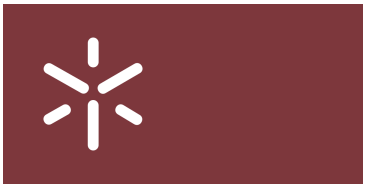


Universidade do Minho
Escola de Engenharia
Licenciatura em Engenharia Informática

Aprendizagem e Decisão Inteligentes

Ano Letivo de 2024/2025

Conceção de Modelos de Aprendizagem e Decisão

Grupo 27

Carlos Eduardo Martins de Sá Fernandes	**A100890**
Tomás Henrique Alves Melo	**A104529**
João Gustavo da Silva Couto Mendes Serrão	**A104444**
Nuno Miguel Barroso Pereira	**A91971**

ADI

Maio, 2025

Índice

Lista de Figuras

1. Introdução

Para abordarmos o processo de realização do trabalho prático de maneira orientada, estruturada e iterativa, recorremos às fases compostas no ciclo de vida do modelo CRISP-DM. Após o **business understanding** do conjunto de dados realizado pelos docentes, o grupo seguiu a lógica proposta pelo modelo CRISP-DM abordado nas aulas da disciplina ao longo do semestre, realizando o **data understanding** ao explorar e analisar os dados de modo a identificar padrões e tendências, **preparação dos dados** ao limpar e transformar os dados para que estes estejam prontos para análise e posterior **modelagem** ao construir e avaliar modelos de mineração de dados para identificar padrões e fazer previsões, e, por fim, **avaliação** e qualidade de desempenho dos diferentes modelos aplicados.

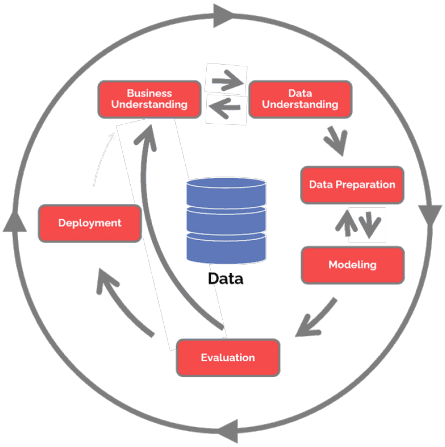


Figura 1: Modelo CRISP-DM

1.1. Estrutura do Relatório

O relatório tem como objetivo mostrar todo o processo realizado ao longo do semestre para a conclusão com sucesso do trabalho prático proposto pelos docentes da disciplina Aprendizagem e Decisão Inteligentes, destacando as principais abordagens e decisões tomadas em ambas as tarefas especificadas no enunciado do trabalho prático.

O processo realizado, de maneira geral, traduz-se na exploração, extração de conhecimento e preparação dos dados, criação de modelos de classificação e regressão, otimização de modelos com variação dos hiperparámetros e análise cuidada e avaliação crítica dos modelos.

2. Tarefa A - Cars Sales

O dataset em análise descreve um conjunto de veículos, contendo múltiplas características técnicas e visuais de cada um, bem como informação identificadora e comercial. O principal objetivo é prever o preço do veículo (primaly_price) com base nas suas propriedades.

2.1. Apresentação do Dataset

Nesta secção, apresentámos as features do dataset que recolhemos, numa tabela, indicando informação relativa a cada uma.

Feature	**Descrição**	**Data Type**
Exterior Color	Cor da carroçaria ou da estrutura externa do veículo	String
Interiol Color	Cor dos materiais e superfícies dentro do veículo	String
Drivetrain	Tipo de tração	String
MPG	Milhas por galão	String
Fuel type	Tipo de combustível usado pelo veículo	String
Transmission	Tipo de transmissão do veículo	String
Engine	Tipo de motor	String
VIN	Número identificador do veículo	String
Stock #	Número de stock	String
Mileage	Número de milhas percorridas do veículo	String
title	Nome do veículo	String
primaly_price	Preço do veículo	Double
currency	Moeda em que se encontra o preço	String
url	URL do veículo	String

2.2. Características do Dataset & Análise dos Dados

2.2.1. Exterior Color

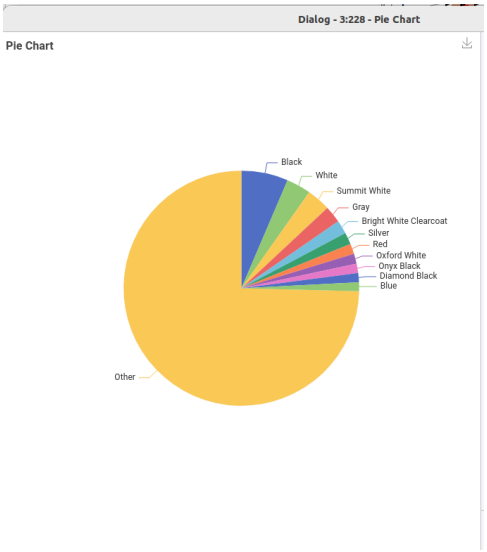


Figura 2: Pie Chart para 'Exterior Color'

Exterior color
No. missings: 0
contains more than 1 000 nominal values
Bottom 20:

Figura 3: Estatisticas apresentadas

O atributo **Exterior Color** contém **0 missing values**, e uma grande quantidade de entradas únicas, de facto acima de 1000, sendo **1120** o valor. Ainda, apresenta uma distribuição elevada e equilibrada com a maioria das entradas no **Other**, porém, verifica-se uma preferência por **Preto** e **Branco** e variantes deles.

2.2.2. Interior Color

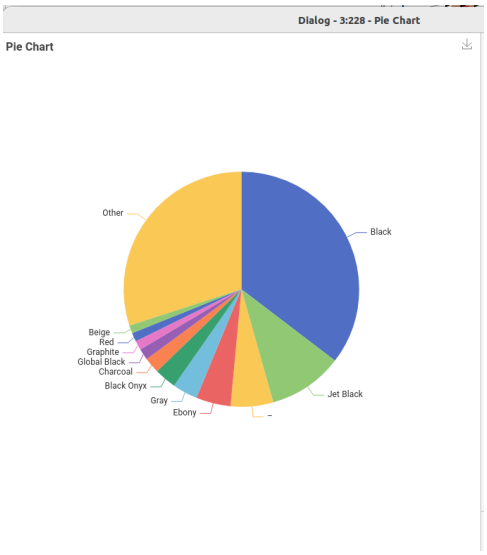


Figura 4: Pie Chart para 'Interior Color'

Interior color
No. missings: 0
Top 20: Black : 2214 Jet Black : 641 – : 366 Ebony : 297 Gray : 217 Black Onyx : 188 Charcoal : 136 Global Black : 98 Graphite : 79 Red : 76 Beige : 68 Titan Black : 48 Medium Dark Slate : 47 Cognac : 36 Tan : 28 Black / Graphite : 25 Medium Earth Gray : 25 Parchment : 25 Light Gray : 23 Cement : 22
Bottom 20: Ultramarine Blue : 1 Light Mountain Brown/Mountain Brown : 1 Black Leatherette : 1 Acorn / Ebony : 1 Cognac W/Contrast Stitch : 1 Palomino Brown / Steel Gray : 1 Dark Rose : 1 Onyx : 1 Charcoal Black w/King Ranch Red : 1 Zinc : 1 Black semi-aniline leather and Black Open-Pore Wood trim : 1 Black W/Medium Dark Slate : 1 Zagora Beige : 1 GRAY : 1 Espresso / Nimbus : 1 Ski Gray : 1 Platinum : 1 BLC : 1 Rosso Corallo : 1 Med Slate Gray : 1

Figura 5: Estatisticas apresentadas

O atributo **Interior Color** apresenta **0 missing values**, e uma quantidade de entradas únicas considerável, sendo **551** o valor. Também, possui uma distribuição elevada, mas não equilibrada com uma grande preferência por **Preto**, porém, a percentagem em **Other** é elevada.

2.2.3. Drivetrain

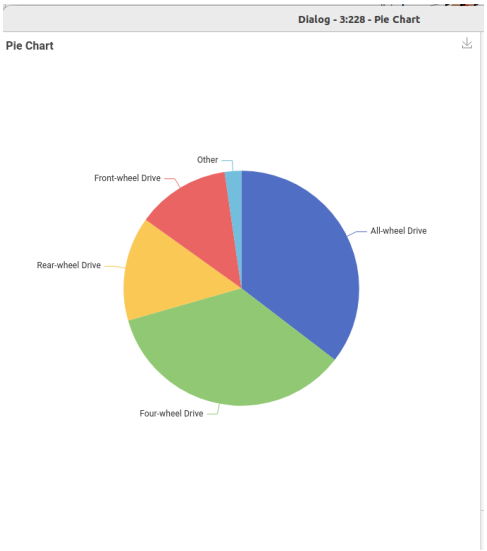


Figura 6: Pie Chart para 'Drivetrain'

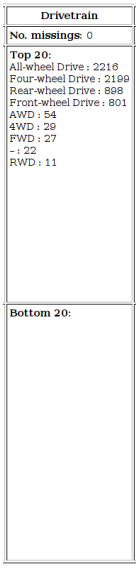


Figura 7: Estatisticas apresentadas

O atributo **Drivetrain** apresenta também **0 missing values** diretos, contudo tem a entrada **'-'** que representa a ausência de dados, e uma quantidade de entradas únicas baixa, sendo esta **9**. Para além disso, apresenta uma distribuição considerada equilibrada, com uma preferência por **All-wheel Drive** e **Four--wheel Drive**.

2.2.4. MPG

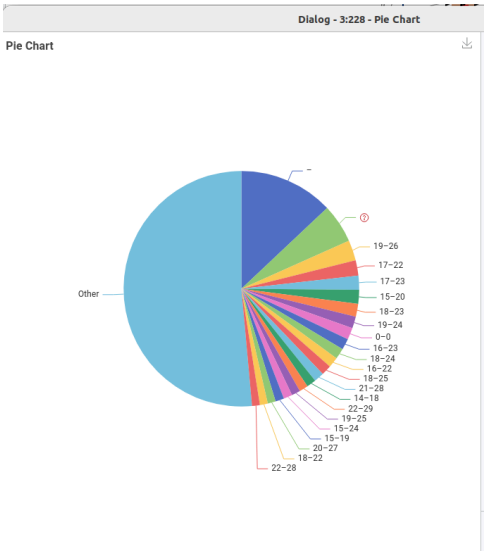


Figura 8: Pie Chart para 'MPG'

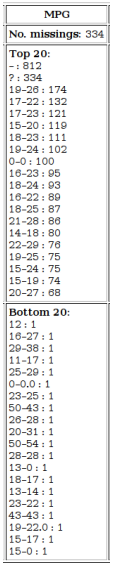


Figura 9: Estatisticas apresentadas

O atributo **MPG** tem **334 missing values** diretos, contudo tem a entrada **'-'** que representa ausência de dados, isto é, valor em falta, e uma quantidade de entradas únicas considerável, sendo esta **300**. Também, apresenta uma distribuição elevada e equilibrada com uma pequena preferência por **19-26** com a percentagem em **Other** elevada, porém, as entrada maiores são as relacionadas a ausência de dados.

2.2.5. Fuel Type

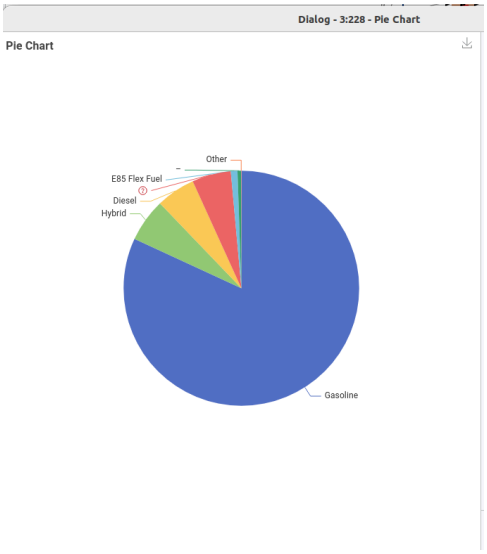


Figura 10: Pie Chart para 'Fuel Type'

Fuel type

No. missings: 334

Top 20:
Gasoline : 5124
Hybrid : 373
Diesel : 336
? : 334
E85 Flex Fuel : 57
– : 30
Electric Fuel System : 1
Gasoline Fuel : 1
MHEV (mild hybrid electric vehicle) : 1

Bottom 20:

Figura 11: Estatisticas apresentadas

O atributo **Fuel Type** contém **334 missing values** diretos, contudo tem a entrada **'-'** que representa ausência de dados, e uma quantidade de entradas únicas baixa, sendo esta **9**. Também, apresenta uma distribuição baixa e não equilibrada com uma enorme preferência por **Gasoline**.

2.2.6. Transmission

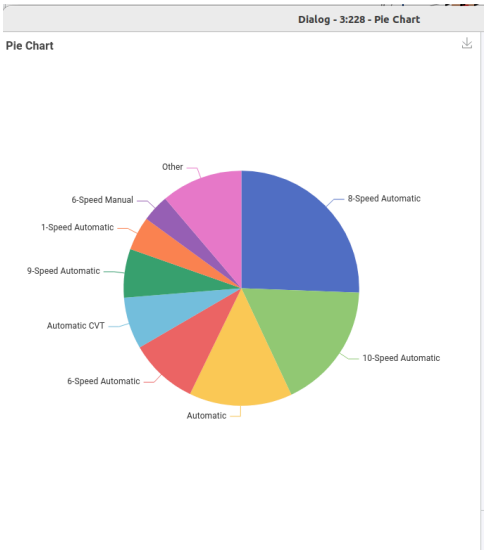


Figura 12: Pie Chart para 'Transmission'

Transmission

No. missings: 0

Top 20:
8-Speed Automatic : 1601
10-Speed Automatic : 1091
Automatic : 886
6-Speed Automatic : 589
Automatic CVT : 445
9-Speed Automatic : 419
1-Speed Automatic : 291
6-Speed Manual : 234
5-Speed Automatic : 127
7-Speed Automatic with Auto-Shift : 120
8-Speed Automatic with Auto-Shift : 93
7-Speed Automatic : 88
4-Speed Automatic : 34
– : 34
Manual : 26
6-Speed Automatic with Auto-Shift : 18
5-Speed Manual : 15
7-Speed Manual : 13
Variable : 13
Automatic with Tiptronic : 12

Bottom 20:
8-Speed A/T : 1
5-Speed Automatic with Overdrive : 1
continuously variable (cvt) : 1
6-Speed Shiftable Automatic : 1
6-Speed Automatic with Sequential Shift ECT : 1
1AT : 1
7-Speed DSGA? Automatic w/ 4MO : 1
Automatic 7G-TRONIC : 1
7-Speed Automatic S tronic : 1
9-speed automatic : 1
8-Spd Auto 8HP75 Trans : 1
6-Speed Automatic Electronic with Overdrive : 1
8-Speed Manual : 1
SHIFTRONIC : 1
6-Spd Aisin F21-250 PHEV Auto Trans : 1
6 Speed Elect Controlled Auto Transmission w/Inte. : 1
ZF 8-Speed Automatic : 1
A : 1
10-Speed Manual : 1
1-Speed CVT with Overdrive : 1

Figura 13: Estatisticas apresentadas

O atributo **Transmission** contém **0 missing values** diretos, contudo tem a entrada **'-'** que representa ausência de dados, e uma quantidade de entradas únicas média, sendo esta **74**. Também, apresenta uma distribuição elevada e pouco equilibrada com uma grande preferência por **8-speed Automatic** e **10-Speed Automatic**, com a percentagem em **Other** baixa.

2.2.7. Engine

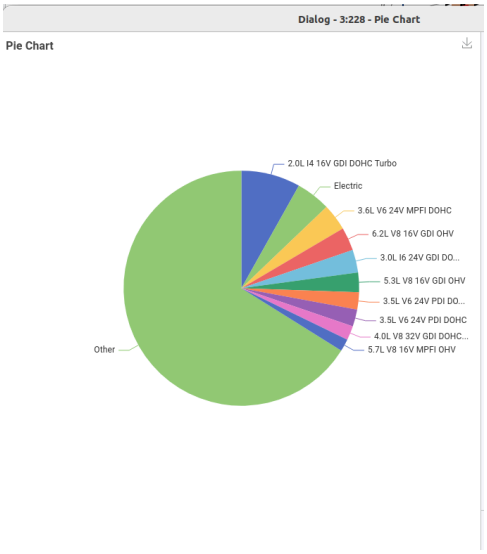


Figura 14: Pie Chart para 'Engine'

Engine
No. missings: 0
Top 20: 2.0L I4 16V GDI DOHC Turbo : 508 Electric : 296 3.6L V6 24V MPFI DOHC : 230 6.2L V8 16V GDI OHV : 199 3.0L I6 24V GDI DOHC Turbo : 197 5.3L V8 16V GDI OHV : 168 3.5L V6 24V PDI DOHC Twin Turbo : 149 3.5L V6 24V PDI DOHC : 146 4.0L V8 32V GDI DOHC Twin Turbo : 122 5.7L V8 16V MPFI OHV : 106 5.0L V8 32V PDI DOHC : 96 3.0L V6 24V GDI DOHC Twin Turbo : 95 3.6L V6 24V GDI DOHC : 90 1.5L I4 16V GDI DOHC Turbo : 88 2.5L I4 16V PDI DOHC Hybrid : 82 3.8L V8 24V GDI DOHC : 82 3.4L V6 24V PDI DOHC Twin Turbo : 82 3.5L V6 24V GDI SOHC : 81 6.4L V8 16V MPFI OHV : 77 4.0L V6 24V MPFI DOHC : 76
Bottom 20: SKYACTIV-G 2.5L I-4 gasoline direct injection, DOHC, variable va : 1 3.6L V-6 DOHC, variable valve control, engine with 283HP : 1 4.4L V-8 gasoline direct injection, DOHC, Double VANOS variable : 1 2.5L I-4 port/direct injection, DOHC, variable valve control, pr : 1 1.8L I-4 i-VTEC variable valve control, engine with 143HP : 1 3.5L V-6 DOHC, VVT-iW variable valve control, regular unleaded. : 1 3.8L V6 : 1 Triton 5.4L V-8 variable valve control, engine with 310HP : 1 1.4L I4 16V MPFI SOHC : 1 EcoBoost 2.7L V-6 gasoline direct injection, DOHC, Ti-VCT variab : 1 5L V-8 engine with 302HP : 1 3.3L V6 24V PDI DOHC : 1 3.5L V-6 gasoline direct injection, DOHC, i-VTEC (w/VTC) variabl : 1 2.7L I-4 DOHC, VVT-iW/VVT-i variable valve control, regular unle : 1 2.5L H-4 DOHC, variable valve control, intercooled turbo, premiu : 1 Regular Unleaded H-4 2.0 L/122 : 1 5.7L V8 : 1 5.7L V8 32V DOHC Flexible Fuel : 1 Intercooled Turbo Regular Unleaded I-4 1.5 L/91 : 1 Cummins 6.7L I-6 diesel direct injection, intercooled turbo, die : 1

Figura 15: Estatisticas apresentadas

O atributo **Engine** contém **0 missing values**, e uma quantidade de entradas únicas considerável, sendo esta **552**. Também, apresenta uma distribuição elevada e equilibrada com uma preferência por **2.0L 14 16V GDI DOHC Turbo** e uma percentagem em **Other** elevada.

2.2.8. Mileage

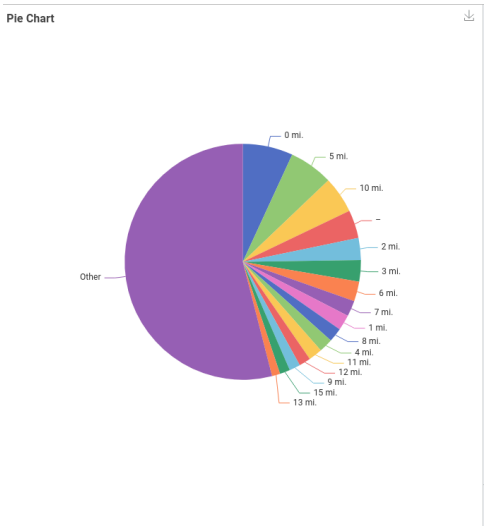


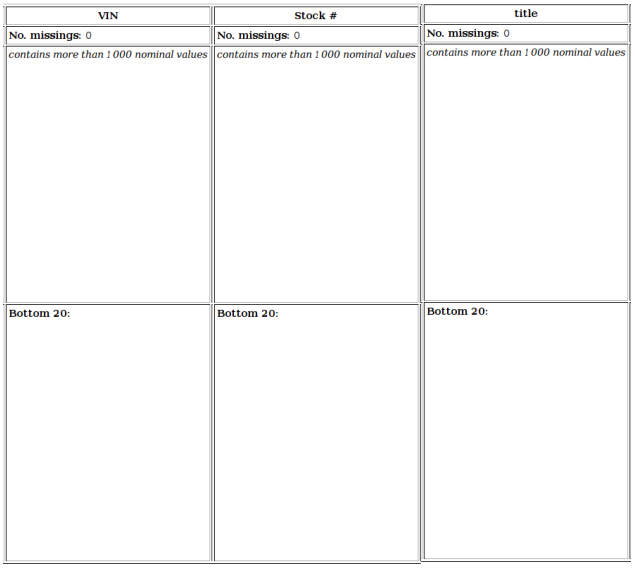
Figura 16: Pie Chart para 'Mileage'

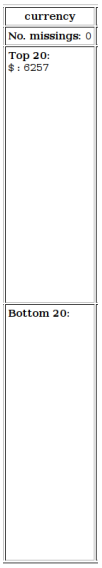
Mileage
No. missings: 0
contains more than 1 000 nominal values
Bottom 20:

Figura 17: Estatisticas apresentadas

O atributo **Mileage** contém **0 missing values**, e uma grande quantidade de entradas únicas, de facto acima de 1000, sendo este **2938**. Também, apresenta uma distribuição elevada e equilibrada com a maioria das entradas no **Other**, porém, verifica-se uma preferência por **0 mi.** e **5 mi.**.

2.2.9. Restantes





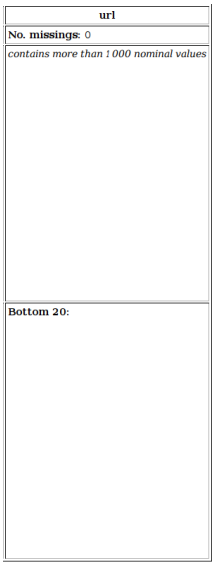


Figura 18: VN Figura 19: Stock Figura 20: Title Figura 21: Currency Figura 22: Url

VN: Contém **0 missing values** e tem uma distribuição 1 para 1 de entradas.
Stock #: Contém **0 missing values** e tem uma distribuição proxima de 1 para 1 de entradas com **6240**.
title: Contém **0 missing values** e tem uma distribuição elevada com **3343** entradas.
currency: Contém **0 missing values** e tem uma distribuição todos para 1 de entradas, com só uma entrada.
url: Contém **0 missing values** e tem uma distribuição 1 para 1 de entradas.

2.2.10. Primaly Price - Target

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis
primaly_price	☐	1999	551986	56380.715	40863.459	1669822267.447	4.219	30.807

Figura 23: Primaly Price

O atributo **Primaly Price** contém **0 missing values**, e uma grande quantidade de entradas únicas, de facto acima de 1000, sendo este **5227**. Também, apresenta uma distribuição elevada e equilibrada com a maioria das entradas com um valor diferente. Por último, é este atributo que será o **Target**, e que tentaremos prever num problema de **Regressão**.

2.2.11. Relação entre Atributos

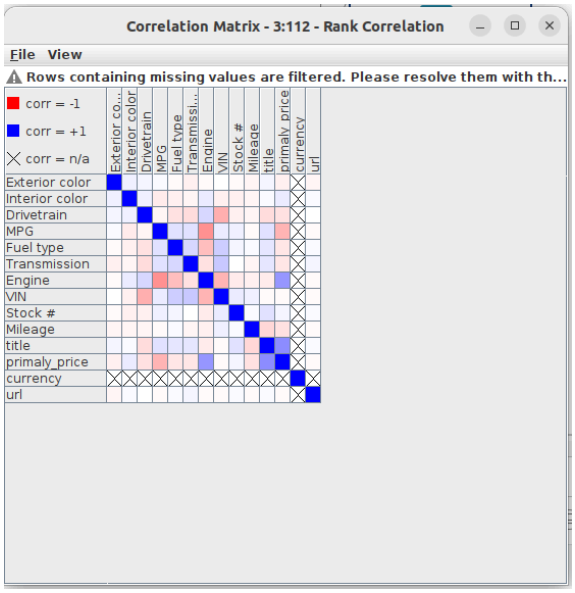


Figura 24: Rank Correlation

A imagem apresenta uma matriz de correlação entre diferentes atributos. A matriz utiliza a **rank correlation**, onde a cor **azul** indica uma correlação **positiva (+1)**, a **vermelha** uma correlação **negativa (-1)**, e os campos com **"X"** representam correlações **indisponíveis**.

Observa-se uma **forte correlação positiva** entre os seguintes pares de variáveis:

- **Exterior color** e **Interior color**
- **Mileage** e **primary_price**
- **title** e **url**

Essas correlações sugerem que essas variáveis tendem a **variar juntas de forma proporcional**.

Além disso, há **correlações negativas moderadas** entre variáveis como **Drivetrain, Fuel type, e Transmission**, o que pode indicar padrões distintos entre diferentes tipos de veículos (por exemplo, carros automáticos a gasolina versus manuais a diesel).

Vale destacar **currency**, não apresentam **correlações válidas** com outras colunas, provavelmente por só ter um valor.

2.3. Exploração, Preparação & Tratamento dos Dados

Nesta secção, destacámos as partes de exploração e tratamento de dados mais relevantes realizadas.

Inicialmente, após analisar nodos de visualização de dados, removemos colunas com muitos valores únicos que não agregavam informação útil para o progresso do workflow, por exemplo removendo a coluna **VIN**, com um valor único para cada linha, **Exterior Color**, **Interior Color** e **Stock #** que possuem imensos valores únicos e a coluna **currency** que possui o mesmo valor em todas as linhas do dataset.

Reparámos em valores no dataset marcados por '–' e para torná-los como missing values recorremos ao nodo String Manipulation (Multi Column).

```
STRING MANIPULATION (MULTI COLUMN) NODE

regexReplace($$CURRENTCOLUMN$$, "^–$", null)
```

Recorremos ainda ao nodo String Manipulation em diversos momentos da preparação e tratamento de dados, tal como visto de seguida:

Figura 25: Tratamento simples de features com String Manipulation

Aqui, recorremos à função 'replace' para padronizar os valores atribuídos a certas colunas e ainda uma expressão regular para extrair o valor da feature **Mileage**.

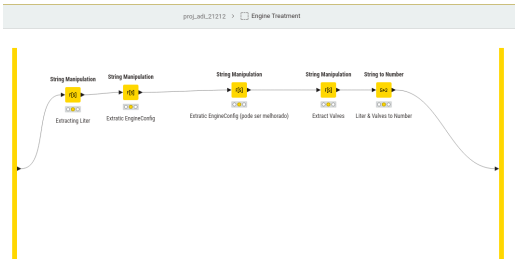


Figura 26: Metanodo Engine Treatment - Expressões regulares para extrair dados

Para a feature **MPG** foi necessário um tratamento extenso, ao termos de separar entre o **valor mínimo de MPG** e **valor máximo de MPG**, garantir que o valor mínimo de MPG era menor que o valor máximo de MPG e vice-versa e ainda transformar esses valores no tipo de dados inteiro. Ainda dentro do tratamento desta feature, se a **Engine** for 'Eletric', então o valor de MPG (Miles per Gallon) deve ser 0.

```
RULE ENGINE NODE

$EngineConfig$ MATCHES ".*Electric.*" => 0
TRUE => $max_swapped$
```

Para além disso, o grupo notou que estes valores poderiam ser calculados com precisão uma vez que veículos com as mesmas características, teriam os mesmos valores para esta coluna. Desta forma, calculamos os valores faltantes (missing values) com a seguinte sequência de nodos:

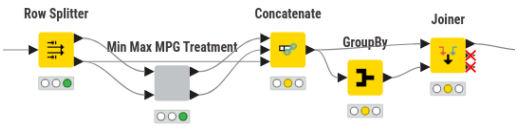


Figura 27: Se Engine for 'Eletric', então MPG = 0

Por análise dos dados, verificámos que, de maneira geral, veículos mais antigos percorreram um número de milhas maior do que veículos mais recentes. Desta forma, atribuímos valores para os missing values desta feature com base no valor de **Mileage** médio para o ano do respetivo veículo com valor em falta nessa coluna.

```
RULE ENGINE NODE

MISSING $Mileage$ => $Mean(Mileage)$
TRUE => $Mileage$
```

A nosso ver, esta é a melhor forma para estimar os valores em falta para esta coluna.

Por fim, após termos tratado dos missing values cujos valores poderiam ser conhecidos com o devido tratamento, tratámos dos restantes de forma geral, aplicando a média para o **min MPG**, **max MPG** e **Valves** e só no fim, remover as linhas com missing values. Ainda, tratámos de remover possíveis outliers dos dados, verificando que melhorava os resultados obtidos no momento da análise dos modelos desenvolvidos.

2.4. Decisões Inteligentes na Preparação & Tratamento dos Dados

- Primeiro foi feito um tratamento específico para os valores em falta atribuindo valores que pudessem ser de facto calculados e só no fim atribuímos os restantes como o valor médio ou remoção da linha.
- Análise da correlação dos dados em vários momentos da construção do workflow.
- Padronizar os dados no início do tratamento dos dados.
- Cuidados a nível lógico no tratamento de dados (valor de **min MPG** deve ser menor do que o valor de **max MPG** e vice-versa, veículos elétricos não devem ter gasolina como **Fuel Type** e devem ter o valor de **MPG** como 0).
- Partição dos dados fornecidos pela feature **Engine** para obtermos mais informação relevante dos veículos que funcionasse como "informação extra" para o melhoramento dos modelos preditivos.

2.5. Data Visualization

2.5.1. Pie Chart - Fuel Type

Este gráfico apresenta a distribuição dos tipos de combustível no dataset. Podemos observar que existe um predominância clara de veículos a gasolina, com uma percentagem significativamente superior (82%) as restantes categorias, o que pode levar a um desequilíbrio no modelo, sendo um ponto a se ter em conta.

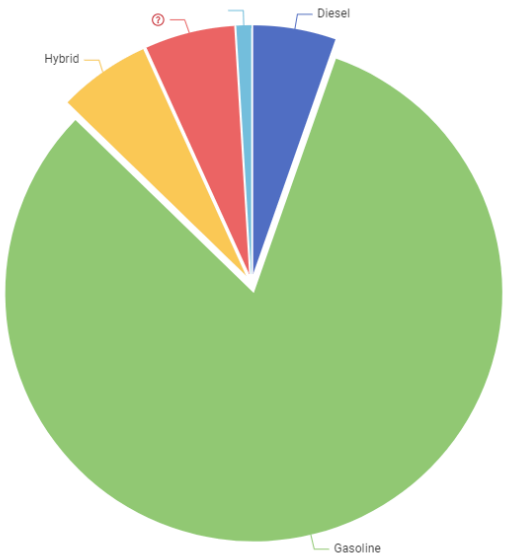


Figura 28: Pie Char - Fuel Type

2.5.2. Bar Chart - Transmission

Através deste gráfico de barras podemos verificar que existe uma grande predominância das transmissões automáticas, especialmente das 8 e 10 velocidades. As restantes têm uma presença significativamente menor, o que poderá nos levar a agrupar todas, de forma a reduzir o numero de categorias.

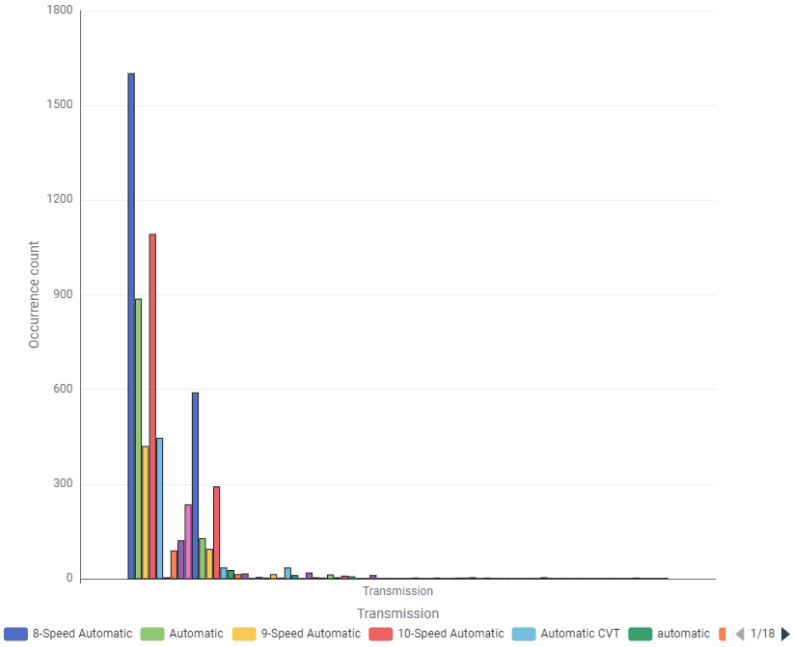


Figura 29: Bar Chart - Transmission

2.5.3. Histograma - Mileage

Como podemos ver neste histograma, que mostra a distribuição da variável Mileage, a maioria dos veículos tem uma quilometragem baixa, o que pode indicar que são carros mais recentes ou poucos usados. Este gráfico pode nos ser útil para deteção de outliers. Conseguimos também visualizar a existência de missing values na coluna de mileage.

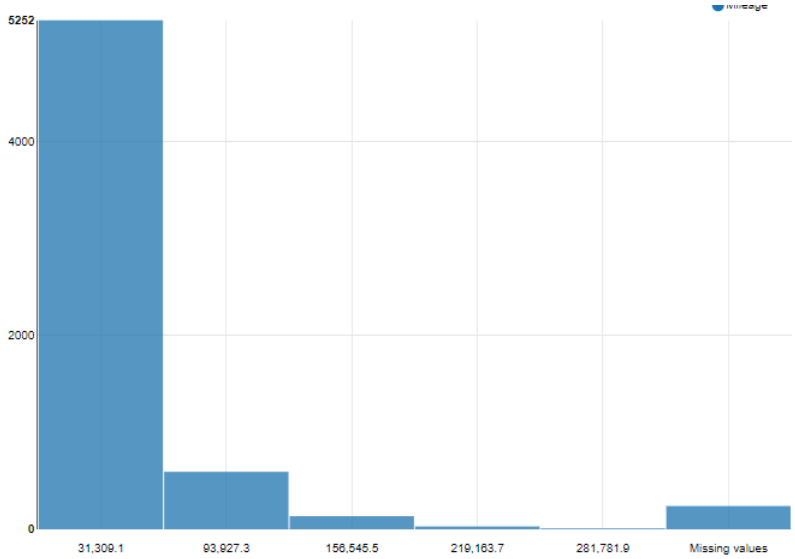


Figura 30: Histogram - Mileage

2.5.4. Box Plot - Primaly Price

O boxplot desta variável permite nos visualizar a dispersão dos preços dos veículos e como podemos observar, existe uma grande quantidade de outlier no topo, correspondentes a veículos com preços muito acima da média. Isto indica-nos que existe uma distribuição assimétrica, com alguns veículos de luxo ou de valor exageradamente elevados.

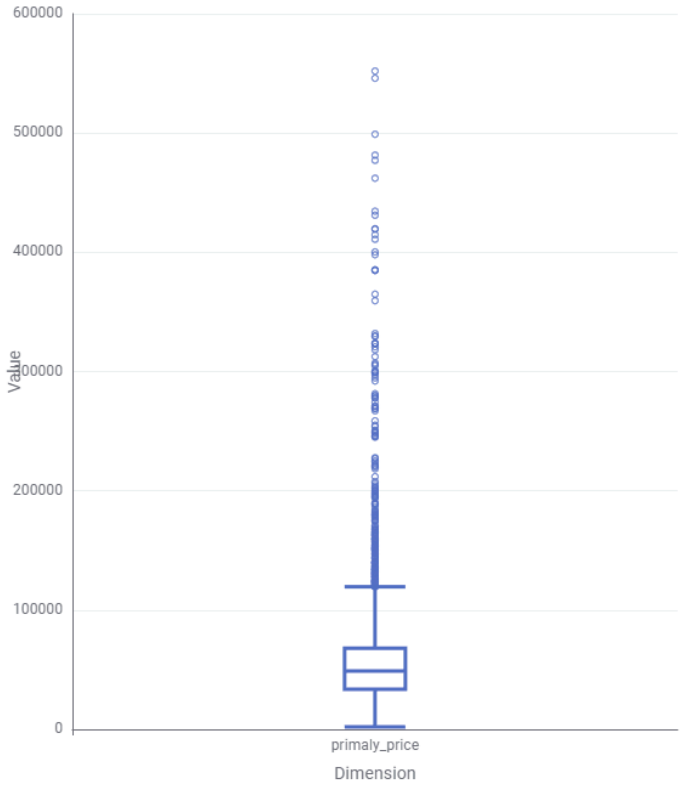


Figura 31: BoxPlot - Primaly_price

2.5.5. Histograma - Primaly Price

Este histograma mostra a distribuição dos preços dos veículos após todos os tratamentos feitos, menos o de remoção e tratamento de outliers, já que estes necessitam de ser tratados ou poderão afetar os modelos de regressão. Podemos observar que ainda existe uma grande dispersão, com muitos carros concentrados na faixa de preços mais baixos e alguns veículos com valores muito altos.

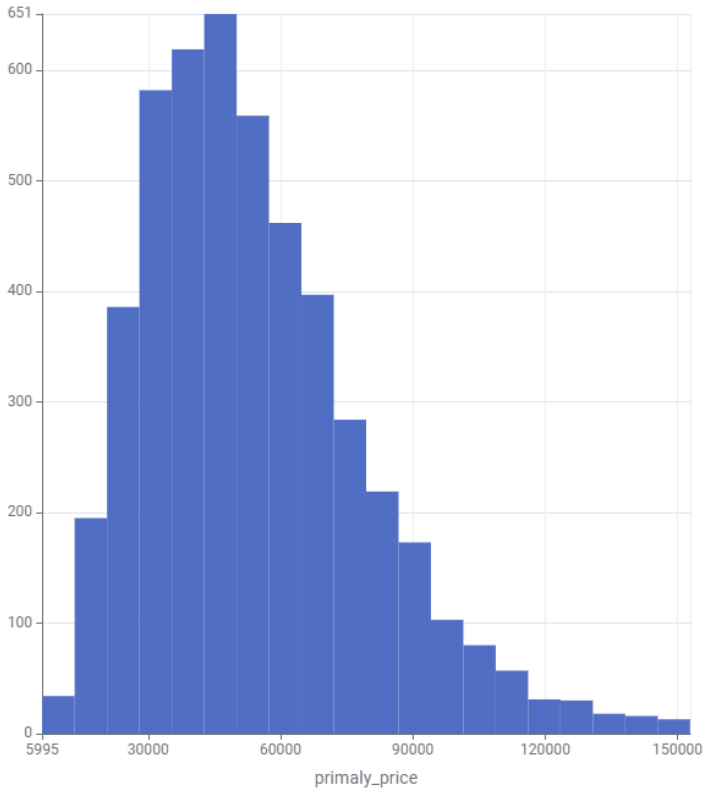


Figura 32: Histogram - Primaly_price

2.6. Modelação - Apresentação dos Modelos Desenvolvidos

Aqui, apresentámos em forma de tabela os vários testes que fizemos a nível dos modelos preditivos, variando os hiperparámetros de cada um, de modo a obter resultados mais positivos em cada um dos modelos desenvolvidos.

2.6.1. Linear Regression

Tentativa	Nº Validations	Type of Sampling	Offset	R^2
0	10	Stratified	UnChecked	0,646
1	10	Stratified	10	0,645
2	10	Stratified	100	0,645
3	10	Stratified	1000	0,645
4	10	Random	UnChecked	0,646
5	100	Stratified	UnChecked	0,646

2.6.2. Regression Tree

Tentativa	Nº Validations	Type of Sampling	Tree Depth	Split Node Size	Node Size	R^2
0	10	Stratified	UnChecked	UnChecked	UnChecked	0,709
1	10	Stratified	10	UnChecked	UnChecked	0,709
2	10	Stratified	100	UnChecked	UnChecked	0,709
3	10	Stratified	1000	UnChecked	UnChecked	0,709

Tentativa	Nº Validations	Type of Sampling	Tree Depth	Split Node Size	Node Size	R^2
4	10	Stratified	UnChecked	10	UnChecked	0,749
5	10	Stratified	UnChecked	20	UnChecked	0,758
6	10	Stratified	UnChecked	30	UnChecked	0,762
7	10	Stratified	UnChecked	40	UnChecked	0,757
8	10	Stratified	UnChecked	30	15	0,751
9	10	Stratified	UnChecked	30	7	0,761
10	10	Stratified	UnChecked	30	3	0,763
11	100	Stratified	UnChecked	30	3	0,779
12	100	Random	UnChecked	30	3	0,779

2.6.3. Random Forest

Tentativa	Nº Validations	Type of Sampling	Tree Depth	Node Size	Nº Models	R^2
0	10	Stratified	UnChecked	10	1000	0,791
1	10	Stratified	10	10	1000	0,791
2	10	Stratified	100	10	1000	0,791
4	10	Stratified	UnChecked	100'	1000	0,642
5	10	Stratified	UnChecked	10	2000	0,791
6	100	Stratified	UnChecked	10	1000	0,796
7	100	Random	UnChecked	10	1000	0,796

2.6.4. Neural Network

Tentativa	Nº Validations	Type of Sampling	Nº Iterations	Nº Hidden Layers	Nº Hidden Neurons	R^2
0	10	Stratified	100	1	10	0,563
1	10	Stratified	300	1	10	0,628
2	10	Stratified	500	1	10	0,644
4	10	Stratified	1000	1'	10	0,668
5	10	Stratified	1000	10	10	0,612
6	10	Stratified	1000	2	10	0,710
7	10	Stratified	1000	3	10	0,708
8	10	Stratified	1000	2	20	0,725
9	100	Stratified	1000	2	20	0,723

2.6.5. Gradient Boosted Tree

Tentativa	Nº Validations	Type of Sampling	Tree Depth	Nº Models	Learning Rate	R^2
0	10	Stratified	4	100	0,1	0,807
1	10	Stratified	10	100	0,1	0,825
2	10	Stratified	20	100	0,1	0,757
3	10	Stratified	10	1000	0,1	0,810
4	10	Stratified	10	200	0,1	0,818

Tentativa	Nº Validations	Type of Sampling	Tree Depth	Nº Models	Learning Rate	R^2
5	10	Stratified	10	100	0,2	0,814
6	100	Stratified	10	100	0,1	0,835

2.7. TOP 5 Modelos - Variação dos hiperparámetros

Modelo	Tentativa	R^2
Gradient Boosted Tree	6	0,835
Random Forest	6	0,796
Regression Tree	11	0,779
Neural Networks	8	0,725
Linear Regression	0	0,646

2.8. Análise Crítica dos Resultados Obtidos - Avaliação

Os modelos foram analisados utilizando R^2 como principal métrica, testando Regressão Linear, Regression Tree, Random Forest, Redes Neuronais e Gradient Boosted Tree. A Regressão Linear apresentou R^2 constante (0,646), indicando inadequação para o problema. A Regression Tree melhorou de 0,709 para 0,779 com Split Node Size igual a 30 e tamanho de nó 3 (tentativa 11), mas o ganho é limitado. O Random Forest alcançou 0,796 (tentativa 6) com 100 validações, mostrando robustez. O Gradient Boosted Tree (R^2 = 0,835) e o Random Forest superaram Redes Neuronais (0,725), sugerindo que relações não lineares são dominantes.

3. Tarefa B - Healthcare, Dataset Grupos Ímpar

Este dataset representa registos clínicos de pacientes admitidos em unidades de saúde. Contém dados demográficos, informações clínicas e detalhes administrativos relativos a cada admissão hospitalar. O objetivo é prever o resultado de um exame médico com base nas demais características disponíveis.

3.1. Apresentação do Dataset

Nesta secção, apresentámos as features do dataset Healthcare, numa tabela, indicando informação relativa de cada uma.

Feature	**Descrição**	**Data Type**
Name	Nome do paciente associado ao registo clínico	String
Age	Idade do paciente à data da admissão	Integer
Gender	Género do paciente (Male ou Female)	String
Blood Type	Tipo de sangue do paciente (e.g., A+, O-, etc.)	String
Medical Condition	Principal condição médica ou diagnóstico associado ao paciente	String
Date of Admission	Data na qual o paciente foi admitido no estabelecimento de saúde	Local Date
Doctor	Nome do doutor responsável pelo paciente durante a admissão	String
Hospital	Nome do estabelecimento de saúde onde o paciente foi admitido	String
Insurance Provider	Seguradora do cliente, e.g. Aetna, Blue Cross, Cigna, Medicare.	String
Billing Amount	Montante cobrado pelos serviços de saúde prestados ao paciente	String
Room Number	O número do quarto onde o paciente ficou alojado durante a sua admissão	Integer
Admission Type	Tipo de admissão, que pode ser: Emergency, Elective ou Urgent	String
Discharge Date	A data em que o paciente teve alta do estabelecimento de saúde	String
Medication	Um medicamento prescrito ou administrado ao paciente durante a sua admissão	String
Test Results	Descreve os resultados de um exame médico realizado durante a admissão do paciente. Os valores possíveis incluem Normal, Anormal ou Inconclusivo	String

3.2. Características do Dataset & Análise dos Dados

3.2.1. Gender

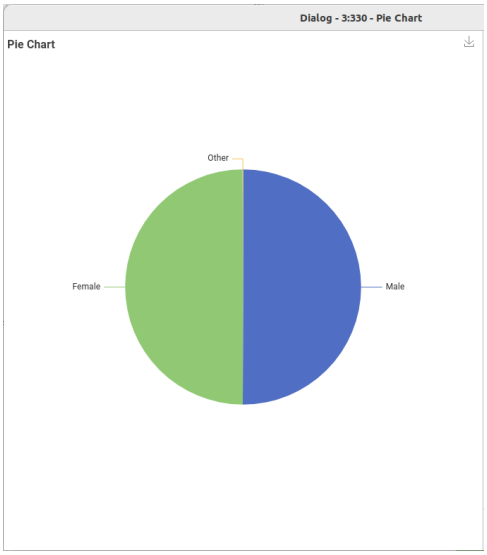


Figura 33: Pie Chart para 'Gender'

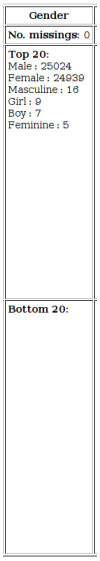


Figura 34: Estatisticas apresentadas

O atributo **Gender** não possui quaisquer **missing values**, possui sim uma pequena quantidade de entradas únicas, sendo **6** o valor. Ainda, apresenta uma distribuição Baixa e equilibrada dentro dos 2 valores mais comuns **Male** e **Female**.

3.2.2. Blood Type

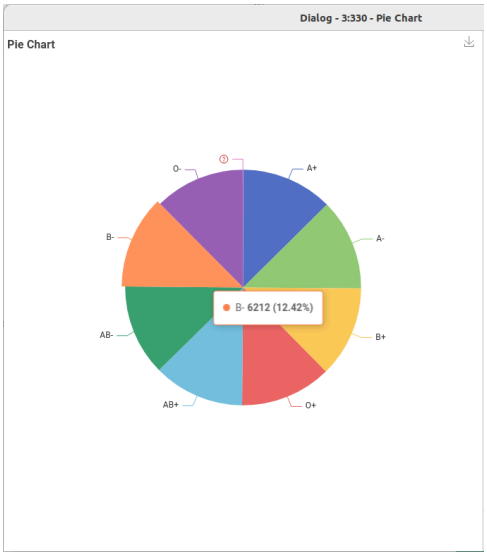


Figura 35: Pie Chart para 'Blood Type'

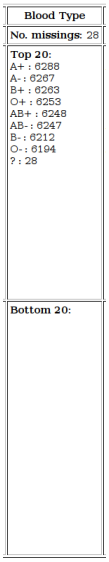


Figura 36: Estatisticas apresentadas

O atributo **Blood Type** contém **28 missing values**, e uma pequena quantidade de entradas únicas, sendo **9** o valor. Ainda, apresenta uma distribuição Alta e equilibrpreferênciaferencia por um valor.

3.2.3. Medical Condition

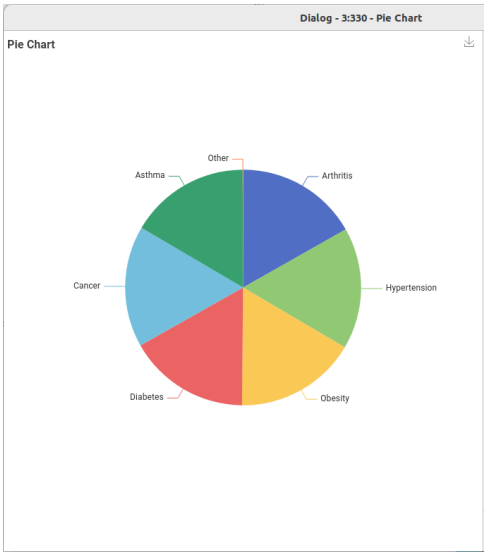


Figura 37: Pie Chart para 'Medical Condition'

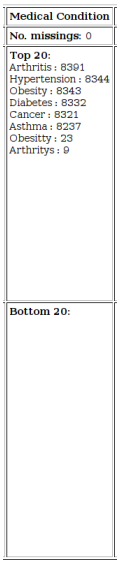


Figura 38: Estatisticas apresentadas

O atributo **Medical Condition** contém **0 missing values**, e uma pequena quantidade de entradas únicas, sendo **8** o valor. Ainda, apresenta uma distribuição Alta e equipreferência preferencia por um valor.

3.2.4. Insurance Provider

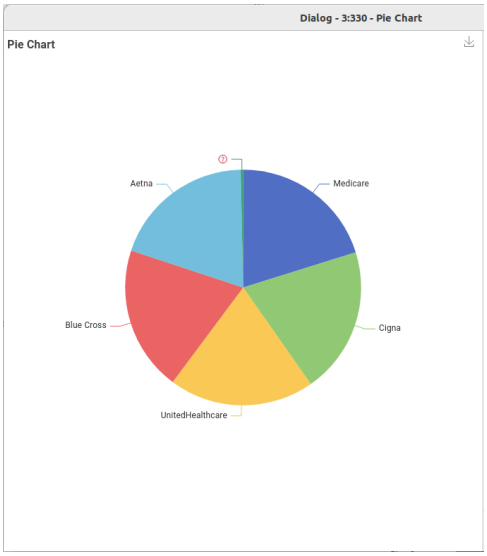


Figura 39: Pie Chart para 'Insurance Provider'

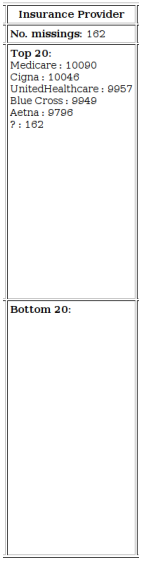


Figura 40: Estatisticas apresentadas

O atributo **Insurance Provider** contém **162 missing values**, e uma pequena quantidade de entradas únicas, sendo **5** o valor. Ainda, apresenta uma distribuição Alta epreferênciaa sem preferencia por um valor.

3.2.5. Admission Type

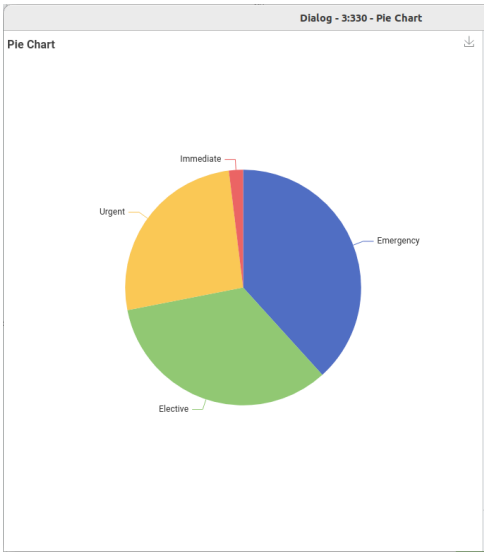


Figura 41: Pie Chart para 'Admission Type'

Figura 42: Estatisticas apresentadas

O atributo **Admission Type** contém **0 missing values**, e uma pequena quantidade de entradas únicas, sendo **4** o valor. Ainda, apresenta uma distribuição Alta e não equilibrada, com uma baixa representação de immediate e uma pequena preferência por **Emergency**.

3.2.6. Medication

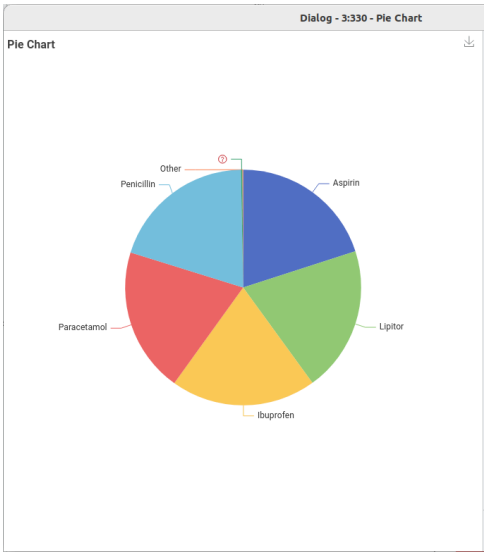


Figura 43: Pie Chart para 'Medication'

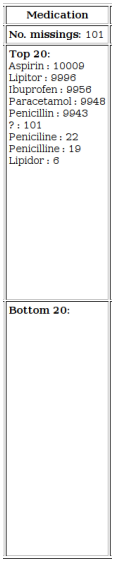


Figura 44: Estatisticas apresentadas

O atributo **Medication** contém **101 missing values**, e uma pequena quantidade de entradas únicas, sendo **9** o valor. Ainda, apresenta uma distribuição Alta e equilibrada sem preferencia por um valor.

3.2.7. Test Results - Target

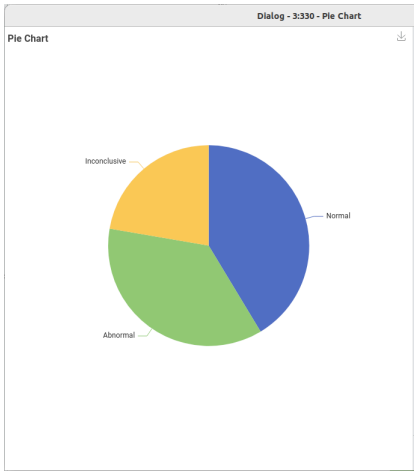


Figura 45: Pie Chart para 'Test Results'

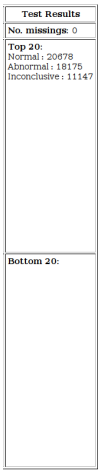
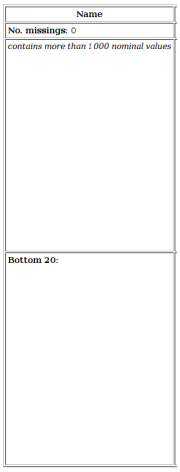
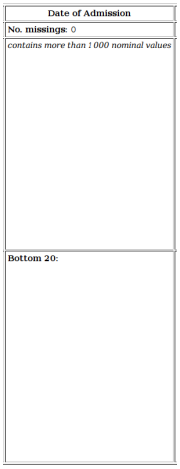


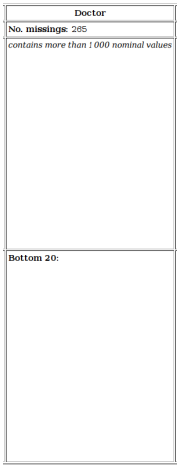
Figura 46: Estatisticas apresentadas

O atributo **Test Results** contém **0 missing values**, e uma pequena quantidade de entradas únicas, sendo **3** o valor. Ainda, apresenta uma distribuição Alta e não muito equilibrada, com ausência de representação de **Inconclusive** e uma pequena preferência por **Normal**.

3.2.8. Restantes Nominais







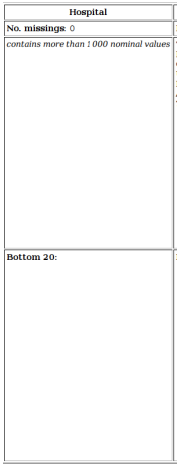
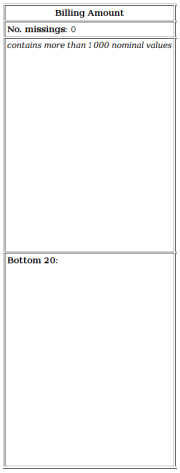
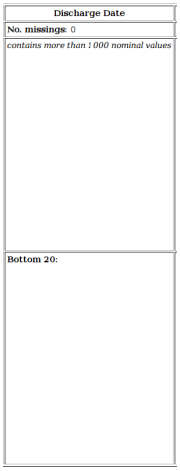


Figura 47: NameFigura 48: Date of AdmissionFigura 49: DoctorFigura 50: Hospital





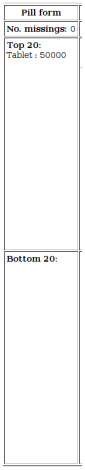


Figura 51: Billing AmountFigura 52: Discharge DateFigura 53: Pill Form

Todos estes atributos não contêm missing values, e apresentam um elevado numero de entradas únicas, com a exeção do Pill Form que só tem um valor. Para todos estes atributos tem de se considerar a possibilidade de os remover, ou um tratamento intensivo.

3.2.9. Restantes Numerais

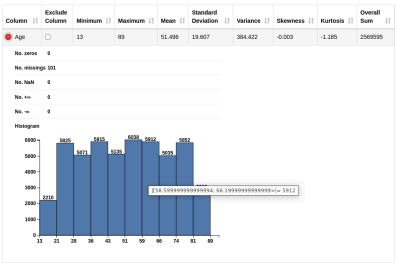


Figura 54: Age

O atributo Age contém 101 missing values, e uma média quantidade de entradas únicas, sendo 78 o valor. Ainda, apresenta uma distribuição Alta e não muito equilibrada, com os valores nos extremos (13-21, 81-89) com baixa representação. A idade média é de 51 anos.

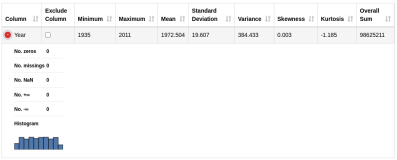


Figura 55: Year

O atributo Year contém 0 missing values, e uma média quantidade de entradas únicas, sendo 77 o valor. Ainda, apresenta uma distribuição Alta e não muito equilibrada, com os valores nos extremos com baixa representação. O ano médio é o 1973.

Figura 56: Month

O atributo Year contém 0 missing values, e uma baixa quantidade de entradas únicas, sendo 12 o valor. Ainda, apresenta uma distribuição Alta e não muito equilibrada, com os valores nos extremos com alta representação. O mês médio é o 6.

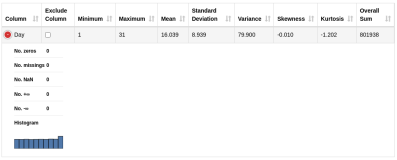


Figura 57: Day

O atributo Year contém 0 missing values, e uma média quantidade de entradas únicas, sendo 31 o valor. Ainda, apresenta uma distribuição Alta e equilibrada, com uma pequena preferência pelos dias finais. O dia médio é o 16.

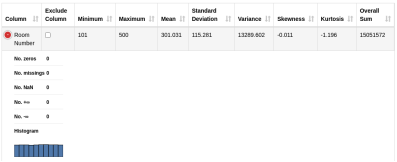


Figura 58: Room Number

O atributo Year contém 0 missing values, e uma alta quantidade de entradas únicas, sendo 400 o valor. Ainda, apresenta uma distribuição Alta e muito equilibrada. O quarto médio é o 301.

Figura 59: Floor

O atributo Year contém 85 missing values, e uma pequena quantidade de entradas únicas, sendo 6 o valor. Ainda, apresenta uma distribuição Alta e equilibrada, com uma ausência de representação do último valor (5). O piso médio é o 3.

3.2.10. Relação entre Atributos

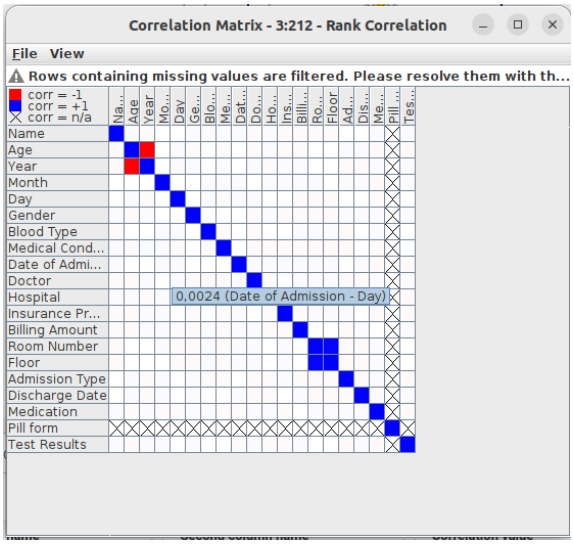


Figura 60: Rank Correlation

A imagem apresenta uma matriz de correlação entre diferentes atributos. A matriz utiliza a **rank correlation**, onde a cor **azul** indica uma correlação **positiva (+1)**, a **vermelha** uma correlação **negativa (-1)**, e os campos com **"X"** representam correlações **indisponíveis**.

Há uma **forte correlação negativa** entre **Age** e **Year**, o que é coerente, pois ano de nascença está muito relacionado à idade.

Há uma **forte correlação positiva** entre **Room Number** e **Floor**, o que é coerente, pois número de quarto estará dependente do andar.

De resto, existe **baixa correlações** entre os atributos, com a última menção relevante do **Pill Form** que apresenta **correlação inválida**, provavelmente por só apresentar um valor.

3.3. Exploração, Preparação & Tratamento dos Dados

Nesta secção, tal como anteriormente feito na primeira parte do relatório, destacámos as partes de exploração e tratamento de dados mais relevantes realizadas.

Inicialmente, após analisar nodos de visualização de dados, começámos por tornar a feature **Billing Ammount** do formato String para o formato Number (double).

De seguida, o grupo entendeu que as idades estavam a ser calculadas com base na data atual do sistema e não com base na data do momento da data de admissão. Desta forma, criámos a coluna **Birth Date** e uma nova coluna **Calculated Age** que resulta da idade calculada desde a **Birth Date** até à **Date of Admission**. Ainda em termos de tratamento de datas, criámos a coluna **Admission Time** calculada pela diferença entre a **Date of Admission** e a **Discharge Date**, representando esta a duração da estadia hospitalar, em dias:

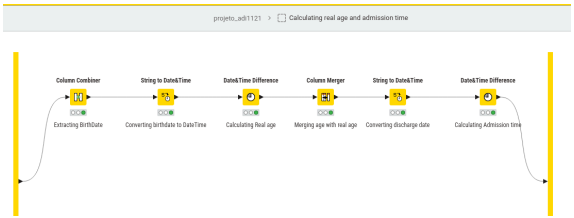


Figura 61: Cálculo da idade real e do tempo de admissão

Ainda, recorremos ao nodo String Manipulation para padronização dos valores de certas colunas tal como visto de seguida:

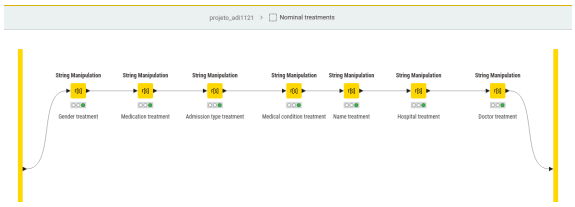


Figura 62: Padronização dos valores das diferentes features

Esta sequência de nodos resume-se no uso da função 'replace' para padronização de valores e ainda o uso da função 'lowerCase' para garantirmos que a comparação entre nomes de pacientes, doutores e hospitais seja insensível a maiúsculas e minúsculas.

Ainda, calculamos o valor do **Floor** com base no valor do **Room Number** para uma dada linha, verificando se estes eram compatíveis em todas as linhas do dataset, isto é, se havia casos onde por exemplo, o número do quarto fosse 205 e o número do piso 1, o que por razões lógicas, não faria sentido.

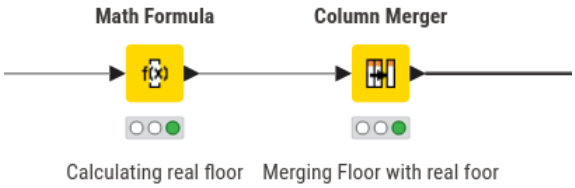


Figura 63: Tratamento da feature Floor

Ainda, criámos uma coluna chamada **Degree** com os graus académicos dos doutores e, apesar de crermos que esta coluna não tenha trazido informação relevante e que proporcionasse melhores resultados, decidimos manter como forma de mostrar esta decisão feita pelo grupo.

```
COLUMN EXPRESSIONS (LEGACY) NODE

column("Doctor").replaceAll("(?i)\\s*(PhD|MD|DDS)\\s*$", "").trim();
```

Para o tratamento dos missing values da feature **Billing Ammount** decidimos colocar como 0 o valor nos casos em que este era negativo. Desta forma, foi possível obter melhores resultados nos modelos preditivos.

```
RULE ENGINE NODE

$Billing Amount$ < 0 => 0
TRUE => $Billing Amount$
```

De seguida, tratamos dos missing values removendo as linhas onde os valores de **Doctor**, **Blood Type**, **Medication**, **Insurance Provider** e **Billing Ammount** eram valores em falta. Apenas no fim, removemos as linhas duplicadas de modo a garantir que todo o tratamento que pudesse levar a que duas ou mais linhas fossem iguais, já tivesse sido feito.

Por fim, normalizámos o conjunto de dados da target **Test Results**, uma vez que os valores obtidos pelos modelos preditivos estavam a ser influenciados por um desbalanceamento dos dados que originava resultados fracos. Para isso, recorremos aos nodos Equal Size Sampling e SMOTE, tendo obtido melhores resultados com o primeiro.

3.4. Decisões Inteligentes na Preparação & Tratamento dos Dados

- Padronizar as strings de **Hospital**, **Doctor** e **Name** para lowerCase, de modo a ser insensível a maiúsculas e minúsculas. Desta forma, ao remover linhas duplicadas, por exemplo, garantimos um resultado mais preciso uma vez que a insensibilidade a maiúsculas e minúsculas não afeta a informação real dos valores.
- Remover linhas duplicadas apenas no final de todo o tratamento, de modo a garantir que de facto, não haja tratamento posterior que seja capaz de tornar duas linhas iguais, resultando nesse caso em informação redundante, falsa e enganadora aos modelos preditivos.
- Calcular o número do **Floor** com base no valor do **Room Number** e verificar se, em todos os casos, havia compatibilidade entre ambos. Acabámos por verificar que sim.
- Normalizar o conjunto de dados da target **Test Results**, resultando assim em melhores resultados nos modelos preditivos desenvolvidos.
- Realização de diferentes tipos de tratamento em casos mais suscetíveis a dúvidas, verificando quais os melhores tratamentos para obter melhores resultados nos modelos preditivos.

3.5. Data Visualization

3.5.1. Pie Chart - Gender

Este gráfico tem como objetivo visualizar a distribuição de género entre os pacientes do dataset. Conseguimos perceber rapidamente que a proporção de indivíduos do sexo masculino e feminino são semelhantes, não existindo uma discrepância nas proporções. Com a utilização deste pie chart conseguimos obter esse feedback rápido de que o dataset possui valores equilibrados, possuindo apenas um minoria de valores que precisarão de ser tratados.

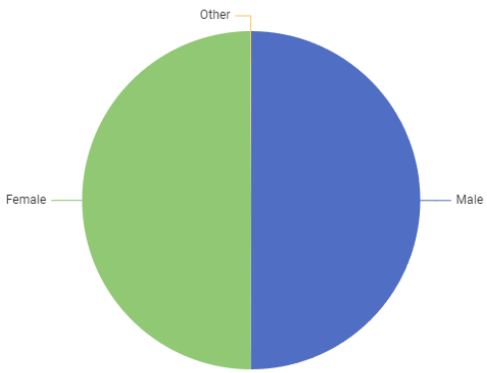


Figura 64: Pie Chart - Gender

3.5.2. Pie Chart - Test Result (Target)

A variável **Test Results**, sendo a variável alvo do nosso modelo, foi também representada através de um Pie Chart. Este gráfico evidenciou um pequeno desbalanceamento entre as classes, sendo a class “Normal” a mais frequente, com uma presença muito reduzida da classe “Inconclusive”, sendo que esta tem cerca de 1 / 5 das 3 classes totais. Isto tudo é essencial para justificar a necessidade de técnicas de balanceamento como o **Equal Size Sampling**, uma vez que um modelo treinado em dados desbalanceados tende a favorecer a classe mais comum.

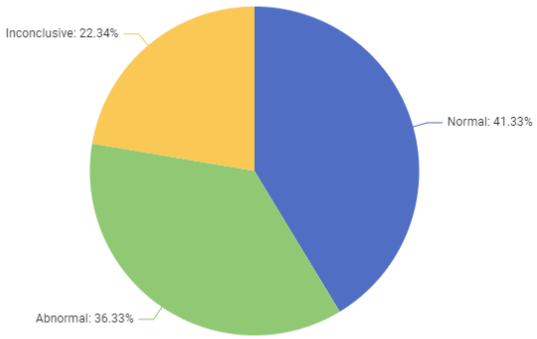


Figura 65: Pie Chart - Test Result

3.5.3. Bar Chart - Blood Type & Medication

Para analisar a distribuição dos diferentes tipos sanguíneos, foi utilizado um **Bar Chart** para a variável em questão, o que revelou que, embora existam 9 categorias diferentes, estes estão distribuídos de forma equilibrada, o que não leva a necessidade de agrupar categorias menos representadas. O mesmo pode ser verificado no Bar chart para a variável **medication**.

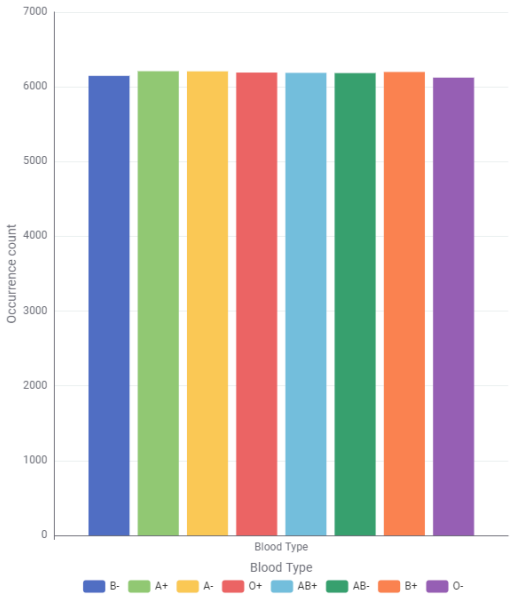


Figura 66: Bar Chart - Blood Type

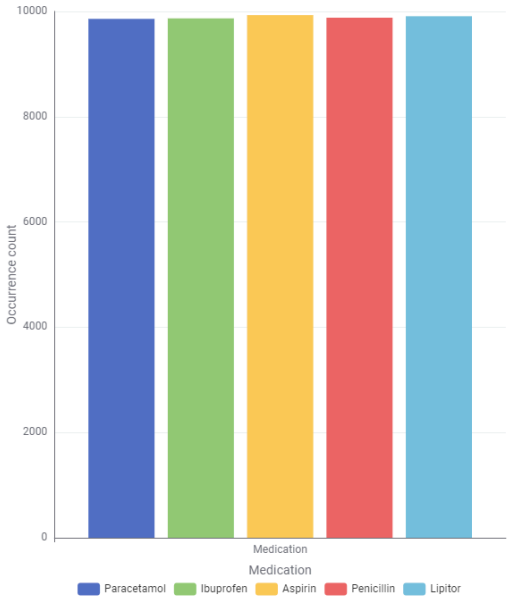


Figura 67: Bar Chart - Medication

3.5.4. Histogram - Age

Aqui para este histograma estamos a usar a Age, que foi calculada através da data de nascimento e data de admissão no hospital. Esta visualização evidenciou-nos que a distribuição das idades é muito semelhante, onde o pico se encontra perto dos 60 anos, tendo uma baixa representatividade nas idades antes dos 20 e depois dos 85. Utilizando este histograma conseguimos perceber de forma fácil se existem algum tipo de tendências ou assimetrias relevantes, o que poderia afetar o comportamento dos modelos.

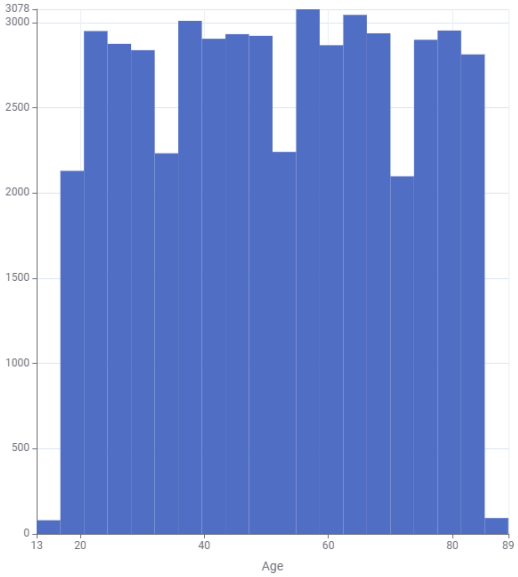


Figura 68: Histogram - Age

3.5.5. Box Plot - Billing Ammount

A variável **Billing Amount** foi explorada com um boxplot de forma a identificar possíveis outlier e a variabilidade dos dados, mas como podemos ver, o gráfico revelou que não existe uma grande dispersão nos valores, sendo que também **não existe** nenhum tipo de **outlier visível**, sendo que embora existam valores de cerca de 52 mil, estes continuam dentro de um intervalo esperado. Este gráfico foi útil na etapa de **limpeza de dados**, já que nos permitiu confirma a eficácia do tratamento aplicado.

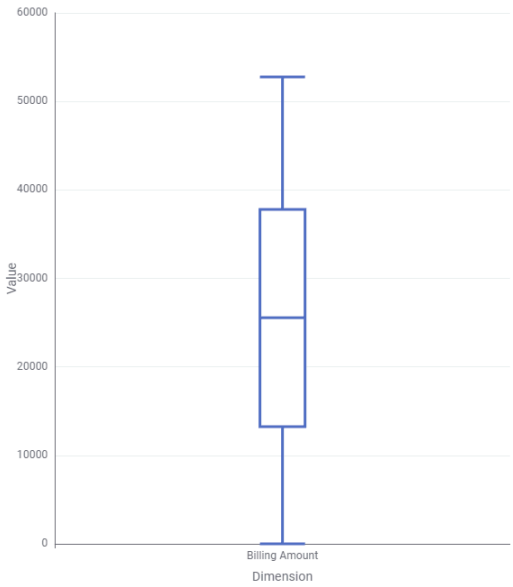


Figura 69: Box Plot - Billing Ammount

3.6. Modelação - Apresentação dos Modelos Desenvolvidos

Aqui, apresentámos em forma de tabela os vários testes que fizemos a nível dos modelos preditivos, variando os hiperparámetros de cada um, de modo a obter resultados mais positivos em cada um dos modelos desenvolvidos.

3.6.1. Neural Network

Tentativa	Nº Iterations	Nº Hidden Layers	Nº Hidden Neurons	Accuracy
0	100	1	10	41,375%
1	1000	1	10	41,375%
2	100	10	10	36,320%
3	100	5	10	41,375%
4	100	1	20	36,320%
5	100	1	5	41,375%
6	100	1	2	41,375%

3.6.2. SVM

Tentativa	Accuracy
0	32,992%

3.6.3. Gradient Boosted Tree

Tentativa	Tree Depth	Learning Rate	Accuracy
0	4	0,1	34,702%
1	10	0,1	36,178%
2	15	0,1	36,562%

Tentativa	Tree Depth	Learning Rate	Accuracy
3	15	0,2	37,250%
4	15	0,3	36,987%

3.6.4. Tree Ensemble

Tentativa	Tree Depth	Split Node SIze	Childe Node Size	Accuracy
0	Unchecked	Unchecked	Unchecked	40,526%
1	20	Unchecked	100	40,384%
2	5	Unchecked	100	37,270%
3	Unchecked	5	100	39,555%
4	Unchecked	2	100	40,202%
5	Unchecked	Unchecked	5	38,438%
6	Unchecked	Unchecked	2	38,200%

3.6.5. Decision Tree

Tentativa	Pruning	Record per Node	Record Store	Accuracy
0	Yes	2	10000	34,317%
1	No	2	10000	36,954%
2	No	4	10000	36,604%
3	No	1	10000	37,329%
4	No	1	20000	37,329%
5	No	1	5000	37,329%

3.6.6. Random Forest

Tentativa	Tree Depth	Node Size	Nº Models	Accuracy
0	Unchecked	Unchecked	100	40,526%
1	20	Unchecked	100	40,384%
2	5	Unchecked	100	37,270%
3	Unchecked	5	100	38,484%
4	Unchecked	2	100	38,625%
5	Unchecked	Unchecked	1000	39,697%
6	Unchecked	Unchecked	500	39,717%
7	Unchecked	Unchecked	200	40,425%

3.7. TOP 6 Modelos - Variação dos hiperparámetros

Modelo	Tentativa	Accuracy
Neural Network	0	41,375%
Random Forest	0	40,526%
Tree Ensemble	0	40,526%

Modelo	Tentativa	Accuracy
Decision Tree	3	37,329%
Gradient Boosted Tree	3	37,250%
SVM	0	32,992%

3.8. Análise Crítica dos Resultados Obtidos - Avaliação

Após testes exaustivos com diferentes modelos preditivos **(Neural Network, Random Forest, Tree Ensemble, Decision Tree, Gradient Boosted Tree, SVM)**, foi possível constatar que nenhum modelo um resultado de elevada precisão, com uma accuracy máximas a rondar os 41%, sendo que de entre todas, a que apresentou um melhor desempenho foi a **Neural Network**, com uma configuração simples (1 hidden layer e 10 hidden neurons) atingindo **41,375%**. O **Random Forest** e o **Tree Ensemble** apresentaram desempenhos muito próximos (ambos com **40,526%**), reforçando a robustez deste métodos mesmo sem grande afinação.

O **Gradient Bossted Tree** alcançou um máximo de **37,250%**, beneficiando ligeiramente do aumento da profundidade das árvores e da taxa de aprendizagem. Já o **Decision Tree**, com um melhor resultado de **37,329%**, mostrou-se sensível a alterações do número mínimo de registos por nó.

A performanca geral sugere que o conjunto de dados apresenta desafios substanciais, como **missing values, distribuições desbalanceadas e pouco viabilidade significaitva** em atributos relevantes como o Test Results. Apesar do uso de técnicas de balanceamento como o **SMOTE** e **Equal Size Sampling**, o impacto sobre os resultados foi limitado.

Este comportamento evidencia que as relações entre os atributos e a variável target são **fraca ou pouco representativas**, e que talvez seja necessário um tratamento de dados **mais profundo** ou a inclusão de **novas variáveis** para melhorar os modelos.

4. Conclusão

Ao longo deste projeto prático, tivemos a oportunidade de aplicar na prática uma variedade de técnicas ligadas à aprendizagem automática, algumas delas discutidas ao longo do semestre, outras que descobrimos e explorámos por iniciativa própria.

O trabalho envolveu não apenas a construção dos modelos, mas também uma fase importante de análise e tratamento dos dados. Este processo foi essencial para compreendermos com mais profundidade o desafio em mãos e ajudou-nos a definir estratégias mais eficazes para enfrentá-lo.

No início, tivemos alguns obstáculos na escolha do conjunto de dados para a tarefa Dataset Grupo, o que nos levou a reavaliar e adaptar a nossa estratégia inicial. Além disso, durante a implementação dos dois workflows, deparamo-nos com várias questões técnicas e decisões metodológicas que exigiram uma pesquisa adicional e discussão lógica e construtiva entre os elementos do grupo. Ainda assim, conseguimos ultrapassar os desafios surgidos, e o resultado final acabou por superar as nossas expectativas!

Desenvolvemos soluções ajustadas aos dados utilizados e documentámos de forma rigorosa todas as fases do trabalho no workflow Knime e num relatório completo, o que consideramos um dos pontos fortes e fundamentais do projeto.

Referências - APA:

Rustamov, E. (2023). Cars Sales. Kaggle.com. https://www.kaggle.com/datasets/elvinrustam/cars-sales

Kaggle. (2024). Kaggle: Your home for data science. Kaggle.com. https://www.kaggle.com/