Universidade do Minho
Escola de Ciências

Airline Passenger Satisfaction

Licenciatura em Estatística Aplicada – Data Mining para Ciência de Dados

2021/2022

Discentes: António Martinho, A93719
Miguel Martins, A90261
Paulo Barros, A92929
Rodrigo Brito, A92320
Sérgio Vieira, A92313

Docente: Paulo Cortez

Índice

1. Introdução

Desde os primórdios da civilização, que o ser humano se fascinou com a ideia de poder voar. Eis que, em 17 de dezembro de 1903, os irmãos Wright efetuaram o primeiro voo controlado da história, dando azo a um mundo novo, um mundo de comunicação aérea, promovida pela sua invenção, o avião.

Hoje em dia partem, de todo o mundo, milhares de voos diariamente, que ligam as populações a todos os cantos do mundo.

É neste contexto que apresentamos o trabalho desenvolvido no âmbito da disciplina de Data Mining para a Ciência de Dados, que consiste no estudo e aplicação de conteúdos lecionados sobre uma base de dados relativa à satisfação dos passageiros de uma companhia aérea, denominada *Airline Passenger Satisfaction.*

Esta base de dados foi retirada do site kaggle (https://www.kaggle.com/code/frixinglife/airline-passenger-satisfaction/data) e é composta por 104.000 observações que dizem respeito a 104.000 passageiros que viajaram na companhia e de disponibilizaram a responder a um questionário, que continha perguntas relativas à satisfação do cliente com vários aspetos da experiência do voo.

Este trabalho tem como principal objetivo a consolidação das competências adquiridas nas aulas da disciplina, ao longo do semestre, bem como incentivar a pesquisa autónoma e melhoria no que toca à linguagem de programação R e ao seu uso, não só como mecanismo de criação, mas como mecanismo de ajuda na análise estatística e aplicação de técnicas de data mining, e será realizado com recurso à ferramenta RStudio (https://rstudio.com/).

O trabalho é composto por quatro fases distintas, inspiradas na metodologia CRISP-DM a fase do esclarecimento sobre a divisão de tarefas, temas discutidos enquanto grupo e a autoavaliação, a fase da apresentação, manipulação, limpeza e análise descritiva dos dados, a fase da utilização dos métodos de data mining e os resultados, e a última fase, que diz respeito às conclusões e à análise das limitações.

Com este projeto, podem ser construídos métodos de previsão mais eficientes, de modo a melhorar a experiência dos clientes nos voos desta empresa comercial, levando ao benefício de ambos.

2. Execução do Projeto

2.1. Funcionamento do Grupo

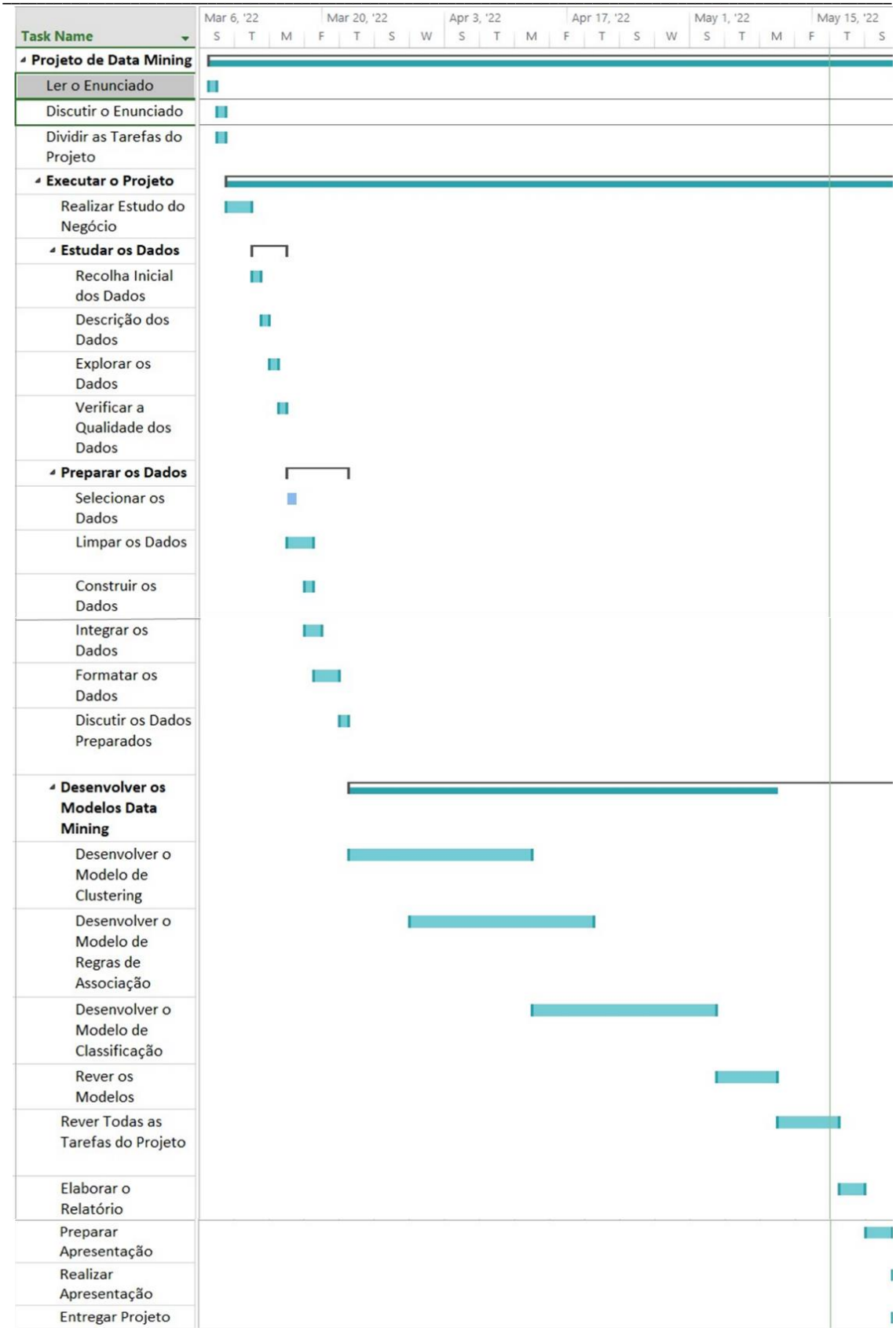
Numa primeira fase, foi elaborado um contrato de trabalho, onde constava a obrigação de comparência de todos os membros do grupo numa reunião à distância, um dia por semana (no mínimo). Nessas reuniões foram definidos os papeis de todos os elementos de grupo, bem como foram discutidos os objetivos do trabalho em questão e a divisão das tarefas.

O passo seguinte foi a escolha da base de dados, onde todos os elementos do grupo apresentaram, após uma pesquisa calculada, diferentes bases de dados e em que o grupo votou, de modo a escolher a mais adequada.

A distribuição de tarefas foi feita tendo em conta os pontos fortes de cada um, não ignorando a necessidade de todos os elementos do grupo estarem familiarizados e serem capazes de executar todas as tarefas propostas. Isto incluiu as aplicações de código em R, a modelação da base de dados, a classificação e a apresentação de resultados e conclusões.

Todo o grupo esteve envolvido na construção e elaboração do relatório.

A progressão do projeto através do semestre pode ser observada através da **Figura.1.** representada através de um diagrama da Gantt.



[Figura.1. Timeline das tarefas do projeto, ilustradas por um diagrama de Gantt**]**

2.2. Divisão do Trabalho

As tarefas deste trabalho foram realizadas por:

- António Martinho, A93719
- Miguel Martins, A90261
- Paulo Barros, A92929
- Rodrigo Brito, A92320
- Sérgio Vieira, A92313

2.2.1. António Martinho

O meu principal foco, neste projeto, tratou-se da pesquisa e análise das regras de associação. A realização desta tarefa requereu alguma pesquisa através do material complementar disponibilizado pelo docente, de apontamentos feitos nas aulas e de recursos disponibilizados na internet. Deste modo, este primeiro contacto com esta nova forma de análise tornou-se mais familiar.

Além das regras de associação, tive uma participação ativa na elaboração de gráficos destinada à análise descritiva das variáveis. Esta tarefa foi desempenhada com mais segurança, uma vez que me encontro mais familiarizado com a elaboração de gráficos no R utilizando a biblioteca *ggplot2*.

Adicionalmente, uma participação, ainda que mais reservada, na classificação e na elaboração do relatório.

2.2.2. Miguel Martins

Neste projeto, coloquei especial ênfase no desenvolvimento do modelo de Clustering, que exigiu imensa pesquisa, mesmo com a minha familiarização pré-existente do R. Para esta tarefa, foram essenciais os apontamentos disponibilizados pelo docente na Blackboard, pois serviram de exemplo, de modo a poder tirar inspiração e aprender com o progresso do trabalho.

Para lá do Clustering tive, também, participação na elaboração de gráficos, o que foi uma parte mais simples, devido à minha experiência na criação de representações visuais utilizando o *ggplot2.*

Tive, também, participação na construção do relatório.

2.2.3. Paulo Barros

A minha participação neste projeto foi centrada nos modelos de classificação, algo que era novo para mim, visto que era, de todo o grupo, aquele com menos experiência no R. Alcancei os objetivos traçados para este projeto, com a ajuda dos apontamentos disponibilizados pelo docente, através da ajuda dos meus colegas de grupo, principalmente na escrita de código em R, e através da pesquisa que realizei, através de fóruns e de livros sobre o assunto.

Acrescento, ainda, que tive parte na construção de gráficos para a análise descritiva, o que foi uma experiência nova, pois nunca tinha trabalhado com a biblioteca.

Além do resto, fiz parte da elaboração do relatório.

2.2.4. Rodrigo Brito

A minha parte neste projeto foi maioritariamente dedicada à análise descritiva e à elaboração do relatório. Estas tarefas ajustaram-se bem à minha capacidade de interpretação e de escrita sendo, claramente, o meu ponto forte, onde usei toda a experiência e conhecimentos adquiridos ao longo dos últimos anos na componente da estatística.

Em adição, participei ativamente na análise das regras de associação e na sua elaboração, além da elaboração de gráficos. O primeiro requereu bastante pesquisa, visto a minha falta de conhecimento sobre o conteúdo, enquanto o segundo foi feito de forma mais natural, devido aos meus conhecimentos mais avançados da ferramenta R.

2.2.5. Sérgio Vieira

Neste projeto dei principal destaque à análise descritiva e à elaboração do relatório, tirando partido da experiência na análise estatística adquirida ao longo dos últimos três anos.

Participei na construção de gráficos, utilizando as minhas capacidades de escrever código em R. Além disso, participei, ainda que com menos preponderância, na elaboração das regras de associação, na modelação da base de dados e na preparação da mesma. Ainda tive participação no modelo de classificação.

Para a execução de tudo isto, foram fundamentais os apontamentos disponibilizados pelo docente e a pesquisa feita através de diversos métodos.

2.3. Autoavaliação

De um modo geral, o grupo trabalhou de forma consistente, alcançando sempre os objetivos definidos na reunião semanal anterior. O grupo foi coerente e honesto no que foi a artilha de informação e a ajuda mútua em todos os aspetos da execução do projeto.

Tendo em consideração o esforço, a resultado final e a realização dos objetivos, o grupo, em conjunto considera que a avaliação indicada é de 17 valores.

Analisando a secção anterior, onde constam as contribuições de cada membro do grupo, na conceção deste projeto, o grupo considera que todos os seus elementos justificam a atribuição da mesma nota, coincidindo com a nota geral indicada anteriormente.

3. Estudo CRISP-DM

A base de dados utilizada neste projeto é denominada "Airline Passenger Satisfaction" e foi retirada do repositório do *kaggle* (https://www.kaggle.com/code/frixinglife/airline-passenger-satisfaction/data?select=test.csv).

A sua análise seguiu a metodologia **CRISP-DM** (Cross Industry Standard Process for Data Mining) e, por isso, o projeto foi dividido nos segmentos:

- Estudo do Negócio
- Estudo dos Dados
- Preparação dos Dados
- Modelação
- Avaliação
- Implementação

3.1. Estudo do Negócio

3.1.1. Objetivos do Negócio

As companhias aéreas procuram, desde a sua criação, novas maneiras de melhorar a experiência dos seus clientes. De assentos mais confortáveis até um serviço de check-in online mais simplificado e que poupe tempo ao potencial passageiro, o objetivo final é sempre a satisfação máxima possível dos seus clientes.

Este trabalho tem como principal finalidade determinar, através da análise da implementação de técnicas de data mining a respostas de um questionário (cuja finalidade foi determinar a satisfação dos clientes de uma companhia aérea com diversos fatores) de modo a poder determinar uma melhor estratégia que leve a uma maior satisfação. Outra finalidade deste projeto é a categorização de diferentes tipo de clientes de modo a poder diferenciar a forma de atuar para que a satisfação seja maximizada, sabendo que nunca poderá ser atingida a satisfação total, visto que essa seria um ideal utópico.

Este projeto será bem-sucedido se for possível determinar, com a maior precisão possível, os fatores que mais influenciam a satisfação dos clientes e se conseguir categorizar os clientes de acordo com as suas características mais vincadas.

3.1.2. Avaliação da Situação Atual

Os dados necessários para a execução deste projeto foram extraídos do ficheiro "Airline Passenger Satisfaction.csv". De maneira a desenvolver este projeto, foi usada a ferramenta RStudio (https://www.rstudio.com/). Os dados utilizados são assumidos como factuais e este modelo e respetivos resultados não poderão ser plagiados.

A data de entrega e apresentação de 24 de maio de 2022 deverá ser respeitada, assim como o contrato "…" presente em anexo.

A elaboração deste projeto foi feita através da utilização dos dispositivos tecnológicos pessoais de cada elemento do grupo

3.1.3. Objetivos do Negócio

O modelo resultante deste trabalho deve ser capaz de permitir os fatores mais influentes na satisfação final dos clientes da companhia aérea e, consequentemente, serão utilizados métodos de data mining como árvores de decisão e determinação das regras de associação mais influentes. Para além destes, será utilizado um modelo de clustering *kmeans*, de modo a agrupar os diferentes tipos de clientes, tendo em conta as suas características.

O critério para decidir se o modelo de classificação teve sucesso é medido através da precisão deste, sendo que deve ter resultados superiores ou equivalentes a 50%, sabendo que o principal objetivo é ser o mais preciso possível.

3.1.4. Planeamento do Projeto

O projeto foi desenvolvido através de reuniões semanais fora do horário das aulas, o que foi acordado por todo o grupo como sendo a forma mais eficiente de realizar o trabalho. Nas reuniões era revisto todo o trabalho feito durante a semana transata e no fim das mesmas era definido um plano de trabalho para a semana seguinte.

As reuniões foram marcadas de acordo com a disponibilidade todos os membros do grupo.

3.2. Estudo dos Dados

Os dados presentes na base de dados *Air Passenger Satisfaction* foram extraídos de 103904 questionários feitos por uma companhia aérea. Em cada um desses questionários existem 24 respostas, cada uma representando uma variável presente na **Tabela.1.** A variável alvo, a que será tida como a variável prevista na classificação trata-se da variável *Satisfaction.*

O grau 0 presente nas variáveis com graus de satisfação de 0 a 5 representam os clientes que não responderam ou não tinham opinião sobre o assunto.

[Tabela.1. Descrição das variáveis da base de dados Airline Passenger Satisfaction**]**

Variável	Descrição da Variável	Tipo de Variável
Id	Nº de identificação do cliente	Qualitativa
Gender	Género do cliente 0: Feminino 1: Masculino	Qualitativa
Customer Type	Assiduidade do cliente: 0: Não regular 1: Regular	Qualitativa

Age	Idade atual do cliente	Quantitativa
Type of Travel	Tipo de Viagem: 0: Business 1: Personal	Qualitativa
Class	Classe do lugar: 1: Economy Plus 2: Economy 3: Business	Qualitativa
Flight Distance	Distância do voo (em km)	Quantitativa
Inflight Wi-Fi Service	Nível de satisfação com o serviço de Wi-Fi (0-5)	Qualitativa
Departure/Arrival Time Convinient	Nível de satisfação com o horário de chegada/partida (0-5)	Qualitativa
Ease of Online Booking	Nível de satisfação com o serviço de booking online (0-5)	Qualitativa
Gate Location	Nível de satisfação com a localização do portão de embarque (0-5)	Qualitativa
Food and Drink	Nível de satisfação com o serviço de alimentação (0-5)	Qualitativa
Online Boarding	Nível de satisfação com o check-in online (0-5)	Qualitativa
Seat Comfort	Nível de satisfação com o conforto do lugar (0-5)	Qualitativa
Inflight Entertainment	Nível de satisfação com o entretenimento a bordo (0-5)	Qualitativa
On-Board Service	Nível de satisfação com o serviço no embarque (0-5)	Qualitativa
Leg Room Service	Nível de satisfação com o espaço para as pernas (0-5)	Qualitativa
Baggage Handling	Nível de satisfação com o manuseamento da bagagem (0-5)	Qualitativa
Check-In Service	Nível de satisfação com o check-in (0-5)	Qualitativa
Inflight Service	Nível de satisfação com o serviço durante o voo (0-5)	Qualitativa
Cleanliness	Nível de satisfação com a limpeza do voo (0-5)	Qualitativa
Departure Delay	Atraso do voo na partida (em minutos)	Quantitativa

Arrival Delay	Atraso do voo na chegada (em minutos)	Quantitativa
Satisfaction	Satisfação do cliente: 0: Não satisfeito/Neutro 1: Satisfeito	Qualitativa

3.2.1. Análise Exploratória

A **Figura.2.** evidencia o número de valores em falta de cada variável, verificando-se que apenas a variável *Arrival Delay* apresenta valores em falta.

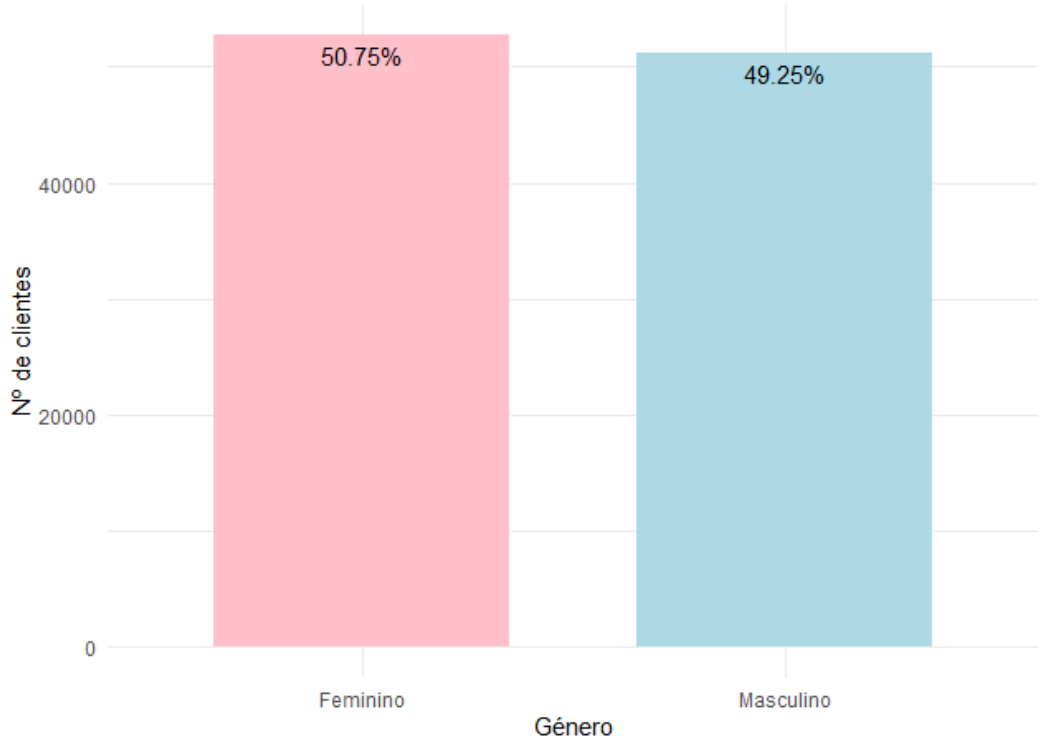
```
> sapply(airline, function(x)sum(is.na(x)))
                Gender          Customer.Type                    Age          Type.of.Travel                  Class        Flight.Distance 
                     0                      0                      0                      0                      0                      0 
                 Wi.Fi     Time.Convinience         Online.Booking          Gate.Location      Food.and.Drink      Online.Boarding 
                     0                      0                      0                      0                      0                      0 
         Seat.Comfort         Entertainment                Service             Leg.Room       Baggage.Handling             Check.In 
                     0                      0                      0                      0                      0                      0 
     Inflight.Service          Cleanliness Departure.Delay.in.Minutes  Arrival.Delay.in.Minutes           Satisfaction 
                     0                      0                      0                    310                      0
```

[Figura.2. Número de valores em falta em cada variável da base de dados *Airline Passenger Satisfaction***]**

```
> summary(airline)
 Gender     Customer.Type      Age        Type.of.Travel Class     Flight.Distance Wi.Fi     Time.Convinience Online.Booking
 0:52727   0:18981       Min.   : 7.00   0:71655        1: 7494   Min.   :  31    0: 3103   0: 5300          0: 4487       
 1:51177   1:84923       1st Qu.:27.00   1:32249        2:46745   1st Qu.: 414    1:17840   1:15498          1:17525       
                         Median :40.00                  3:49665   Median : 843    2:25830   2:17191          2:24021       
                         Mean   :39.38                            Mean   :1189    3:25868   3:17966          3:24449       
                         3rd Qu.:51.00                            3rd Qu.:1743    4:19794   4:25546          4:19571       
                         Max.   :85.00                            Max.   :4983    5:11469   5:22403          5:13851       

 Gate.Location Food.and.Drink Online.Boarding Seat.Comfort Entertainment Service   Leg.Room  Baggage.Handling Check.In
 0:    1       0:  107        0: 2428         0:    1      0:   14       0:    3   0:  472   1: 7237          0:    1   
 1:17562       1:12837        1:10692         1:12075      1:12478       1:11872   1:10353   2:11521          1:12890   
 2:19459       2:21988        2:17505         2:14897      2:17637       2:14681   2:19525   3:20632          2:12893   
 3:28577       3:22300        3:21804         3:18696      3:19139       3:22833   3:20098   4:37383          3:28446   
 4:24426       4:24359        4:30762         4:31765      4:29423       4:30867   4:28789   5:27131          4:29055   
 5:13879       5:22313        5:20713         5:26470      5:25213       5:23648   5:24667                    5:20619   

 Inflight.Service Cleanliness Departure.Delay.in.Minutes Arrival.Delay.in.Minutes Satisfaction
 0:    3          0:   12     Min.   :   0.00            Min.   :   0.00          0:58879     
 1: 7084          1:13318     1st Qu.:   0.00            1st Qu.:   0.00          1:45025     
 2:11457          2:16132     Median :   0.00            Median :   0.00                      
 3:20299          3:24574     Mean   :  14.82            Mean   :  15.18                      
 4:37945          4:27179     3rd Qu.:  12.00            3rd Qu.:  13.00                      
 5:27116          5:22689     Max.   :1592.00            Max.   :1584.00                      
                                                         NA's   :310
```

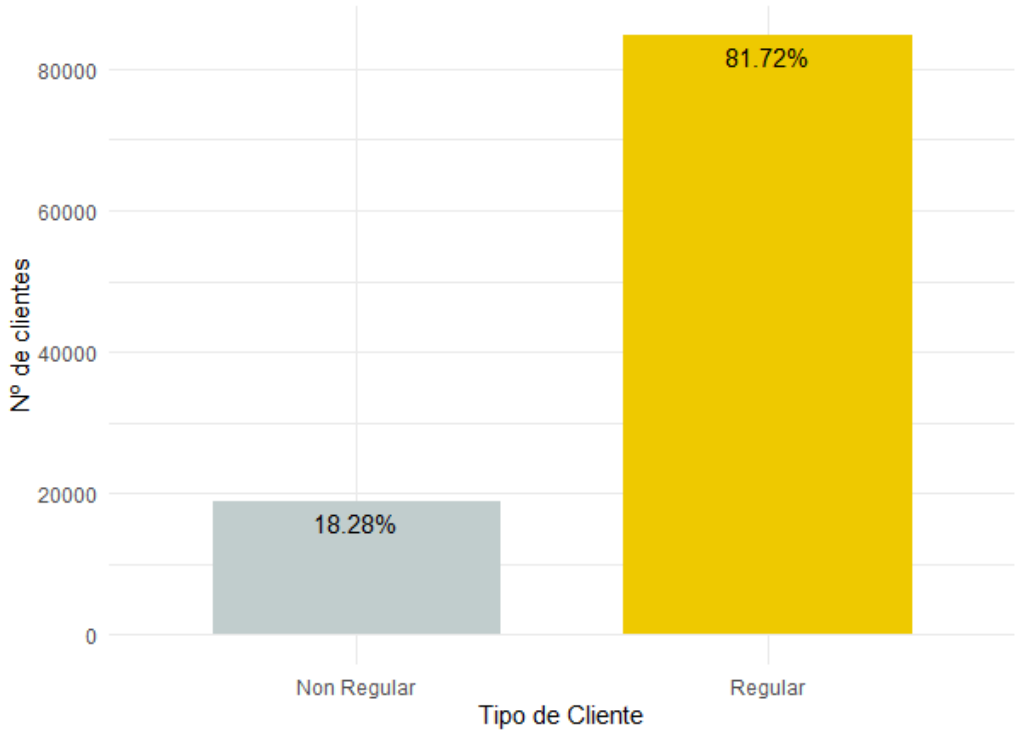
[Figura.3. *Summary* de cada variável da base de dados *Airline Passenger Satisfaction***]**

O *summary,* presente na **Figura.3.**, apresenta, também, os 310 valores em falta na variável *Arrival Delay*, bem como um pequeno resumo dos valores de cada variável.



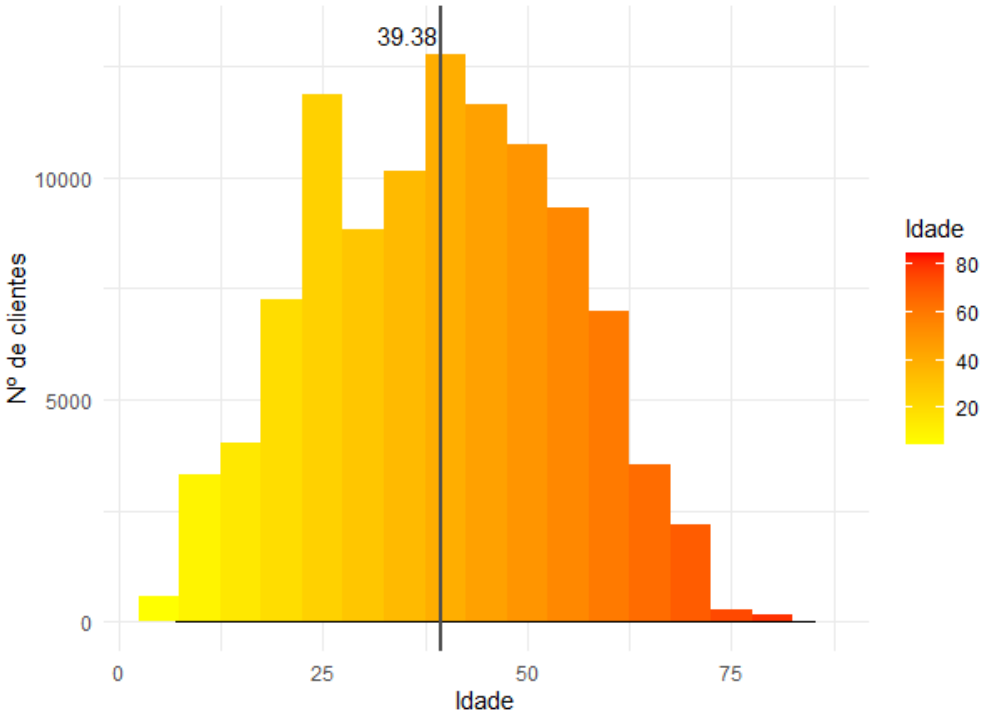
[Figura.4. Gráfico de barras da variável *Gender***]**

Analisando a variável *Gender,* com recurso à **Figura.3.** e **Figura.4.**, observa-se um ligeiro desequilíbrio no que toca ao número de clientes do género feminino (50.75%) e masculino (49.25%), sendo que o primeiro se apresenta em maior número.



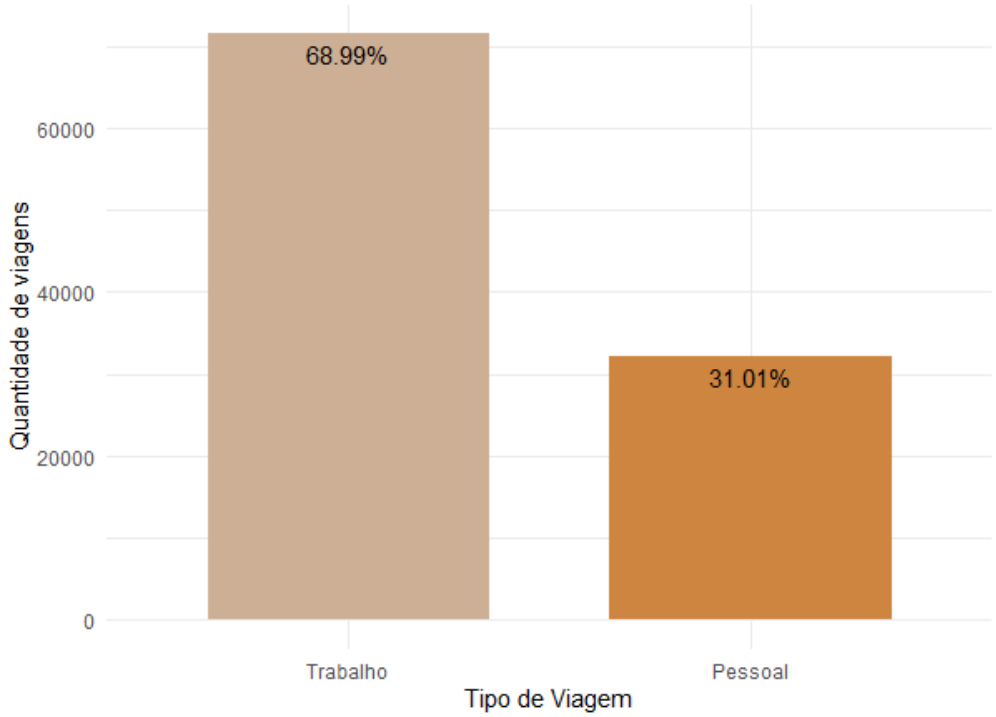
[Figura.5. Gráfico de barras da varável *Customer Type***]**

De acordo com a **Figura.5.**, os clientes desta companhia aérea mostram-se bastante leais a esta, uma vez que 81.72% dos clientes é regular.

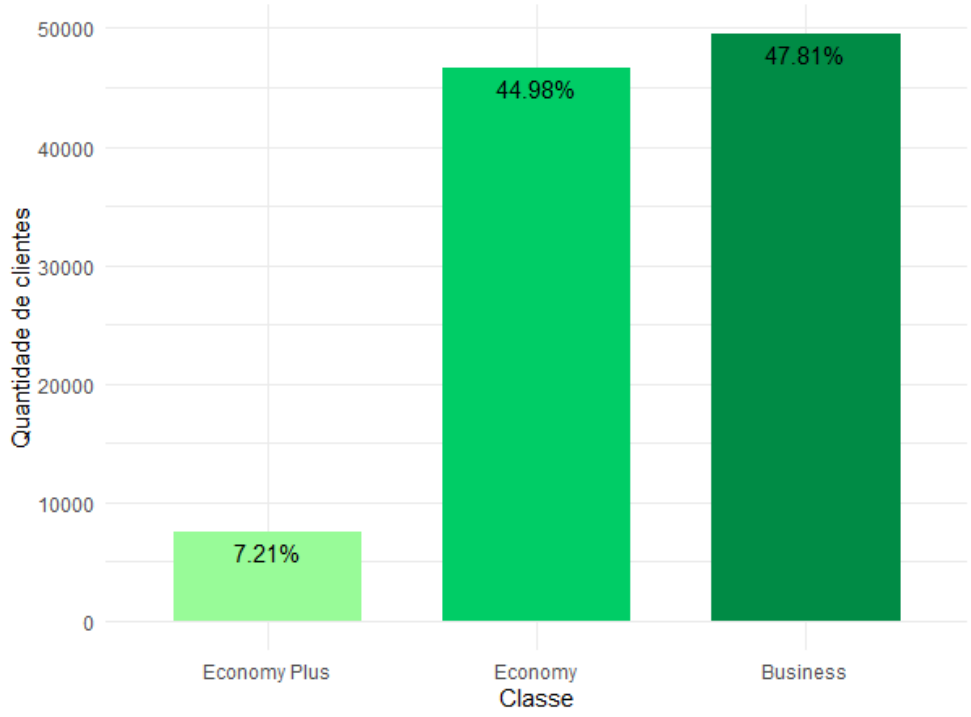


[Figura.6. Histograma da variável *Idade***]**

A idade média representada na **Figura.6.** (aproximadamente 39 anos de idade) e a tendência da figura está de acordo com o panorama de viagens em companhias aéreas, com um maior número de pessoas entre os 25 e 50 anos de idade, ou seja, com um maior número de clientes em idade adulta, o que justifica, também, os resultados da **Figura.7.,** pois este é o intervalo de idades com mais trabalhadores.

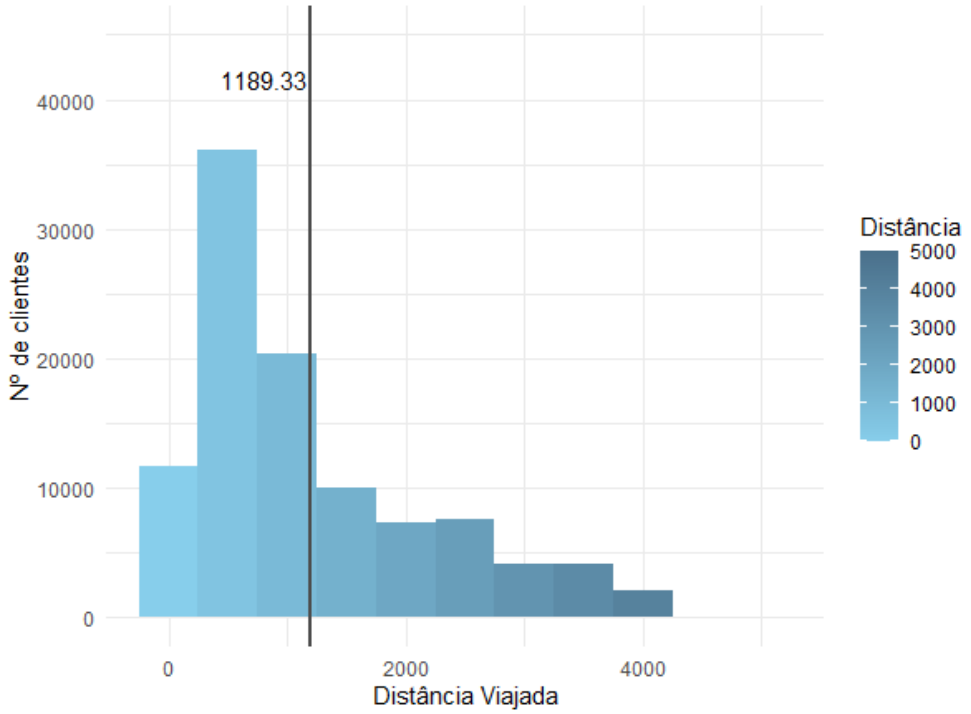


[Figura.7. Gráfico de barras da variável *Type of Travel***]**



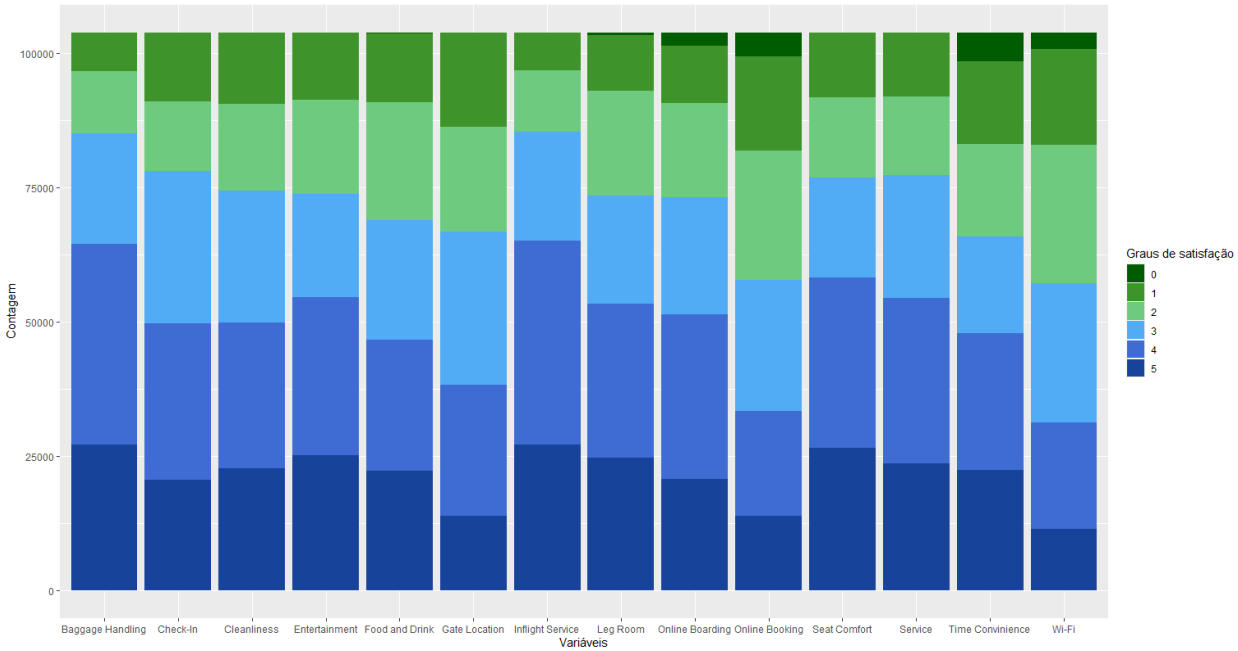
[Figura.8. Gráfico de barras da variável *Class***]**

A **Figura.7.** e a **Figura.8.** caracterizam o tipo de viagem dos clientes, mostrando o tipo de viagem que estes estão a realizar e em que classe. Isto mostra que uma grande parte dos clientes desta companhia aérea viaja em negócios (68.99%), o que pode explicar a maioria dos lugares comprados serem das classes *Business* (47.81%) e *Economy* (44.98%), visto os bilhetes poderem ter sido comprados pelas empresas destes clientes.



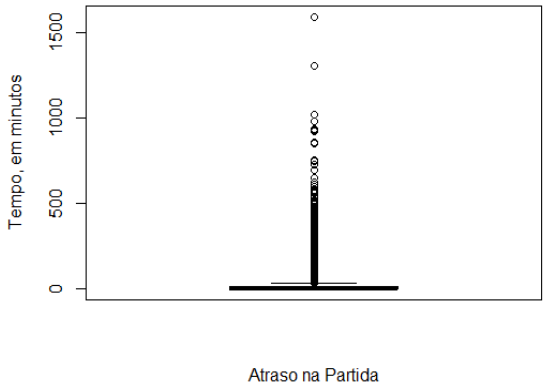
[Figura.9. Histograma da variável *Flight Distance***]**

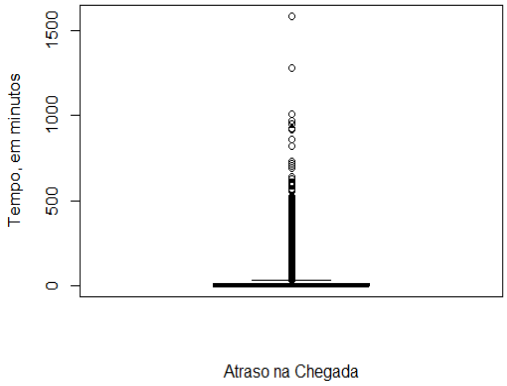
Na **Figura.9.** observa-se uma distância média de viagem de, aproximadamente, 1189.33 km**,** mostrando uma tendência para viagens mais curtas. Isto pode ser explicado pelo maior número de voos regionais e de curta distância, ao contrário dos voos de longa distância que são feitos em números muito inferiores, pois são mais demorados e mais dispendiosos.



[Figura.10. Gráfico de barras empilhado das variáveis com graus de satisfação**]**

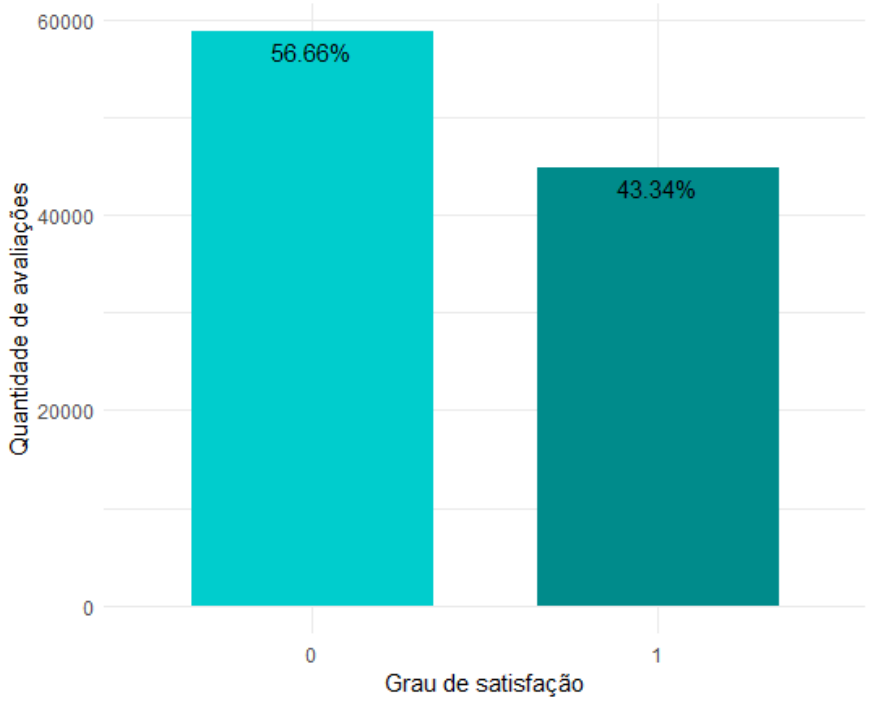
Tendo como referência a **Figura.10.**, podemos contemplar que os clientes se encontram, na sua maioria, bastante satisfeitos com o serviço a bordo. Por outro lado, os clientes encontram-se pouco satisfeitos com o serviço de *Wi-Fi* e com o *Online Booking,* o que pode demonstrar duas zonas do seu serviço onde esta companhia pode melhorar.





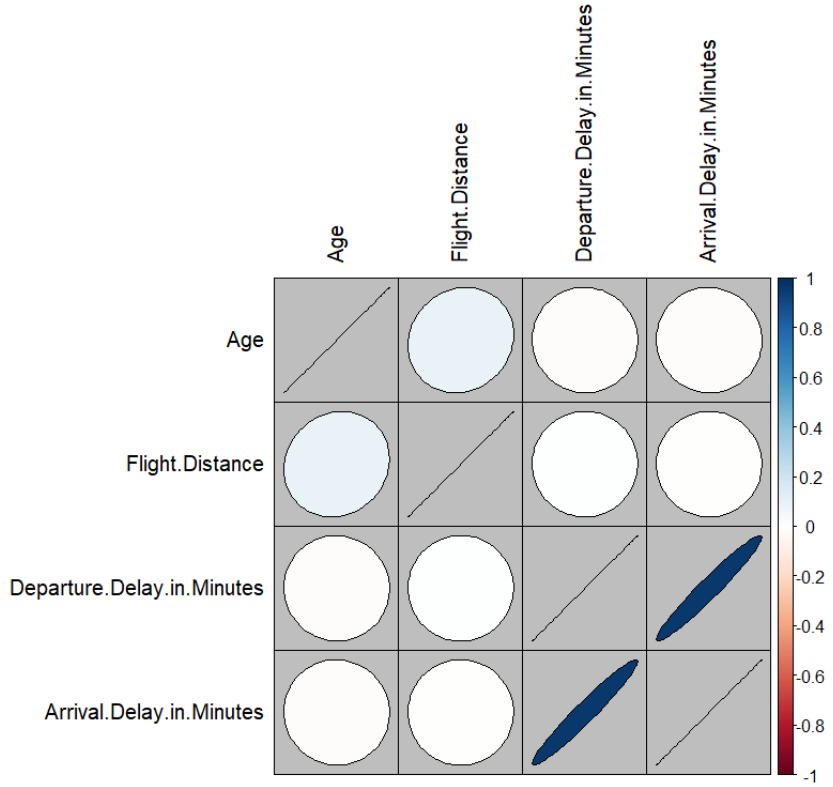
[Figura.11. Boxplot da variável *Departure Delay***]** **[Figura.12.** Boxplot da variável *Arrival Delay***]**

As figuras 11 e 12 mostram uma grande tendência para atrasos de 0 minutos, ou seja, não existiram atrasos, o que demonstra a pontualidade da companhia. Existem, no entanto, algumas exceções, algumas até bastante radicais, sendo o máximo registado de um atraso de partida de 1592 minutos, o que equivale a 26 horas e meia, ou seja, mais de um dia de atraso.



[Figura.13. Gráfico de barras da variável *Satisfaction***]**

Embora existam vários indicadores positivos, a companhia aérea tem, segundo a **Figura.13.,** uma predominância de clientes pouco satisfeitos ou neutros (56.66%) opostamente aos seus clientes satisfeitos (43.34%).



[Figura.14. Gráfico de correlações utilizando o método da elipse**]**

A **Figura.14.** mostra uma correlação acentuada ($R^2 = 0.96548$) entre as variáveis *Departure Delay* e *Arrival Delay.* Isto pode ser explicado pois um atraso na partida leva, muito frequentemente, a um atraso na chegada e uma partida pontual resulta, muito frequentemente, numa chegada pontual.

3.3. Preparação dos Dados

A fase de preparação dos dados é fundamental para o sucesso do modelo, pois os valores em falta e os dados mal agrupados podem ter uma preponderância significativa na eficiência do modelo.

Como, na **Figura.2.,** existem valores em falta e a variável categórica *Id* não apresentava relevância para o estudo, serão esses os primeiros passos.

3.3.1. Limpeza dos Dados

Nem todos os atributos de uma base de dados influenciam aquilo que serão os resultados do seu estudo, pois apenas aumentam o ruído nos dados, o que pode dificultar a aprendizagem, sendo, por isso, irrelevantes. Este é o caso da variável *Id,* como se pode concluir pela **Tabela.2.**

[Tabela.2. Atributos excluídos da base de dados**]**

Atributo	Motivo de Exclusão
Id	Sem relevância para o estudo

Como os valores omissos, apresentados na **Figura.2.** são em número bastante reduzido em relação ao número total de observações então, a sua remoção não apresenta um risco significativo no que toca à modelação e posterior conclusão do estudo.

[Tabela.3. Tratamento de valores omissos**]**

Atributo	Técnica	Observação
Arrival Delay	Case Deletion	Eliminação de 310 registos indefinidos (vazios)

De modo a facilitar a elaboração das regras de associação, o clustering e a classificação, foi necessária a transformação de atributos numéricos em atributos categóricos. A transformação foi feita utilizando os decis dos atributos numéricos alterados.

[Tabela.4. Recodificação do atributo *Age***]**

Código Antigo	Novo Código	Descrição	Freq
[7-20]	0	Idade até aos 20 anos	11296
(20-30]	1	Idade entre os 20 e 30 anos	21363
(30-40]	2	Idade entre os 30 e 40 anos	21137
(40-48]	3	Idade entre os 40 e 48 anos	18845
(48-59]	4	Idade entre os 48 e 59 anos	21271
(59-85]	5	Idade entre os 59 e 85 anos	9682

[Tabela.5. Recodificação do atributo *Flight Distance***]**

Código Antigo	Novo Código	Descrição	Freq
[31-236]	0	Distância do voo até aos 236 km	10486
(236-481]	1	Distância do voo entre os 236 e os 481 km	20599
(481-842]	2	Distância do voo entre os 481 e os 842 km	20715

(842-1512]	3	Distância do voo entre os 842 e os 1512 km	20717
(1512-2751]	4	Distância do voo entre os 1512 e os 2751 km	20723
(2751-4983]	5	Distância do voo entre os 2751 e os 4983 km	10354

[Tabela.6. Recodificação do atributo *Departure Delay***]**

Código Antigo	Novo Código	Descrição	Freq
[0]	0	Atraso na partida de 0 minutos	58552
(0-2]	1	Atraso de partida entre os 0 e os 2 minutos	5205
(2-8]	2	Atraso de partida entre os 2 e os 8 minutos	9736
(8-19]	3	Atraso de partida entre os 8 e os 19 minutos	10027
(19-44]	4	Atraso de partida entre os 19 e os 44 minutos	9950
(44-1592]	5	Atraso de partida entre os 44 e os 1592 minutos	10124

[Tabela.7. Recodificação do atributo *Arrival Delay***]**

Código Antigo	Novo Código	Descrição	Freq
[0]	0	Atraso na chegada de 0 minutos	58552
(0-2]	1	Atraso de chegada entre os 0 e os 2 minutos	5205
(2-9]	2	Atraso de chegada entre os 2 e os 8 minutos	10990
(9-19]	3	Atraso de chegada entre os 8 e os 19 minutos	8773
(19-44]	4	Atraso de chegada entre os 19 e os 44 minutos	9950

(44-1584]	5	Atraso de chegada entre os 44 e os 1592 minutos	10123

Devido ao elevado número de valores "0" nos atributos *Departure Delay* e *Arrival Delay,* a divisão foi feita a partir do decil 50%.

3.4. Modelação

3.4.1. Regras de Associação

As regras de associação são métodos de machine learning com base em regras que procuram detetar padrões em dados transacionais.

No caso de um projeto como este, as regras de associação servem para detetar padrões entre variáveis na base de dados, de modo a encontrar ligações interessantes para a companhia.

Nesta fase de modelação, foi utilizada a função *apriori* de modo a detetar padrões dentro da base de dados.

A associação entre atributos poderá evidenciar situações que, estudadas devidamente, poderão trazer benefícios à companhia.

3.4.2. Clustering

Uma das tarefas mais difíceis ao analisar uma base de dados é conseguir agrupar a informação da mesma, o clustering é uma técnica estatística usada para classificar/agrupar elementos em grupos, de forma que esses elementos sejam parecidos entre si e distintos entre clusters.

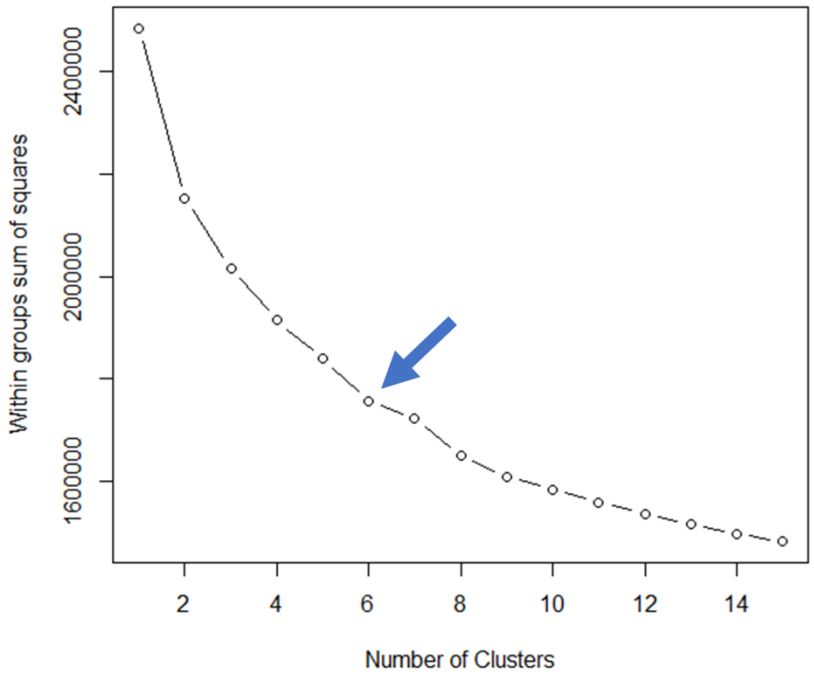
Neste trabalho começamos por aplicar o método gaussiano de misturas de modelos que como o próprio nome indica envolve a mistura de várias distribuições gaussianas e obtivemos a partição da nossa base de dados em 2 grupos.

```
> fit1 <- Mclust(arl2)
fitting ...
  |======================================================================================================================================================| 100%
> summary(fit1)
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust EEV (ellipsoidal, equal volume and shape) model with 2 components:

 log-likelihood      n  df      BIC      ICL
      -2374555 103594 529 -4755218 -4755220

Clustering table:
    1     2
52703 50891
```

[Figura.15. Resultado do summary resultante da aplicação do método gaussiano**]**

Como este método separa os clusters por distribuições gaussianas achamos melhor o uso do método kmeans, que também costuma ser o mais utilizado para a análise de clusters, que reúne os dados e os atribui ao centroide mais próximos, ou seja, significa que cada ponto só pertence a um cluster.É de boa prática

estandardizar os dados antes de uma análise de clustering e assim o fizemos partindo então para a obtenção do número recomendado de grupos que usando o elbow method pareceu-nos indicado a divisão em 6 grupos.



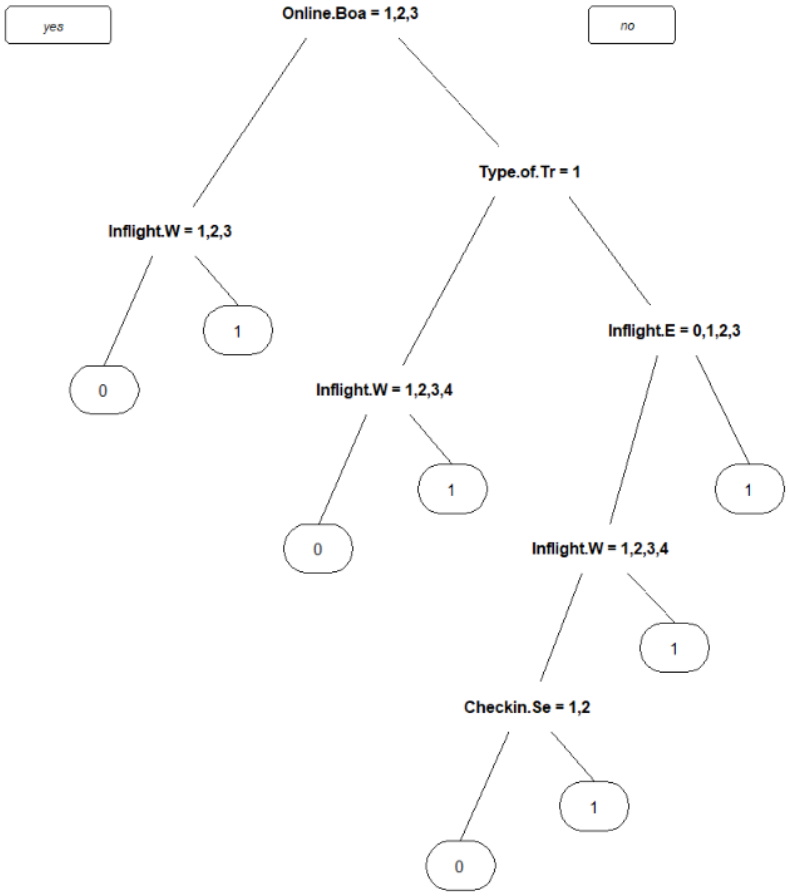
[Figura.16. Gráfico com o número ideal de clusters.**]**

3.4.3. Classificação

A classificação é um processo de previsão que tem como finalidade prever uma variável discreta. Neste caso, essa variável discreta a prever é a variável *Satisfaction* e será usado um modelo linear generalizado.

O modelo linear generalizado (mlg) é uma generalização da regressão linear simples. Este modelo permite-nos obter várias variáveis de resposta, cada qual com modelos de distribuição do erro diferentes de uma normal.

Os atributos inseridos no modelo usado para a classificação são aqueles que são mais preditivos para a variável *Satisfaction* quando é usada a árvore de decisão dada pela função *rpart().* Estes atributos são o *Online Boarding,* o *Type of Travel,* o *Inflight Wi-Fi Service,* o *Inflight Entertainment* e o *Check-In Service.*



[Figura.17. Árvore de decisão da base de dados, para a variável *Satisfaction***]**

3.5. Avaliação dos Resultados

3.5.1. Regras de Associação

De modo a identificar as relações mais importantes na base de dados, podemos utilizar medidas tais como o *support,* a *confidence* e o *lift.*

O alvo de cada regra de associação é um valor da variável *Satisfaction* pois é esta que queremos prever.

As regras vão ser ordenadas por lift e elaboradas pelo algoritmo *apriori*. Os valores mínimos de *support, confidence* e *maxlen* (número máximo de atributos por regra) serão, respetivamente, 0.02, 0.5 e 3.

[Tabela.8. Resultados das regras de associação para a variável *Satisfaction* ordenadas por *lift***]**

lhs	support	confidence	coverage	lift	count
{Inflight.Wi.Fi.Service=5}	0.1093693	0.9907310	0.1103925	2.285983	11330
{Online.Boarding=5}	0.1737938	0.8717800	0.1993552	2.011519	18004

{Inflight.Entertainment=5, Leg.Room.Service=5}	0.1066085	0.8686487	0.1227291	2.004294	11044
{On.board.Service=5, Leg.Room.Service=5}	0.1046586	0.8661820	0.1208275	1.998603	10842
{Leg.Room.Service=5, Inflight.Service=5}	0.1077476	0.8324881	0.1294283	1.920858	11162

Observando a **Tabela.8.** podemos reparar que três das cinco regras de associação com maior *lift* apresentam a variável *Leg Room Service*, o que é um indicador de que esta variável tem uma relação importante na satisfação dos clientes. A regra de associação com maior *lift* tem como *lhs* o atributo *Inlight Wi-Fi Service,* o que é natural dada a era tecnológica em que vivemos hoje em dia. Esta pode ser também a razão para a colocação da segunda regra de associação com maior *lift,* o atributo *Online Boarding.*

Atentando nestas três variáveis, obtêm-se os resultados das tabelas 9, 10 e 11.

[Tabela.9. Resultados das regras de associação para a variável *Satisfaction* com *rhs Inflight Wi-Fi Service***]**

lhs	rhs	support	confidence	coverage	lift	count
{Inflight.Wi.Fi.Service=0}	{Satisfaction=1}	0.02980868	0.9974160	0.0298859	2.301408	3088
{Inflight.Wi.Fi.Service=5}	{Satisfaction=1}	0.10936927	0.9907310	0.1103925	2.285983	11330
{Inflight.Wi.Fi.Service=1}	{Satisfaction=0}	0.11578856	0.6745965	0.1716412	1.190591	11995
{Inflight.Wi.Fi.Service=4}	{Satisfaction=1}	0.11411858	0.5989765	0.1905226	1.382061	11822
{Inflight.Wi.Fi.Service=2}	{Satisfaction=0}	0.18674827	0.7511551	0.2486148	1.325709	19346
{Inflight.Wi.Fi.Service=3}	{Satisfaction=0}	0.18656486	0.7494281	0.2489430	1.322661	19327

A **Tabela.9.** mostra que aqueles que não responderam e aqueles que responderam positivamente tendem a estar satisfeitos.

[Tabela.10. Resultados das regras de associação para a variável *Satisfaction,* com *rhs Online Boarding***]**

lhs	rhs	support	confidence	coverage	lift	count
{Online Boarding=0}	{Satisfaction=1}	0.01300268	0.5566116	0.02336043	1.284309	1347
{Online Boarding=1}	{Satisfaction=0}	0.08868274	0.8619816	0.10288241	1.521306	9187
{Online Boarding=2}	{Satisfaction=0}	0.14901442	0.8846925	0.16843640	1.561389	15437
{Online Boarding=5}	{Satisfaction=1}	0.17379385	0.8717800	0.19935518	2.011519	18004
{Online Boarding=3}	{Satisfaction=0}	0.18138116	0.8641464	0.20989633	1.525127	18790
{Online Boarding=4}	{Satisfaction=1}	0.18446049	0.6230315	0.29606927	1.437564	19109

A **Tabela.10.** mostra que, não só os cliente que responderam *Online Boarding* positivamente, mas também os que não responderam ou não sabiam acabaram satisfeitos.

[Tabela.11. Resultados das regras de associação para a variável *Satisfaction* , com *rhs Leg Room Service***]**

lhs	rhs	support	confidence	coverage	lift	count
{Leg.Room.Service=1}	{Satisfaction=0}	0.07935788	0.7973812	0.09952314	1.407293	8221
{Leg.Room.Service=2}	{Satisfaction=0}	0.13633994	0.7254610	0.18793559	1.280362	14124
{Leg.Room.Service=3}	{Satisfaction=0}	0.14074174	0.7274723	0.19346680	1.283912	14580
{Leg.Room.Service=5}	{Satisfaction=1}	0.14586752	0.6142933	0.23745584	1.417402	15111
{Leg.Room.Service=4}	{Satisfaction=1}	0.16143792	0.5826366	0.27708168	1.344358	16724

A **Tabela.11.** apresenta os resultados das regras de associação entre as variáveis *Leg Room Service* e *Satisfaction,* mostrando que apenas os clientes que responderam positivamente ao questionário se mostraram satisfeitas.

3.5.2. Clustering

Com a divisão em 6 grupos e efetuando o comando kmeans conseguimos obter grupos com as seguintes dimensões: 18263, 18597, 18907, 13568, 19244, 15015. O que não torna uma visualização fácil, mas são muitos dados a serem agrupados.

[Figura.18. Representação gráfica dos clusters num plano de duas dimensões**]**

Passamos então a uma análise dos 6 grupos formados.

O nosso objetivo é entender o que leva uma pessoa a qualificar o seu voo como satisfeito ou insatisfeito.

```
> table(km6$cluster, arl$Satisfaction)

        0     1
1 15519  2744
2   525 18072
3 17817  1090
4 12109  1459
5  1249 17995
6 11478  3537
```

```
#GRAU Satisfação
#Grupo1: 2744/(15519+2744) = 0.1502491
#Grupo2: 18072/(525+18072) = 0.9717696
#Grupo3: 1090/(17817+1090) = 0,0577035
#Grupo4: 1459/(12109+1459) = 0,1075324
#Grupo5: 17995/(1249+17995) = 0,9350976
#Grupo6: 3537/(11478+3537) = 0,2355644
```

[Figura.19. Tabela com o número de clientes satisfeitos e insatisfeitos por cluster e graus de satisfação de cada cluster **]**

Percebemos então que o grupo com melhor grau de satisfação é o segundo.

Por último só nos falta perceber quais as características de cada grupo e, neste caso, como é que elas influenciaram na decisão da classificação de satisfeito ou insatisfeito com o seu voo.

Depois de utilizar uma simples função de agregação da nossa base de dados com os grupos obtidos a partir da função kmeans, tendo como comparação as médias das variáveis dos grupos dados, conseguimos concluir que as pessoas do grupo 2 deram mais importância à facilidade de check-in online, do conforto do assento, das comodidades a bordo do avião bem como do serviço e a maneira como a sua bagagem foi tratada.

Para o grupo 3 que foi o que considerou a sua viagem mais insatisfeita concluímos que os motivos que os levaram a tomar essa decisão foram os atrasos tanto na partida como na chegada sendo que os seus voos eram de distâncias mais curtas.

3.5.3. Classificação

A classificação foi feita utilizando o modelo linear generalizado, pois este é um modelo mais simples. Para ser realizada esta etapa, a base de dados foi dividida em dados de treino e dados de teste, sendo a divisão feita com 80% e 20% dos dados, respetivamente.

Após a divisão, foi feita a modelação com o *glm*, utilizando as variáveis mais preditivas, dadas pela árvore de decisão da **Figura.15.**

Assim, foi obtida uma matriz de confusão, como forma de observar os dados de maneira mais simples e de fácil interpretação.

A matriz apresentada na **Tabela.12.** apresenta uma percentagem de sucesso de 91.98%, provando que é um bom modelo.

Na diagonal, a negrito estão representados os resultados previstos corretamente.

[Tabela.12. Matriz de confusão do modelo de classificação**]**

		Referência	
		0	1
Previsões	0	**10759**	819
	1	833	**8184**

3.6. Implementação

Os modelos aqui trabalhados não serão utilizados no quotidiano, contudo, seria possível estes serem implementados, onde clientes de companhias aéreas pudessem ter parte ativa na melhoria dos serviços prestados pela empresa.

O primeiro passo seria detetar possíveis motivos de descontentamento com quaisquer atributos de uma aeronave por parte dos clientes, dando oportunidade aos responsáveis de ajudar e providenciar soluções para melhorar a viagem de cada passageiro.

Visto que o modelo prevê a satisfação de clientes de forma bastante eficiente será, também possível encontrar as zonas de maior descontentamento, mas também as zonas de maior contentamento, sendo mais simples efetuar um estudo de forma a melhorar todos os serviços, de acordo com a sua necessidade ou urgência.

Uma vez que são mostrados pelo modelo, adicionalmente, grupos de clientes com características semelhantes e gostos semelhantes, poderá ser traçado um plano para cada um desses grupos, de modo a melhorar a experiência de várias pessoas conjuntamente e, por isso, tratar de forma mais rápida os problemas, uma vez que se estariam a tratar de grupo de pessoas e não de individualidades.

Ao fim do tempo necessário para implementar estes modelos, podem ser inseridos novos dados aos modelos, o que resultará em previsões ainda melhores, melhorando o algoritmo preditivo.

4. Conclusão

Os objetivos do trabalho foram alcançados. A base de dados foi trabalhada utilizando a metodologia CRISP-DM e foram modelados utilizando métodos de data mining, tais como as regras de associação, o clustering e a classificação. Isto permitiu a previsão de resultados utilizando o modelo *glm* de classificação, resultando num modelo eficiente, com 91.98% de sucesso. Com isto foram consolidados os conceitos lecionados nas aulas da unidade curricular de Data Mining para Ciência de Dados.

Existiu uma melhoria significativa na pesquisa de conteúdo, na construção de gráficos utilizando o *ggplot*, na implementação de regras de associação, na construção de modelos de clustering e na obtenção de previsões utilizando métodos de classificação. A metodologia CRISP-DM foi estudada e utilizada, podendo ser recordada para projetos futuros.

O balanço final é positivo, podendo estes modelos ser aplicados no quotidiano.

O trabalho poderia, obviamente, ser melhorado, recorrendo a outros modelos, mais regras de associação, estudos mais profundos e outros métodos.

No entanto, visto os objetivos terem sido alcançados e completos, e de forma a não tornar este trabalho tão extenso, nada mais se acrescenta.

5. Bibliogafia

- MUKHERJEE S., PACKT PUBLISHING (2016) F# for Machine Learning Essentials: Get up and running with machine learning with F# in a fun and functional way, 1st Edition;

- KASSAMBARA A., STHDA (2018) Machine Learning Essentials, 1st Edition;

- PENG R., Lulu.com (2016) Exploratory Data Analysis with R, 1st Edition;

- CORTEZ P., SILVA M., Using data mining to predict secondary school student performance, 2008;

- https://smolski.github.io/livroavancado/analise-de-clusters.html;

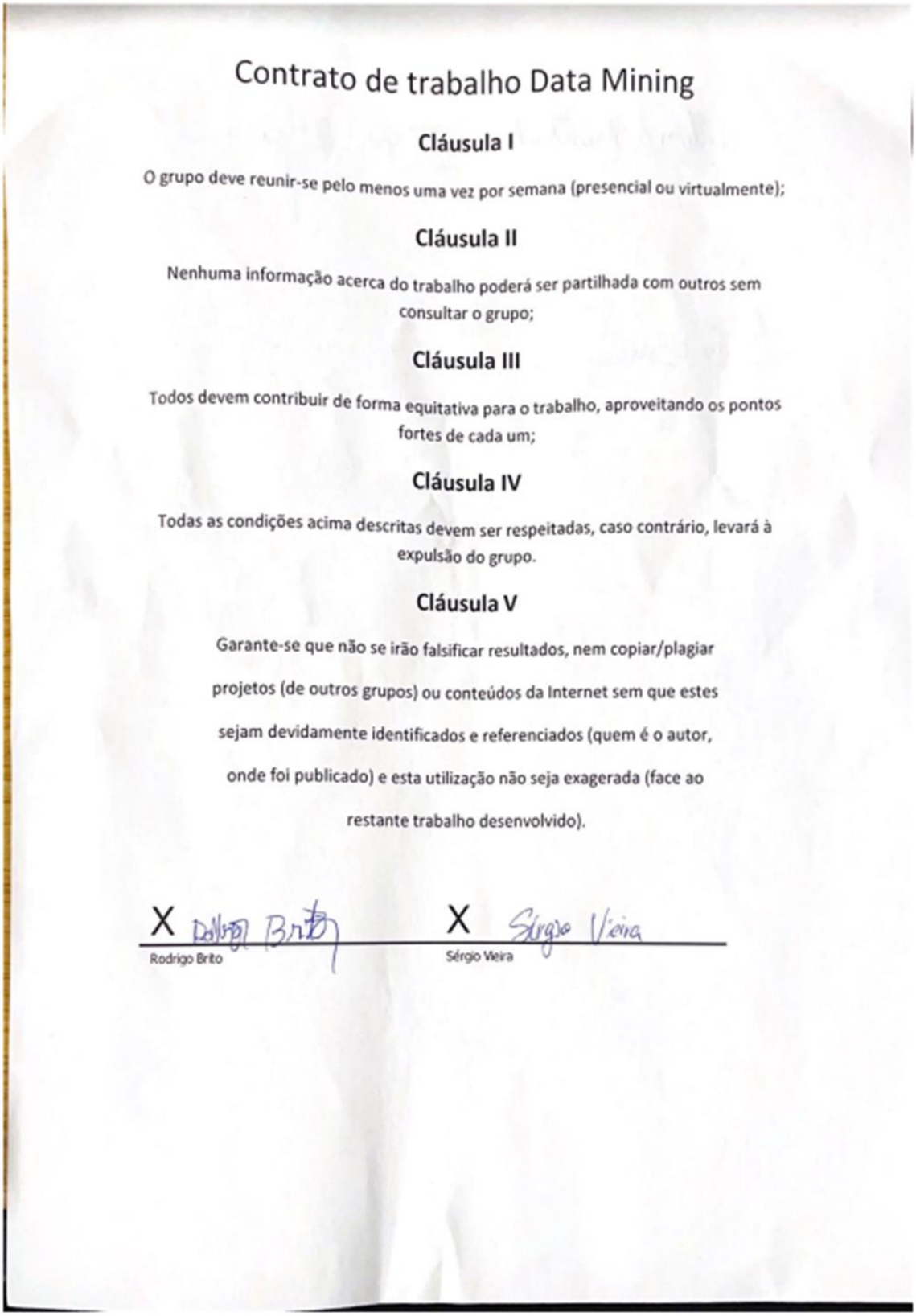
- https://www.kaggle.com/datasets/vardhansiramdasu/fraudulent-transactions-prediction;

- https://stackoverflow.com/questions/27926131/how-to-get-items-for-both-lhs-and-rhs-for-only-specific-columns-in-arules;

- http://www.stat.columbia.edu/~tzheng/files/Rcolor.pdf;

- https://www.kaggle.com/code/frixinglife/airline-passenger-satisfaction/data.

6. Anexos



Contrato de trabalho Data Mining

Cláusula I

O grupo deve reunir-se pelo menos uma vez por semana (presencial ou virtualmente);

Cláusula II

Nenhuma informação acerca do trabalho poderá ser partilhada com outros sem consultar o grupo;

Cláusula III

Todos devem contribuir de forma equitativa para o trabalho, aproveitando os pontos fortes de cada um;

Cláusula IV

Todas as condições acima descritas devem ser respeitadas, caso contrário, levará à expulsão do grupo.

Cláusula V

Garante-se que não se irão falsificar resultados, nem copiar/plagiar projetos (de outros grupos) ou conteúdos da Internet sem que estes sejam devidamente identificados e referenciados (quem é o autor, onde foi publicado) e esta utilização não seja exagerada (face ao restante trabalho desenvolvido).

X Rodrigo Brito — Rodrigo Brito

X Sérgio Vieira — Sérgio Vieira

X António Martinho

X Miguel Mesa

X Paulo Barros