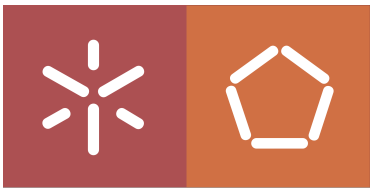
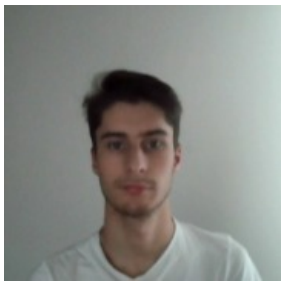


Universidade do Minho
Escola de Engenharia

Aprendizagem e Decisão Inteligentes

Relatório Trabalho Prático
LEI - 3º Ano - 2º Semestre

Braga, 19 de fevereiro de 2025


Lucas Oliveira A98695


Mike Pinto A89292


Rafael Gomes A96208

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Este relatório visa documentar o trabalho desenvolvido no âmbito da Unidade Curricular de Aprendizagem e Decisão Inteligente do 2.º semestre do 3.º ano do curso de Engenharia Informática da Universidade do Minho.

Neste contexto, o trabalho desenvolvido abrange a exploração e análise detalhada de ambos os conjuntos de dados, destacando os aspetos relevantes e as estratégias de pré-processamento adotadas. Além disso, são apresentados os modelos de Machine Learning elaborados para cada *dataset*, explicando as decisões tomadas durante o processo de desenvolvimento.

Este relatório propõe-se a oferecer uma visão abrangente do trabalho realizado, contribuindo assim para o entendimento e aprofundamento dos conceitos abordados na disciplina.

Capítulo 2

Tarefa 1: *Músicas Spotify*

2.1 *Dataset*

O *dataset* escolhido pelo grupo consiste num conjunto de dados sobre músicas do *Spotify*, que contem 49986 linhas e 18 atributos.
Optou-se por este *dataset* devido a ser um tema interessante e à oportunidade de testar um modelo de regressão. Sendo que a equipa docente focou-se no de classificação, decidimos explorar a outra vertente.
A metodologia utilizada nesta tarefa para a resolução do problema é o **CRISP-DM**, que consiste nestas 6 fases seguintes:

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

Sendo que a última fase não será explorada, por não se enquadrar nos conteúdos lecionados.

2.2 Estudo do Negócio

O objetivo é prever a popularidade de uma música e identificar os atributos que contribuem para a tornar popular.

2.3 Estudo dos Dados

Os atributos presentes no conjunto de dados são os seguintes:

- **id**: Código de identificação de cada música.

- **artist_name**: Nome do artista.
- **track_name**: Nome da música.
- **popularity**: Indica a popularidade da música, entre 0 e 99.
- **acousticness**: Indica a presença de elementos acústicos na música, entre 0 e 1.
- **danceability**: Indica o quão adequada uma música é para dançar, entre 0 e 1.
- **duration_ms**: Duração da música em milissegundos.
- **energy**: Indica o nível de energia na música, entre 0 e 1.
- **instrumentalness**: Indica a quantidade de vozes presentes na música, entre 0 e 1.
- **key**: Tonalidade da música.
- **liveness**: Indica a presença de audiência durante a gravação da música, entre 0 e 1.
- **loudness**: Indica o volume médio da música em decibéis, entre 0 e 1.
- **mode**: Indica caso a música seja *Major* ou *Minor*.
- **speechiness**: Indica a presença de palvras referidas numa música, entre 0 e 1.
- **tempo**: Indica o ritmo da música em batidas por minuto.
- **obtained-date**: Indica a data em que a música foi obtida.
- **valence**: Indica a positividade da música, entre 0 e 1.
- **genre**: Indica o género da música.

2.3.1 Análise de Cada Atributo

artist_name: As músicas onde o nome do artista, em média, possui mais popularidade é o *"Duki"* com 82, já com menor média é o *"O Rappa"* e o *"Jorma Kaukonen"* com 1. O que indica que o nome do artista pode ter uma boa influência na popularidade de uma música.

track_name: Ao analisar o atributo, não se conseguiu concluir a sua influência relativamente à popularidade, pois cada música possui um nome diferente, portanto considerou-se irrelevante para a resolução do problema.

acousticness: Para analisar este atributo, foi feita uma subdivisão em grupos de intervalos de 0.1, em que calculou-se a média de cada grupo para poder visualizar a influência na popularidade, sendo que a com menor média é a *acousticness* de 0 que possui uma média de 23 de popularidade, já a com maior média é a *acousticness* de 0.2 (valores entre 0.11 até 0.2) com uma média de 52.47 de

popularidade, também observou-se que a correlação entre os dois atributos tem um valor de -0.29. Então com isso chegou-se à conclusão que seria um atributo a considerar para a previsão da popularidade.

danceability: Para analisar este atributo, foi feito o mesmo processo que o atributo anterior, onde vimos que a com menor média é a *danceability* de 0.1 que contém uma média de 36 de popularidade, já com maior média é a *danceability* de 1 com uma média de 59 de popularidade, ou seja, quanto maior for a *danceability*, maior será a popularidade da música, concluindo assim, que este atributo possui uma grande influência na popularidade.

duration_ms: Ao analisar este atributo, observou-se que quando a música passa de uma média de 250000ms (4.2 minutos), a popularidade tende a piorar. Levando então este atributo em consideração para a previsão da popularidade.

energy: Analisando este atributo foi usado o mesmo processo que o *acousticness*, onde vimos que a energia da música com menor média foi a *energy* de 0.1 com uma média de 37 de popularidade, já a com maior média é a *energy* de 0.6 com uma média de 53 de popularidade, o que também assimilou-se, é que, quando a energia de música é muito alta ou muito baixa, a popularidade tende a ser pior do que quando a energia é moderada. Concluindo que este atributo também ajuda-nos a prever a popularidade da música.

instrumentalness: Repetindo o mesmo processo anterior, vimos que a com menor média é a *instrumentalness* de 1 com uma média de 38 de popularidade, já a com maior é a *instrumentalness* de 0 com uma média de 65 de popularidade, ou seja, quanto mais vozes presentes na música, pior a popularidade dela. Concluindo assim, que devemos também considerar este atributo.

key: Analisando este atributo, vimos que não iria possuir grande influência relativamente à popularidade, pois a *key* com maior popularidade é a **C#** com uma média de 61 de popularidade, já a menor é a **D#** com uma média de 55.2 de popularidade, sendo a diferença entre a maior e a menor apenas de 6 valores de popularidade e também ao analisar a correlação entre as duas e vermos que não existia correlação entre elas, conclui-se que este atributo é irrelevante para a análise do problema.

liveness: Repetindo o processo habitual, observou-se que com menor média é a *liveness* de 1 com a popularidade de 41.5, já com maior média é a *liveness* de 0.2 com a popularidade de 52, ou seja, quanto mais presença de público na música, pior a popularidade. O que levou também a considerar este atributo.

loudness: Analisando este atributo, vimos que quanto menor os decibéis de uma música pior a popularidade, já quanto maior os decibéis, maior a popularidade. Considerando assim, este atributo.

mode: Explorando este atributo, observou-se que não possui grande influência relativamente à popularidade, pois o *mode* **Major** possui uma popularidade média de 65 e o *mode* **Minor** possui uma popularidade média de 63.3, sendo a diferença muito minima e vendo que a correlação entre os dois atributos não

exisitia, considerou-se irrelevante para a previsão.

speechiness: Repetindo o processo habitual, observou-se que o *speechiness* com menor média foi o 1 com uma média de 46 de popularidade, já com maior foi o 0.5 com uma média de 55.5 de popularidade, ou seja, tanto muita ou pouca presença de palavras referidas numa música piora a popularidade, já quando é moderada, tende a ser melhor a popularidade. Levando assim também em consideração este atributo para a previsão.

tempo: Averiguando este atributo, inicialmente considerou-se que iria ter influência na popularidade, mas após observar a figura 2.1, verificou-se que poderia induzir em erro os modelos de previsão, pois entre os *tempos* 120 e 135 existe uma variância grande entre as popularidades, o que poderia levar a um erro para nos modelos. Sendo então essa a razão, não se considerou o *tempo* para a previsão da popularidade.

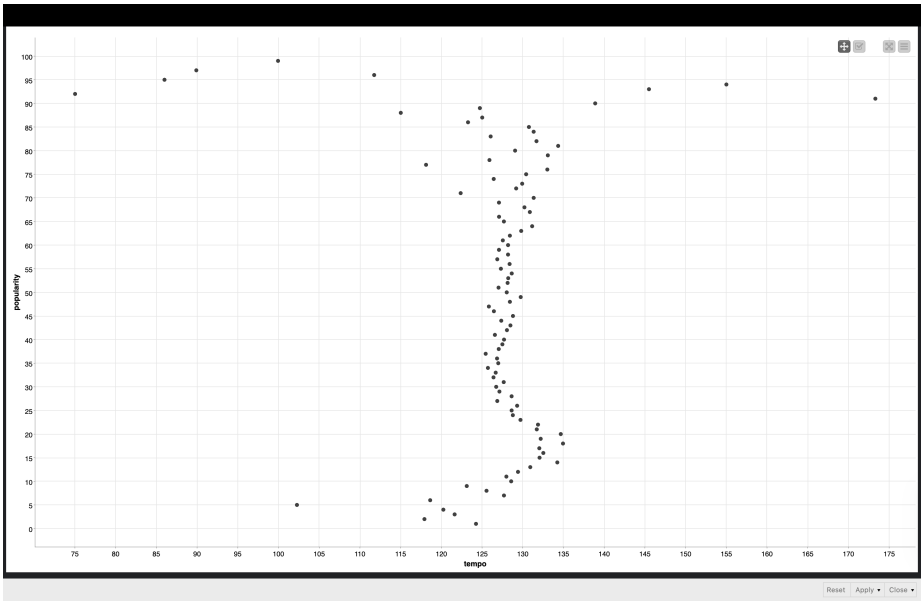


Figura 2.1: *Scatter plot* das médias dos tempos relativamente à popularidade

obtained-date: Analisando este atributo, verificamos que não iria ter grande influência para a popularidade de uma música.

valence: Repetindo o processo, observou-se que o *valence* com menor média foi o valor 0 com uma média de 46 de popularidade, já com maior foi o valor 0.5 com uma média de 51 de popularidade, e analisando melhor averiguou-se que normalmente quando o *valence* é superior a 0.3, a popularidade é maior do que quando é inferior a 0.3. Concluindo assim, que este atributo também será considerado para a previsão da popularidade.

genre: Por fim, antes de analisar-se este atributo, já considerou-se que iria influenciar pelos conhecimentos sobre o assunto, e após executar a análise, verificou-

se que existiam 3 géneros que normalmente têm uma popularidade considerável, sendo esses ***Rock***, ***Rap*** e ***Hip-Hop*** que possuem uma média de popularidade entre 64 e 69, já pelo contrário existem 2 géneros que normalmente têm uma popularidade tão boa, sendo esses ***Anime*** e ***Classical*** que possuem uma média de popularidade entre 42 e 45, os restantes géneros que são ***Jazz***, ***Blues***, ***Electronic***, ***Alternative*** e ***Country*** possuem uma média de popularidade entre 51 e 57. Concluindo então que este atributo também influencia bastante na popularidade de uma música.

2.4 Preparação dos Dados

2.4.1 Análise Prévia ao Tratamento de Dados

Para iniciar a análise, utilizou-se o nodo *"Data Explorer"* para verificar a presença de *missing values*. Observou-se que todas as variáveis do conjunto de dados apresentam 5 valores em falta, o que indica a necessidade de tratamento desses valores ausentes. Além disso, identificou-se a necessidade de converter o tipo de dados de algumas variáveis, conforme apresentado na figura 2.2.

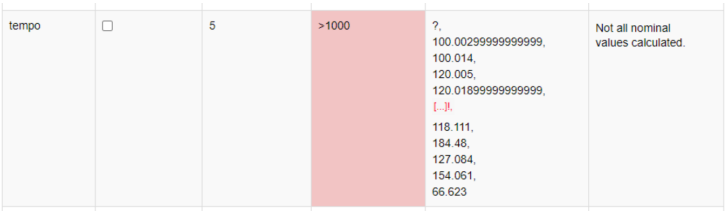


Figura 2.2: Visualização de um atributo através da utilização do *DataExplorer*.

Utilizamos o *Box Plot* para visualizar a distribuição dos dados e identificar possíveis valores discrepantes. Observamos a presença de valores muito altos que exigem correção.

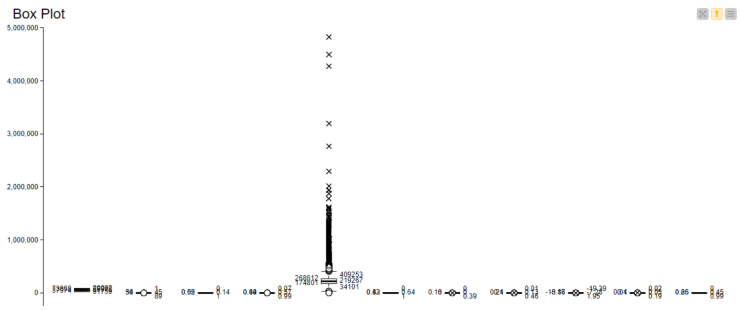


Figura 2.3: *Box Plot*

Por meio do *Line Plot*, também analisou-se a distribuição dos dados ao longo do tempo. Observou-se uma inclinação irregular tanto para o lado positivo quanto negativo, o que sugere um desequilíbrio na distribuição dos dados.

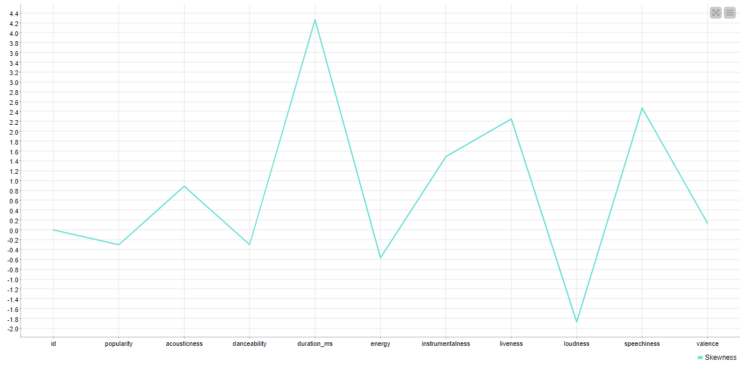


Figura 2.4: *Line Plot*

2.4.2 Tratamento Geral dos Dados

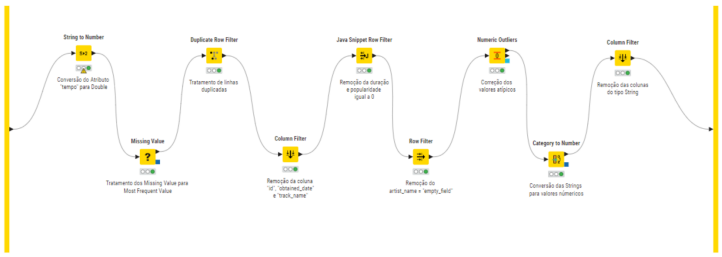


Figura 2.5: *Tratamento de Dados*

Conversão do atributo *"tempo"* para *Numeric*

Inicialmente, realizou-se a conversão do atributo *"tempo"* para o formato numérico, facilitando assim, a sua manipulação e análise.

Tratamento de *Missing Values*

Corrigiu-se os valores ausentes utilizando a estratégia *"Most Frequent Value"*, através do nodo *"Missing Value"*.

Remoção de Linhas Duplicadas e com Valores Nulos

Eliminaram-se as linhas duplicadas do conjunto de dados para evitar redundâncias e inconsistências. Além disso, excluíram-se as linhas que possuíam duração ou popularidade igual a zero, pois esses valores podem afetar negativamente a qualidade da análise.

Filtragem de Colunas

Foram removidas as colunas *"id"*, *"obtained_date"*, *"track_name"*, *"key"*, *"mode"* e *"tempo"*, pois foram consideradas como irrelevantes.

Remoção de Linhas com o Valor de *empty-field*

Foram eliminadas todas as linhas que possuíam o valor *"empty-field"* no atributo *"artist_name"*.

Correção de *Outliers*

Foram corrigidos os valores exponenciais que representavam *outliers*, substituindo-os pelo valor mais próximo permitido.

Conversão de Atributos para Numérico

Converteram-se os atributos *"artist_name"* e *"genre"* para valores numéricos, e posteriormente as suas colunas de origem foram filtradas.

2.4.3 Análise Posterior ao Tratamento de Dados

Após o tratamento de dados, realizou-se uma verificação abrangente para avaliar a qualidade do conjunto de dados. Sendo feitas as seguintes observações:

Todos os valores ausentes foram corrigidos durante o tratamento de dados, garantindo que não haja *missing values* no *dataset*. Os valores atípicos identificados anteriormente foram tratados e normalizados, garantindo a não distorção dos dados, como apresentado na figura 2.6.

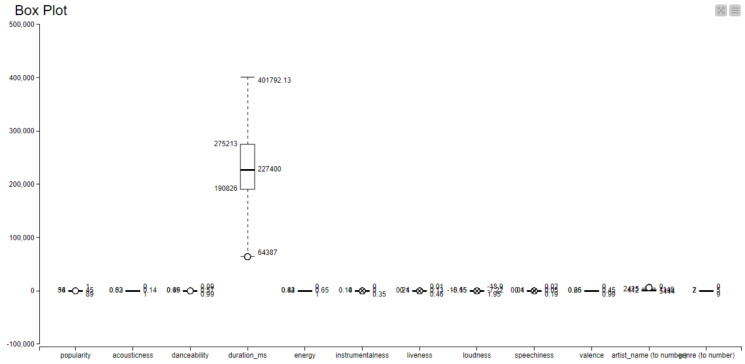


Figura 2.6: *Box Plot*

Ao analisar a assimetria dos dados usando a medida estatística *"Skewness"*, observou-se uma melhoria relativamente à irregularidade detetada anteriormente. Sugerindo assim, uma distribuição mais balanceada e adequada dos dados, como resultado na figura 2.7.

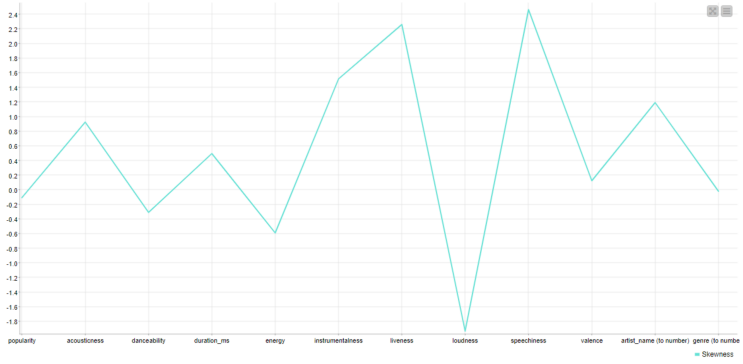


Figura 2.7: *Line Plot*

2.5 Modelação

2.5.1 Modelação com Todos os Atributos

Inicialmente, criamos um modelo de previsão que considerava todos os atributos disponíveis. Dividimos a previsão de dados em duas partes: uma com dados sem nenhuma alteração e outra com os dados normalizados. Em cada parte, aplicamos dois métodos de particionamento diferentes: um com o nodo *"X-Partitioner"* e outro com o nodo *"Partitioning"*. Posteriormente, em cada partição, utilizamos três algoritmos de aprendizagem: *Linear Regression Learner*, *Simple Regression Tree Learner* e *Gradient Boosted Tree Learner (Regression*.

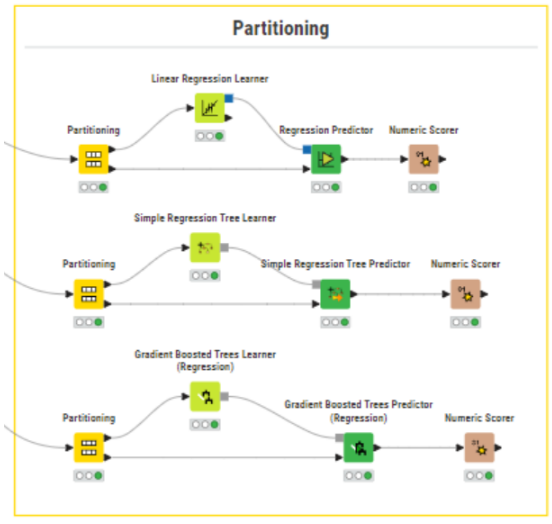


Figura 2.8: Exemplo de um dos métodos de previsão

Sendo que o algoritmo que teve o melhor valor foi o *Gradient Boosted Tree Learner (Regression)*, normalizado e com uso do nodo *"Partitioning"*, com o resultado apresentado na figura 2.9.

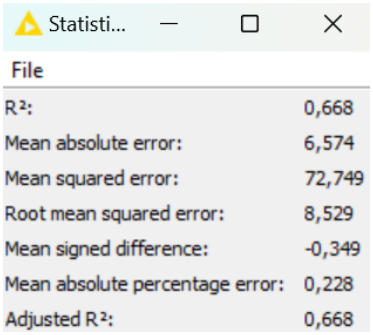


Figura 2.9: Resultado do *Gradient Boosted Tree Leraner (Regression)*

2.5.2 Modelação com Seleção dos Melhores Atributos

Para identificar os atributos mais significativos, aplicamos uma estratégia de *Feature Selection*, apresentado na figura 2.10. Após a análise detalhada, foram selecionados os seguintes atributos: *acousticness*, *danceability*, *duration_ms*, *instrumentalness*, *valence*, *artist_name* e *genre*.

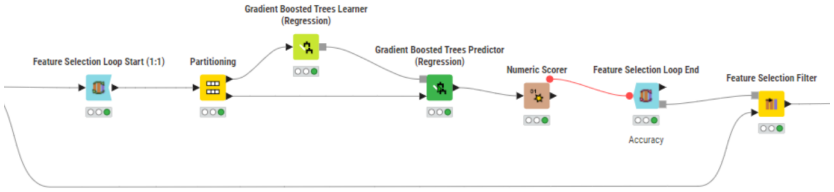


Figura 2.10: Modelação com o uso do *Feature Selection*

Ao analisar o R^2, observou-se que o seu valor foi *0,669*. Embora haja uma melhoria mínima relativamente ao modelo anterior, não se notou grande alteração.

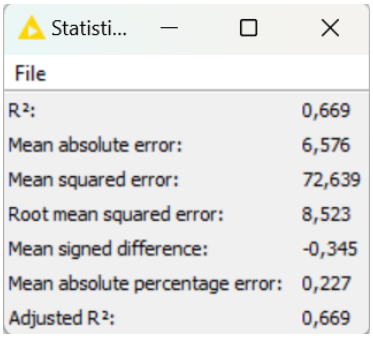


Figura 2.11: Melhor resultado do uso do *Feature Selection*

2.5.3 Modelação com o Objetivo de Regressão

Com o intuito de converter este problema de regressão num problema de classificação, através do nodo *"Rule Engine"*, foi feita uma conversão da popularidade

dade para um *"Star_Rating"*. Sendo essa nova variável, delimitada entre 1 a 5 estrelas. Em seguida, foram aplicados os seguintes algoritmos: *Decision Tree Learner*, *Random Forest Learner* e *Gradient Boosted Trees Learner*.
Sendo o *Gradient Boosted Trees Learner* o que apresenta uma melhor *accuracy* de 66.07%.

2.6 Avaliação

De um modo geral, foram identificados os atributos com mais influência para a previsão dos dados mediante uma análise cuidadosa dos mesmos. O estudo e tratamento dos dados provenientes do *dataset* foi minuciosamente pensado, para a melhor previsão possível. Embora os modelos não tenham alcançado valores altos de precisão, considerou-se que os resultados foram razoáveis, dadas as circunstâncias. É importante notar a dimensão do conjunto de dados e a medida da popularidade possui uma escala considerável, o que pode influenciar os resultados obtidos.

Capítulo 3

Tarefa 2: *Liver Disease*

3.1 *Dataset*

O conjunto de dados atribuído para a nossa análise possui 583 linhas e 17 atributos, abrangendo informações sobre a presença ou ausência de doenças hepáticas em indivíduos. Este conjunto contém dados relevantes para a investigação e análise de condições relacionadas ao fígado, permitindo-nos explorar padrões e tendências que possam estar associados a essas doenças.
A metodologia utilizada para a resolução desta tarefa foi novamente o método **CRISP-DM**.

3.2 Estudo do Negócio

O objetivo principal do estudo é desenvolver um modelo capaz de prever se uma pessoa possui doença hepática. Além disso, procuramos identificar quais atributos do conjunto de dados que podem ter maior influência para contrair a doença.

3.3 Estudo dos Dados

- **id_code:** Código de identificação do paciente.
- **Age:** Idade do paciente.
- **birth_year:** Ano de nascimento do paciente.
- **birth_month:** Mês de nascimento do paciente.
- **birth_date:** Data de nascimento do paciente.
- **Gender:** Género do paciente.
- **TB:** Bilirrubina total.
- **DB:** Bilirrubina direta.
- **Alphos:** Fosfatase alcalina.

- **Sgpt:** Alanina aminotransferase (ALT).
- **Sgot:** Aspartato aminotransferase (AST).
- **TB(#1):** Bilirrubina total (#1).
- **ALB:** Albumina.
- **CHOL:** Colesterol.
- **AG_Ratio:** Proporção entre albumina e globulina.
- **Bilmg:** Bilirrubina em miligramas.
- **Selector:** Indicador de presença ou ausência de doença hepática.

3.3.1 Análise de Cada Atributo

Selector: Antes de analisar os outros atributos, verificamos que o *Selector* tinha 3 classificações, sendo que duas delas consideramos que seriam idênticas, que foi o caso de *"2=no liver disease"* e *"3=without liver disease"*, então decidimos utilizar o nodo *"Rule Engine"*, para redefinir a coluna *"Selector"* para representar apenas dois valores: *"YES"* para casos de doença hepática e *"NO"* para ausência de doença hepática. Em seguida, renomeamos a coluna *"Selector"* para *"Liver Disease"* para maior clareza e melhor análise dos atributos seguintes.

Age: Ao analisar o atributo idade, observamos uma tendência significativa numa certa faixa etária ser mais propensa a contrair a doença, que abrange pessoas entre os 25 e os 65 anos. Essa descoberta sugere que a idade pode ser um atributo importante na identificação de indivíduos com maior risco de desenvolver a doença.

bith_year, birth_month, birth_data: Ao analisar os atributos, observamos que a única informação relevante que poderíamos extrair deles é a idade. No entanto, como já temos o atributo "Age"que fornece essa informação de maneira mais direta e precisa, consideramos esses atributos irrelevantes para o problema proposto.

Gender: Analisando este atributo, apesar de no nosso conjunto de dados, existirem mais homens do que mulheres, observamos que os homens tendem a contrair mais a doença do que as mulheres, como podemos ver na figura 3.1, o que indica que este atributo tem uma relevância grande para a nossa previsão.

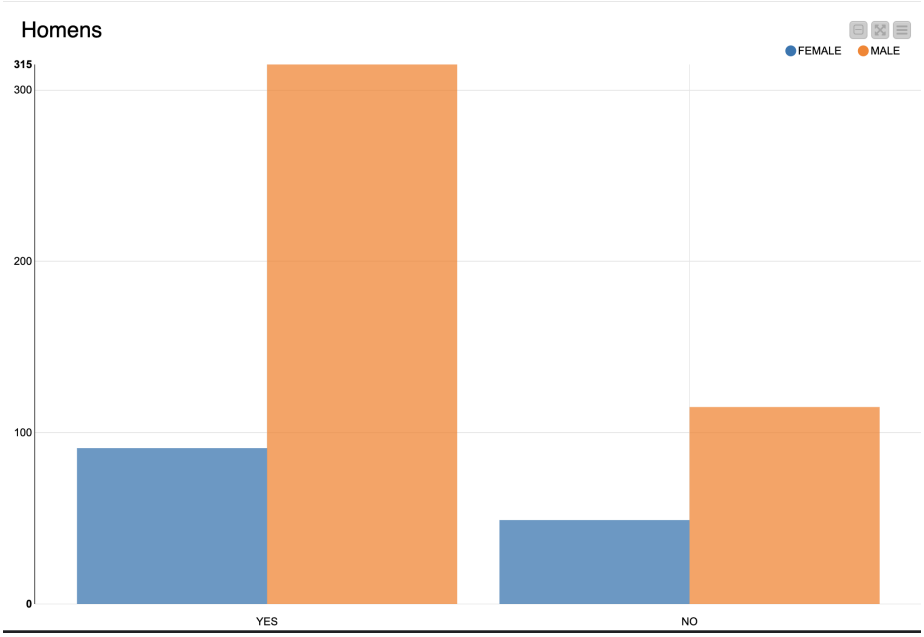


Figura 3.1: Gráfico de barras dos géneros com doença e sem doença

TB, DB, Alphos, Sgpt, Sgot, ALB, AG_Ratio, Bilmg: Analisando estes atributos, verificamos que grande parte deles teria boa influência para a previsão da doença, como podemos ver na figura 3.2.

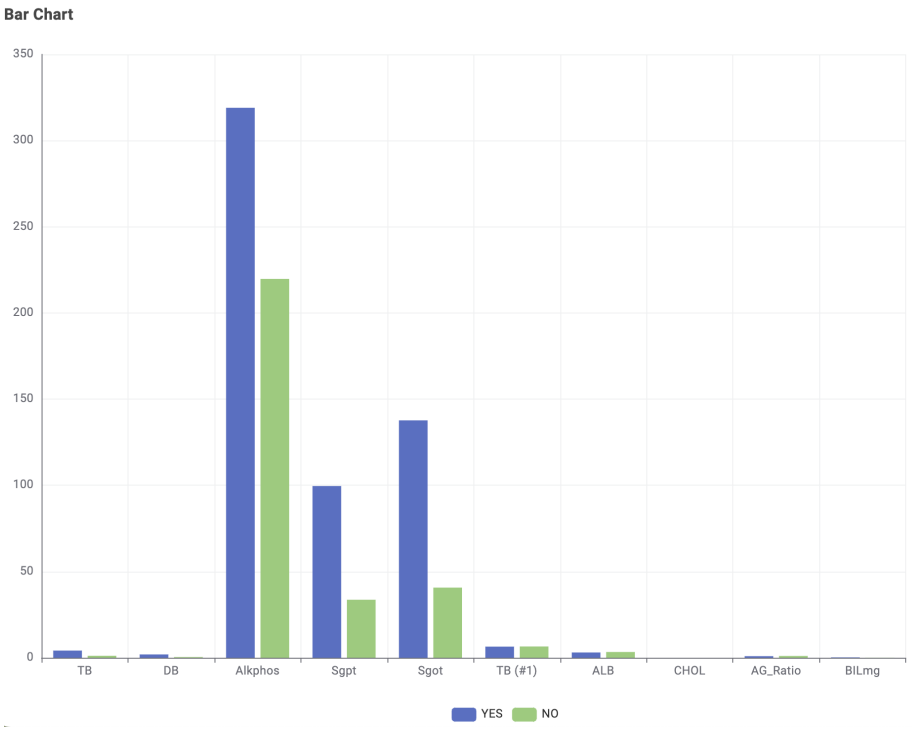


Figura 3.2: Média de todas as análises de doentes e não doentes

TB#1, CHOL: Analisando estes dois atributos e olhando para a figura 3.2, verificamos que o *CHOL* não continha valores associados, portanto consideramos irrelevante, já o *TB#1* decidimos também excluir da previsão, pois já contínhamos um atributo semelhante que é o *TB* e também analisando a figura 3.2 verificamos que não existia grande diferença nas médias dos valores entre o **YES** e **NO** nesse atributo.

3.4 Preparação dos Dados

3.4.1 Análise Prévia ao Tratamento de Dados

Segue-se então, uma análise abrangente do conjunto de dados atribuído, com o objetivo de examinar os valores contidos nele. Esta etapa crítica do processo de análise de dados é fundamental para garantir uma maior precisão no tratamento de dados. Ao examinar minuciosamente cada valor, iremos identificar eventuais inconsistências ou anomalias que necessitem de posteriormente de correção.

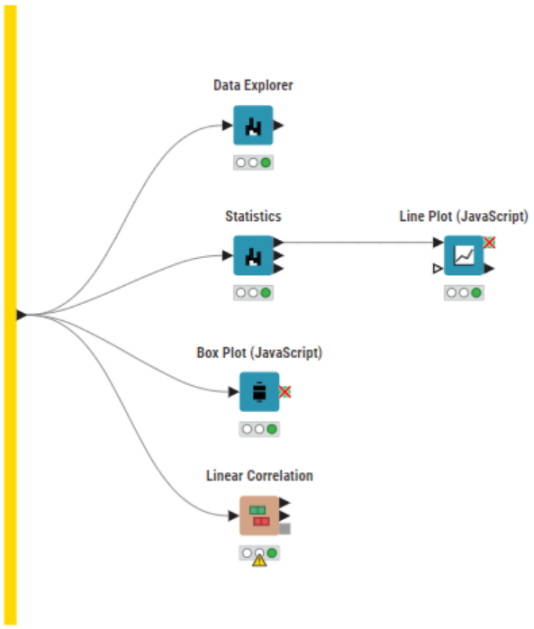


Figura 3.3: Análise Prévia dos Dados

Primeiramente, através do nodo *"Data Explorer"*, identificamos que algumas variáveis estão classificadas como tipo *"String"* quando deveriam ser do tipo *"Number"*. Além disso, observamos a presença de valores em falta em certos atributos, analisando estatísticas como média, desvio padrão, valores máximos e mínimos, como mostrado na figura 3.4.

Numeric | Nominal | Data Preview

Search:

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis	Overall Sum	No. zeros	No. missings	No. NaN
id_code		1	583	292	168.442	28372.667	0	-1.200	170236	0	0	0
Age		4	90	44.746	16.190	262.111	-0.029	-0.560	26087	0	0	0
birth_year		1933	2019	1978.254	16.190	262.111	0.029	-0.560	1153322	0	0	0
birth_month		1	11	5.909	2.900	8.413	0.053	-1.148	3445	0	0	0
Alkphos		63	2110	290.576	242.938	59018.867	3.765	17.753	169406	0	0	0
Sgpt		10	2000	80.714	182.620	33350.194	6.549	50.579	47056	0	0	0
Sgot		10	4929	109.911	288.919	83473.916	10.546	150.920	64078	0	0	0
CHOL		0	0	0	0	0	0	0	0	568	15	0

Showing 1 to 8 of 8 entries

Figura 3.4: *Data Explorer*

Utilizando os nodos *"Box Plot"* e *"Line Plot"*, conseguimos visualizar a presença de valores atípicos em certas variáveis, figura 3.5, indicando a necessidade de atenção no tratamento de dados. Além disso, a análise do *"Skewness"* revelou uma clara inclinação positiva na distribuição dos dados, indicando uma assi-

metria para o lado positivo, sendo demonstrado mais especificamente na figura 3.6.

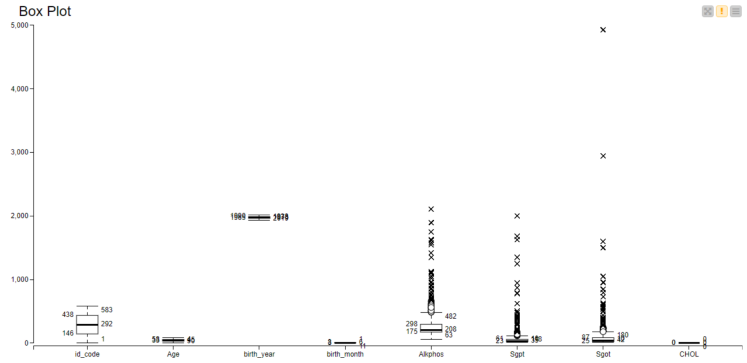


Figura 3.5: *Box Plot*

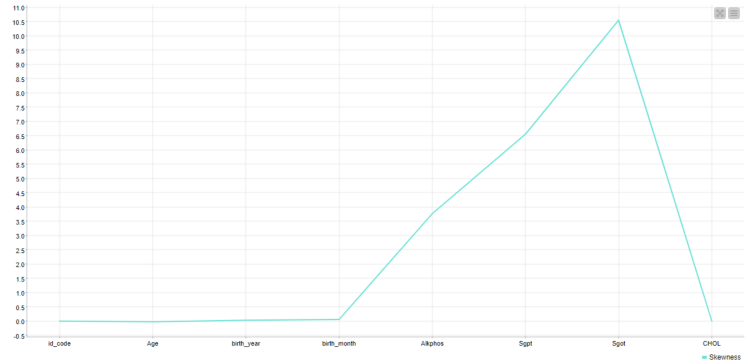


Figura 3.6: *Line Plot*

O uso do nodo *"Linear Correlation"* permitiu investigar a relação direta entre os atributos do *dataset*. Na qual se destacou, uma correlação positiva entre os atributos *Sgot* e *Sgpt*, como demonstrado na figura 3.8

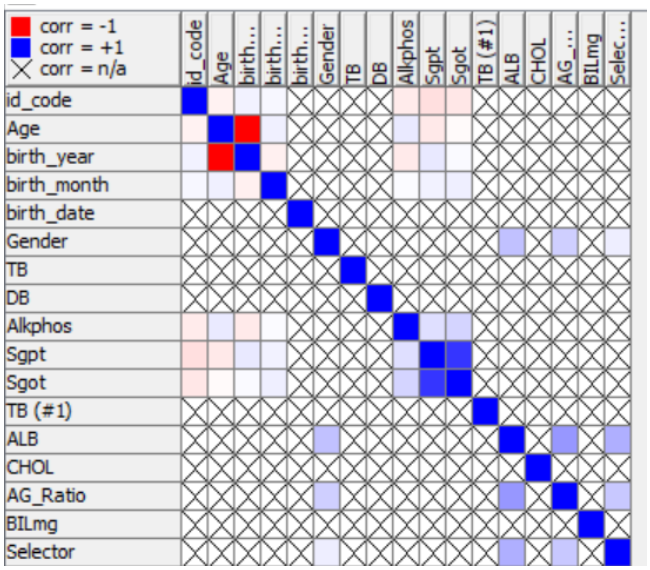


Figura 3.7: *Linear Correlation*

3.4.2 Tratamento Geral dos Dados

Para o processo de tratamento de dados realizado no conjunto de dados em estudo. Identificamos e corrigimos várias questões, como colunas irrelevantes, valores em falta e erros de formatação.

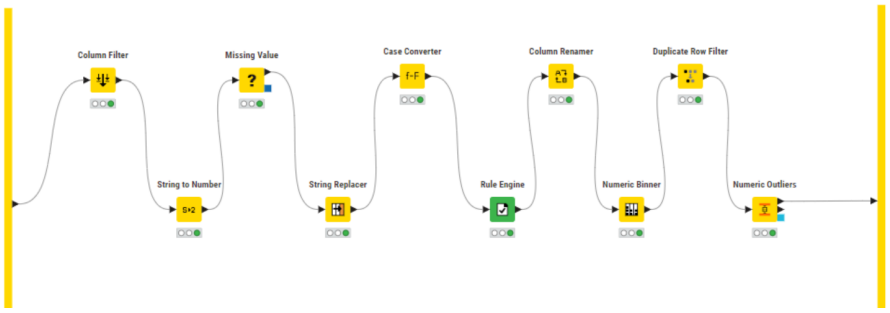


Figura 3.8: Tratamento de Dados

Identificação e Remoção de Colunas

Inicialmente, procedemos à identificação e remoção de colunas com informações irrelevantes. As colunas *"id_code"*, *"birth_year"*, *"birth_month"*, *"birth_date"*, *"TB(#1)"* e *"CHOL"* foram consideradas não significativas para a análise. A coluna "id_code"foi eliminada, uma vez que a informação de identificação já está disponível na coluna *"RowID"*. As colunas relacionadas com a data de nascimento foram consideradas redundantes, sendo que o cálculo da idade foi verificado antes da sua remoção. Adicionalmente, foi investigada a diferença entre as colunas *"TB"* e *"TB(#1)"*.

A coluna *"CHOL"* foi encontrada com todos os seus valores a 0.

Conversão de Dados e Tratamento de Valores em Falta

As colunas *"TB"*, *"DB"*, *"ALB"*, *"ALB_RATIO"* e *"BILmg"* foram convertidas de formato de *string* para numérico através do nodo *"String to Number"*. Foram detetados valores em falta nas colunas *"AG_Ratio"* e *"BILmg"*, os quais foram substituídos pela média dos restantes valores das respetivas colunas, utilizando o nodo *"Missing Value"*.

Normalização de Dados

Foram identificados erros na coluna *Gender*, os quais foram corrigidos utilizando o nodo *"String Replacer"* para uniformizar as identificações. Além disso, foram aplicadas transformações para garantir que os valores na coluna de género estivessem em maiúsculas, através do nodo *"Case Converter"*.

Criação de Novas Variáveis

Para facilitar a análise dos dados, foi criada uma nova coluna utilizando o nodo *"Numeric Binner"* para categorizar as pessoas com base na sua idade em quatro grupos:

- Criança : $]\ -\infty, 13\ [$
- Jovem : $[\ 13\ ,\ 29\ [$
- Adulto : $[\ 29\ ,\ 60\ [$
- Sénior : $[\ 60\ ,\ +\infty\ [$

Remoção de Dados Duplicados e Valores Atípicos

Utilizamos o nodo *"Duplicated Row Filter"* para remover todas as linhas duplicadas e o nodo *"Numeric Outliers"* para lidar com valores atípicos nos conjuntos de dados numéricos, em que esses valores são substituídos pelo valor mais próximo permitido.

3.4.3 Análise Posterior ao Tratamento de Dados

Após a conclusão do tratamento de dados, procedemos a uma análise através dos mesmos métodos utilizados anteriormente ao tratamento geral.
Sendo que, verifica-se que não foram detetados valores em falta nas colunas existentes e ao converter variáveis do tipo *String* para *Number*, expandimos o leque de variáveis de estudo disponíveis para análise. Além disso, todas as variáveis foram normalizadas, garantindo uma representação consistente e comparável dos dados. Além disso, os valores atípicos (*oulayers)* foram identificados e posteriormente normalizados, como resultado na figura 3.10.

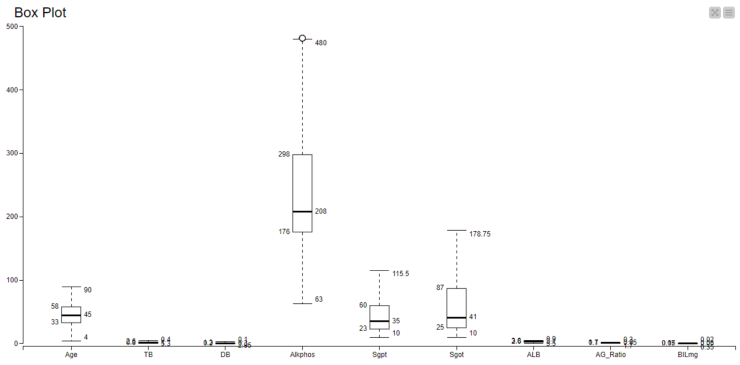


Figura 3.9: *Box Plot*

Ao analisar a assimetria dos dados utilizando o critério *"Skewness"*, notou-se uma redução significativa no desvio em comparação com os dados originais, 3.9. Esta melhoria indica uma distribuição mais equilibrada e simétrica dos dados após o tratamento.

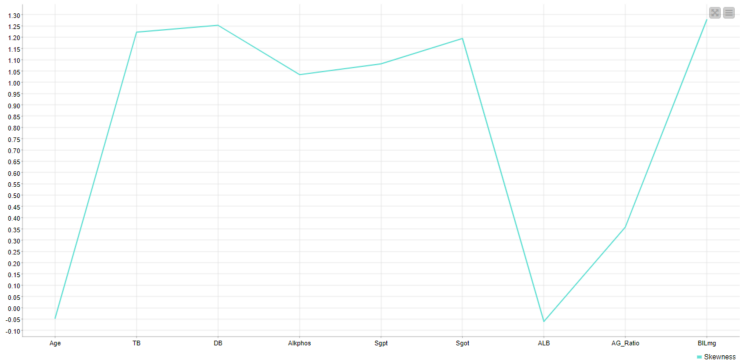


Figura 3.10: *Line Plot*

Uma vez que possuimos mais variáveis, conseguimos observar que existem relações que anteriormente nao conseguiamos verificar, tais como, um correlacao positiva com os pares TB-DB, TB-BILmg, DB-BILmg, e uma correlacao negativa com os pares AG_Ratio-TB, AG_Ratio-DB, AG_Ratio-BILmg e ALB-Age.

Com o aumento do número de variáveis disponíveis após a conversão, conseguimos identificar e analisar novas relações entre os dados. Observamos correlações que não eram evidentes anteriormente. Notavelmente, identificamos correlações positivas entre os pares de variáveis *TB-DB*, *TB-BILmg* e *DB-BILmg*. Além disso, observamos correlações negativas entre os pares de variáveis *AG_Ratio-TB*, *AG_Ratio-DB*, *AG_Ratio-BILmg* e *ALB-Age*, sendo assim apresentado na figura 3.11

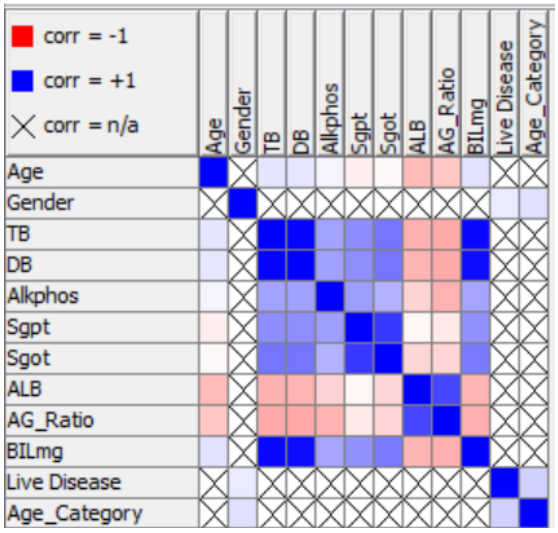


Figura 3.11: *Linear Correlation*

Começamos a análise comparando passo a passo as estatísticas entre homens e mulheres para entender melhor as diferenças. Inicialmente, examinamos o número de homens e mulheres no conjunto de dados para determinar a melhor abordagem de comparação. Dado o número relativamente maior de homens no conjunto de dados, como podemos observar na figura 3.12, optamos por utilizar médias para análise comparativa.

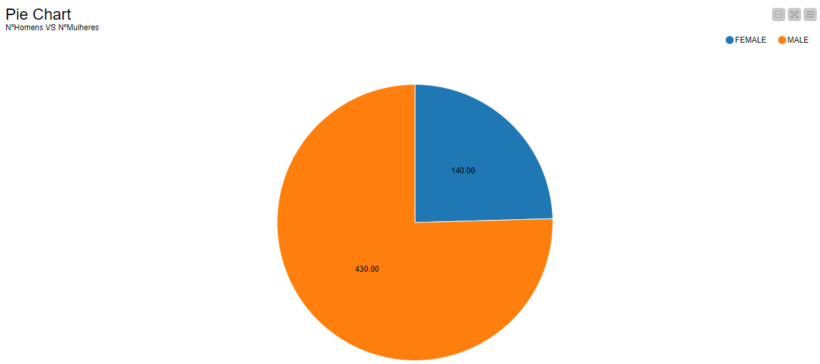


Figura 3.12: *Pie Chart (JavaScript)*

3.5 Modelação

3.5.1 Modelação com Todos os Atributos

Inicialmente, criamos um modelo de previsão que considerava todos os atributos disponíveis, utilizando o mesmo método de normalização e de partição do *dataset* da Tarefa 1. Sendo que os algoritmos aplicados foram os seguintes: *Decision Tree Learner*, *Random Forest Learner* e *Gradient Boosted Trees Learner*.

Sendo que o algoritmo que obteve o melhor desempenho foi o *Gradient Boosting*, através do *"Partitioning"* normalizado, alcançando uma *accuracy* de 87.72%.

3.5.2 Modelação com Seleção dos Melhores Atributos

Para identificar os atributos mais significativos, aplicamos novamente a estratégia de *Feature Selection*. Após a análise detalhada, foram selecionamos os seguintes atributos como os mais relevantes para a previsão dos dados: *Age*, *Gender*, *TB*, *DB*, *Alkphos*, *Sgot*, *ALB*, *AG_Ratio* e *Age_Category*.
Com a utilização desses atributos selecionados, foi alcançada uma melhoria na precisão do modelo, com uma *accuracy* de 88.60%.

3.5.3 Modelação com *Clustering*

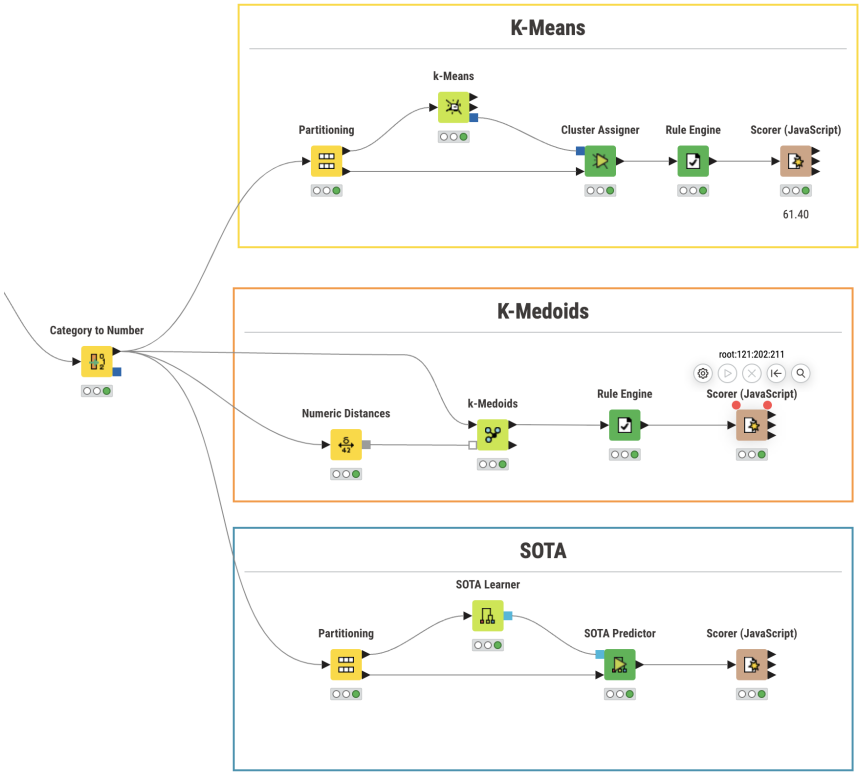


Figura 3.13: Modelos Clustering utilizados

Na modelação com *Clustering* foram utilizados três modelos, sendo eles o *K-Means*, o *K-Medoids* e o *SOTA*, em que em ambos os dois primeiros usamos 11 clusters para a previsão, sendo que no primeiro alcançou-se uma *accuracy* de 67.54%, no outro alcançou-se uma accuracy de 68.42%, já no *SOTA* obteve-se o melhor resultado com uma *accuracy* de 75.44%.

3.5.4 Modelação com Separação entre *MALE* e *FEMALE*

Devido à maior percentagem de homens no *dataset*, optou-se por realizar a previsão separadamente para cada género. Isso permitiu ajustar os modelos de previsão de forma mais específica e potencialmente obter valores mais precisos. Sendo que foi obtido uma *accuracy* de 94.19% para o dos homens e uma *accuracy* de 75% para as mulheres.

Scorer View

Confusion Matrix

	NO (Predicted)	YES (Predicted)	
NO (Actual)	22	3	88.00%
YES (Actual)	2	59	96.72%
	91.67%	95.16%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
94.19%	5.81%	0.857	81	5

(a) Melhor *accuracy* Homens.

Scorer View

Confusion Matrix

	NO (Predicted)	YES (Predicted)	
NO (Actual)	6	4	60.00%
YES (Actual)	3	15	83.33%
	66.67%	78.95%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
75.00%	25.00%	0.443	21	7

(b) Melhor *accuracy* Mulheres.

Figura 3.14: Melhores resultados obtidos entre Homens e Mulheres

3.5.5 Modelação com o Objetivo de Regressão

De modo a converter este problema de classificação para um problema de regressão, alteramos o objetivo da previsão para um valor numérico. Em seguida, aplicamos três algoritmos de regressão: Linear *Regression Learner*, *Simple Regression Tree Learner*, *Gradient Boosted Trees (Regression)*, sendo que os resultados obtidos não se enquadrarem com o que esperávamos, o melhor resultado obtido foi de 0.49 no R^2 no modelo *Gradient Boosted Trees (Regression)*.

3.5.6 Modelação com Redes Neuronais

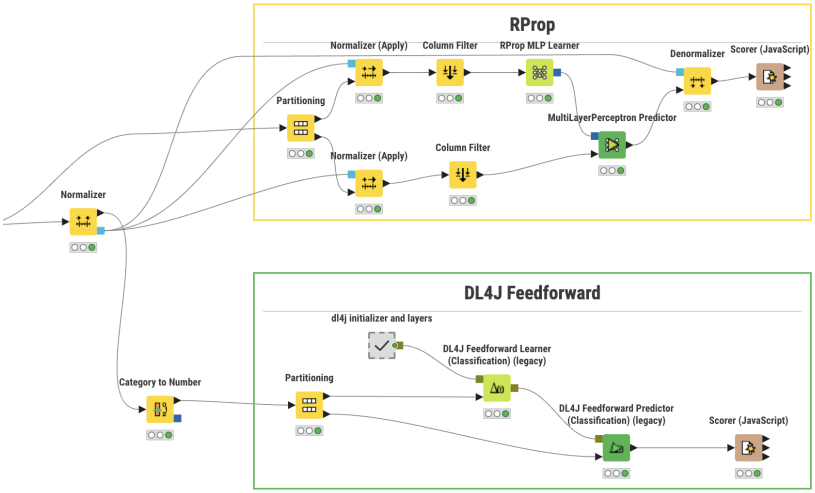


Figura 3.15: Modelos de Redes Neuronais.

Na modelação com Redes Neuronais, antes de inserção dos dados nos modelos, é necessário normalizá-los para melhorar o seu desempenho e também garantir estabilidade no modelo. Sendo que então os modelos utilizados foram *RProp MLP Learner* e o *DL4J Feedforward Learner (Classification)* e com isso obtivemos no primeiro uma *accuracy* de 84.21%, já no outro obteve-se uma *accuracy* de 81.58%.

3.6 Avaliação

Após o estudo que se realizou no conjunto de dados, apesar de algum tratamento realizado, permitiu-nos garantir um bom estudo e previsão do problema.

Na modelação foram utilizados diferentes modelos, visando encontrar o melhor resultado e concluimos que a utilização de algoritmos de árvores de gradiente aumentado (*Gradient Boosted Trees)* com os dados normalizados, levaram-nos a melhores resultados tanto na inclusão das análises todas, mas também como a separação de géneros entre Homens e Mulheres, tendo sido nos Homens o melhor resultado obtido, pois também o *datatset* continha mais homens do que mulheres. No entanto, a utilização da técnica de *Feature Selection*, proporcionou-nos encontrar os melhores atributos para o problema, levando a uma melhoria ligeira em comparação com as análises todas. A seguir, a modelação com bons resultados também foi a das Redes Neuronais, o que também nos leva a conclusão que é uma boa forma de prever o problema. Por outro lado, a modelação por *Clustering*, apesar do melhor resultado não estar muito longe dos outros resultados, levou-nos à conclusão de que não será uma boa forma de prever este problema. Por fim, o pior modelo foi a modelação com objetivo de regressão, o que nos faz concluir que também não é o melhor caminho para obter uma previsão correta.

Capítulo 4

Conclusão

Durante a elaboração deste projeto, aplicaram-se os conceitos aprendidos nas aulas de forma a explorar e tratar os dados com sucesso. Embora a escolha do *dataset*, nao tenha sido a mais acertada, principalmente devido ao seu tamanho, conseguiu-se superar esses desafios com o avançar do projeto.
A exploração dos dados foi fundamental para compreender a natureza dos problemas e identificar os atributos relevantes para as suas respetivas análises. O tratamento dos dados, de uma forma geral, foi bem sucedido.
Apesar das dificuldades apresentadas, os modelos de aprendizagem foram construídos. As técnicas e algoritmos utilizados demonstraram resultados promissores na previsão e análise dos dados. No entanto, reconheceu-se a importância de refletir sobre as escolhas durante a realização do projeto, nomeadamente na seleção de *datasets* de estudo.
Em suma, o processo de exploração e tratamento de dados foi bem alcançado e os modelos de aprendizagem foram construídos de forma satisfatória, proporcionando uma boa análise e previsão dos dados.