0 | [PRODUTO A] + [PRODUTO D] | [PRODUTO E] | 78% |
| 0 | [PRODUTO F] | [PRODUTO G] | 73% |
| 1 | [PRODUTO F] | [PRODUTO G] | 65% |
| 2 | [PRODUTO A] + [PRODUTO D] | [PRODUTO E] | 100% |
| 2 | [PRODUTO A] + [PRODUTO D] + [PRODUTO C] | [PRODUTO E] | 100% |
| 2 | [PRODUTO A] + [PRODUTO B] | [PRODUTO E] | 96% |
| 2 | [PRODUTO D] + [PRODUTO C] + [PRODUTO E] | [PRODUTO A] | 94% |
| 2 | [PRODUTO B] | [PRODUTO E] | 93% |
| | Zona Lisboa | | |
| 0 | [PRODUTO A] + [PRODUTO D] | [PRODUTO E] | 80% |
| 0 | [PRODUTO A] + [PRODUTO C] | [PRODUTO E] | 77% |
| 0 | [PRODUTO D] | [PRODUTO E] | 71% |
| 1 | [PRODUTO F] | [PRODUTO G] | 88% |
| 1 | [PRODUTO D] | [PRODUTO G] | 80% |
| 1 | [PRODUTO H] | [PRODUTO G] | 67% |
| 1 | [PRODUTO D] + [PRODUTO C] | [PRODUTO E] | 66% |
| 2 | [PRODUTO D] + [PRODUTO B] | [PRODUTO E] | 93% |
| 2 | [PRODUTO A] + [PRODUTO D] + [PRODUTO C] | [PRODUTO E] | 90% |
| 2 | [PRODUTO A] + [PRODUTO B] | [PRODUTO E] | 86% |
| 2 | [PRODUTO D] + [PRODUTO C] | [PRODUTO E] | 84% |
| 2 | [PRODUTO A] + [PRODUTO G] | [PRODUTO E] | 81% |