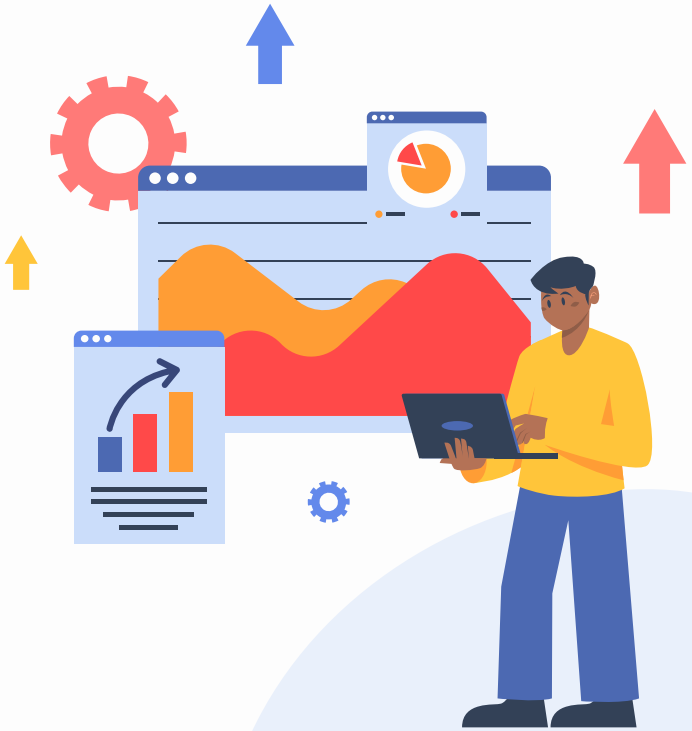
Conceção de modelos de aprendizagem e decisão

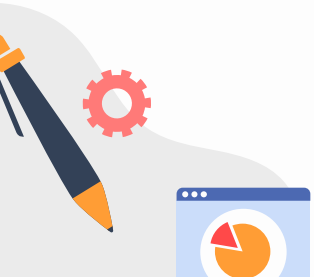


Aprendizagem e Decisão Inteligentes
3º Ano, 2º Semestre
Ano letivo 2024/2025

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

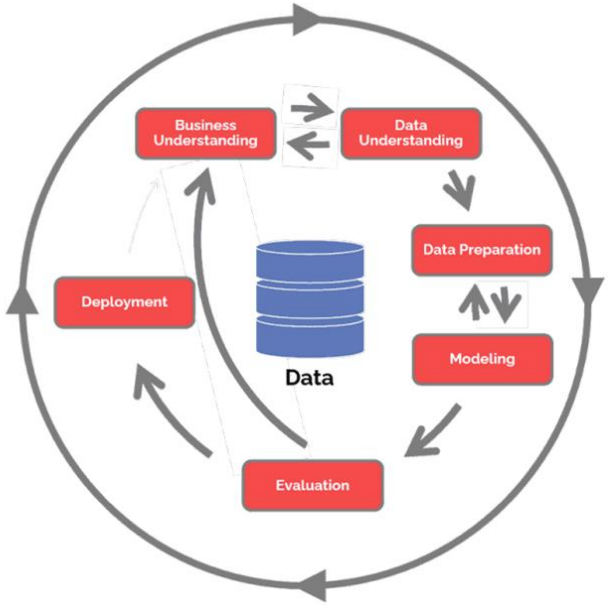
Índice



01
Introdução

Modelo CRISP-DM seguido de forma orientada, estruturada e iterativa.

- **Business Understanding** realizado previamente pelos docentes.
- **Data Understanding** com exploração e análise de padrões e tendências nos dados.
- **Preparação dos Dados** através de limpeza e transformação para análise.
- **Modelagem** com construção e avaliação de modelos preditivos.
- **Avaliação Final** com comparação de desempenho e qualidade dos modelos aplicados.



02

Tarefa Dataset Grupo - Cars Sales

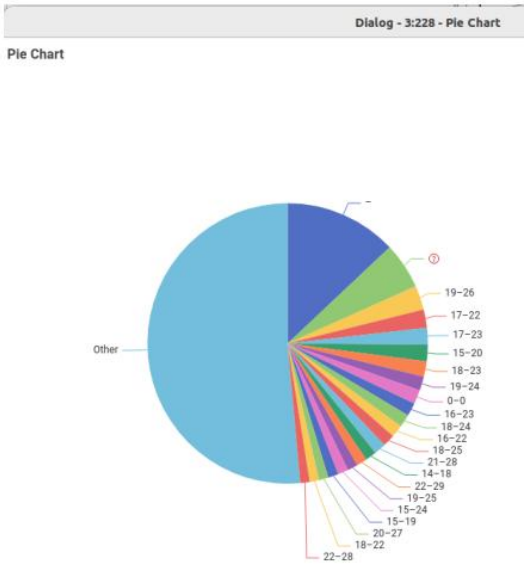
Apresentação do Dataset

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

Target: primaly_price

Características do Dataset & Análise dos Dados

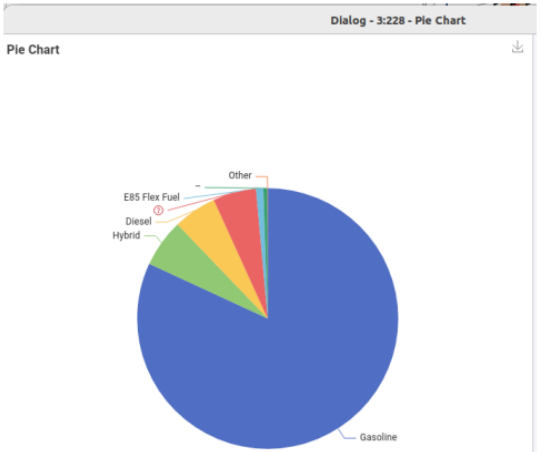
- O atributo **MPG** tem 334 missing values diretos e entradas com ‘–’ que representa ausência de dados.
- Quantidade de entradas únicas considerável (300).
- Apresenta uma distribuição elevada



MPG	
No. missings: 334	
Top 20:	
-: 812	
?: 334	
19-26: 174	
17-22: 132	
17-23: 121	
15-20: 119	
18-23: 111	
19-24: 102	
0-0: 100	
16-23: 95	
18-24: 93	
16-22: 89	
18-25: 87	
21-28: 86	
14-18: 80	
22-29: 76	
19-25: 75	
15-24: 75	
15-19: 74	
20-27: 68	
Bottom 20:	
12: 1	
16-27: 1	
29-38: 1	
11-17: 1	
25-29: 1	
0-0.0: 1	
23-25: 1	
50-43: 1	
26-28: 1	
20-31: 1	
50-54: 1	
28-28: 1	
13-0: 1	
18-17: 1	
13-14: 1	
23-22: 1	
43-43: 1	
19-22.0: 1	
15-17: 1	
15-0: 1	

Características do Dataset & Análise dos Dados

- O atributo **Fuel Type** tem 334 missing values diretos e entradas com ‘–’ que representa ausência de dados.
- Quantidade de entradas únicas baixa (9).
- Apresenta uma distribuição baixa e não equilibrada com uma enorme preferência por Gasoline



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:

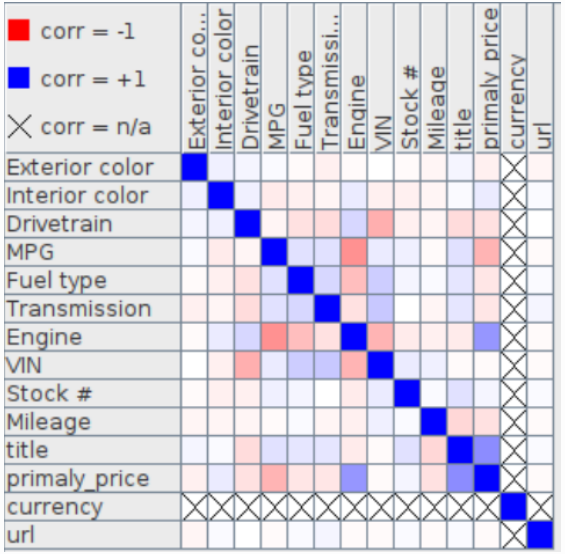
Características do Dataset & Análise dos Dados

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

- Engine e primaly_price
- Title e primaly_price

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

- Engine e MPG
- MPG e primaly_price



Exploração, Preparação & Tratamento dos Dados

Limpeza e Pré-processamento de Dados

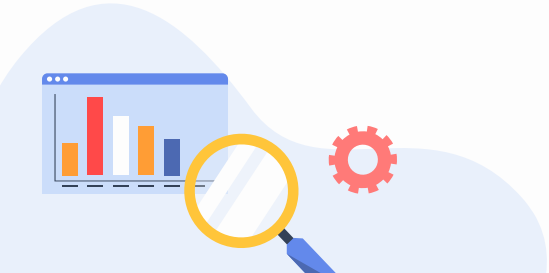
Remoção de colunas não informativas:

- Colunas com valores únicos ou com um único valor:
    - `VIN`, `Stock #`, `Exterior Color`, `Interior Color`, `currency`.

Tratamento de valores inválidos, nomeadamente para a feature MPG:

- Substituição de '–' por missing values:
    - `String Manipulation (Multi Column)` com expressão regular:

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



Exploração, Preparação & Tratamento dos Dados

Tratamento de Features e Missing Values

Tratamento do atributo MPG:

- Separação em **`min_MPG`** e **`max_MPG`**.
- Conversão para inteiro e validação de consistência.
- Para veículos elétricos:
    - Usado `Rule Engine`: `$EngineConfig$ MATCHES ".*Electric.*" => 0`

Imputação de valores em falta:

- **`Mileage`**: Média por ano do veículo.
- **`min_MPG`**, **`max_MPG`**, **`Valves`**: Substituição por média geral.
- Remoção de linhas com missing values restantes, devido à pequena perda de informação.

Exploração, Preparação & Tratamento dos Dados

Tratamento de Features e Missing Values

Tratamento do atributo Engine:

- Sequência de nodos **String Manipulation** com expressões regulares**.**
- Extração de valores relevantes, criando novas features com informação útil.
- Extrair a config da feature Engine:
    - Usado String Manipulation:

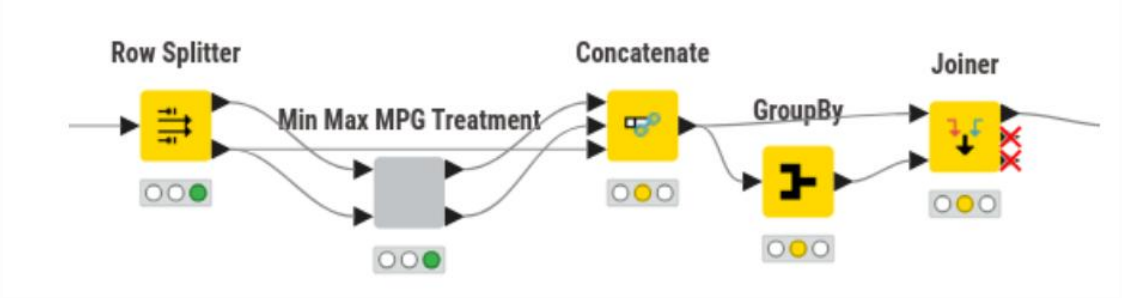
```
regexReplace($Engine$, "(?i).*?(?<![A-Za-z])([A-Z])\\-?(\\d+).*", "$1$2")
```

Feature Engineering - Feature Extraction + Data Type Conversion**:**

- Partição dos dados fornecidos pela feature **Engine** para obtermos mais informação relevante dos veículos (**Liter, EngineConfig, Valves**) com String Manipulation.
- Conversão para inteiro (**Liter, Valves**).

Exploração, Preparação & Tratamento dos Dados

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

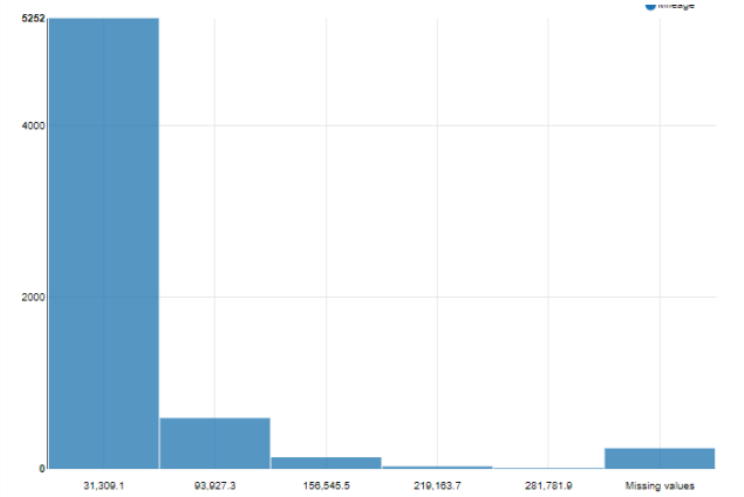


Exploração, Preparação & Tratamento dos Dados

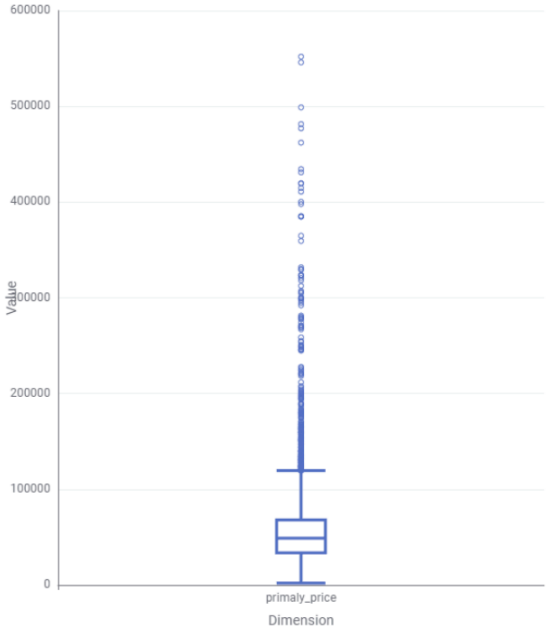
Decisões Inteligentes

- Padronização dos valores das diferentes features.
- Substituição direta de missing values e posterior imputação pela média, mediana e remoção de linhas.
- **Feature Engineering**.
- Análise da **correlação** dos dados e visualização de estatísticas em vários momentos.
- Cuidados a nível lógico - **MPG** e **Fuel Type**.

Exploração, Preparação & Tratamento dos Dados



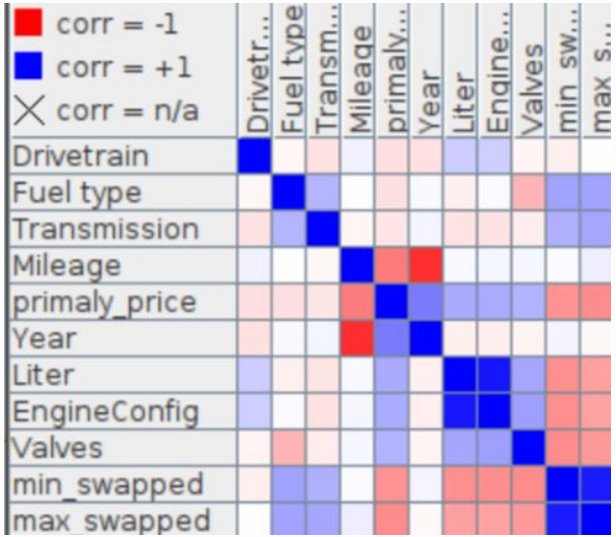
Histograma - Mileage



Box Plot - Primaly Price

Exploração, Preparação & Tratamento dos Dados

Relação entre atributos após
preparação e tratamento dos dados



Modelação e Análise de Resultados

Concluímos que a **Regressão Linear** (R^2 = 0,646) é inadequada, devido à sua limitação em captar relações não lineares. A **Regression Tree** mostrou ligeira melhoria (até R^2 = 0,779), mas com ganhos limitados. O **Random Forest** (R^2 = 0,796) demonstrou maior robustez e desempenho consistente. O melhor resultado foi obtido com o **Gradient Boosted Tree** (R^2 = 0,835), superando inclusive as **Redes Neuronais** (R^2 = 0,725), o que sugere que o problema possui fortes **relações não lineares** que são melhor modeladas por **métodos baseados em árvores**.

R^2:	0,835
Mean absolute error:	6 588,261
Mean squared error:	101 254 790,617
Root mean squared error:	10 062,544
Mean signed difference:	108,741
Mean absolute percentage error:	0,131
Adjusted R^2:	0,835

TOP 5

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

03
Tarefa Dataset Atribuído - Healthcare

Apresentação do Dataset

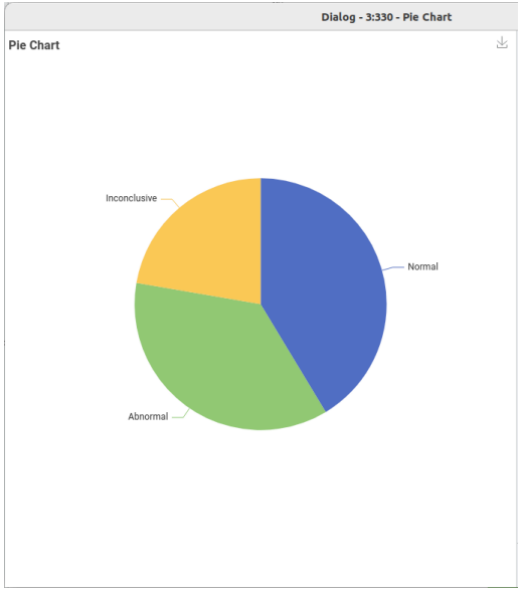
Feature	Descrição	Data Type
Name	Nome do paciente associado ao registo clínico	String
Age	Idade do paciente à data da admissão	Integer
Gender	Género do paciente (Male ou Female)	String
Blood Type	Tipo de sangue do paciente (e.g., A+, O-, etc.)	String
Medical Condition	Principal condição médica ou diagnóstico 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

Target: Test Results

Características do Dataset & Análise dos Dados

Atributos nominais com distribuição equilibrada, exceto:

- Test Results, menor representação de **Inconclusive**



Características do Dataset & Análise dos Dados

Atributos numerais com distribuição equilibrada, exceto:

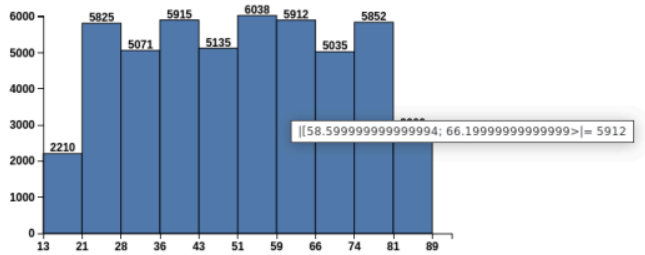
- **Age**
- **Year**
- **Month**

Todos estes apresentam baixa representação nos extremos, por exemplo, na Age: 13-21 e 81-89

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis	Overall Sum
Age	☐	13	89	51.496	19.607	384.422	-0.003	-1.185	2569595

No. zeros	0
No. missings	101
No. NaN	0
No. +∞	0
No. -∞	0

Histogram



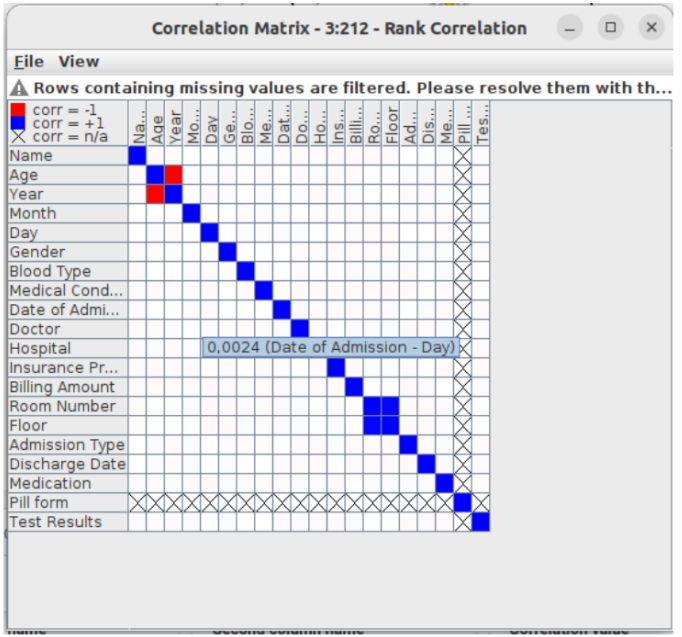
Características do Dataset & Análise dos Dados

Observa-se uma forte **correlação positiva** entre:

- Room Number e Floor

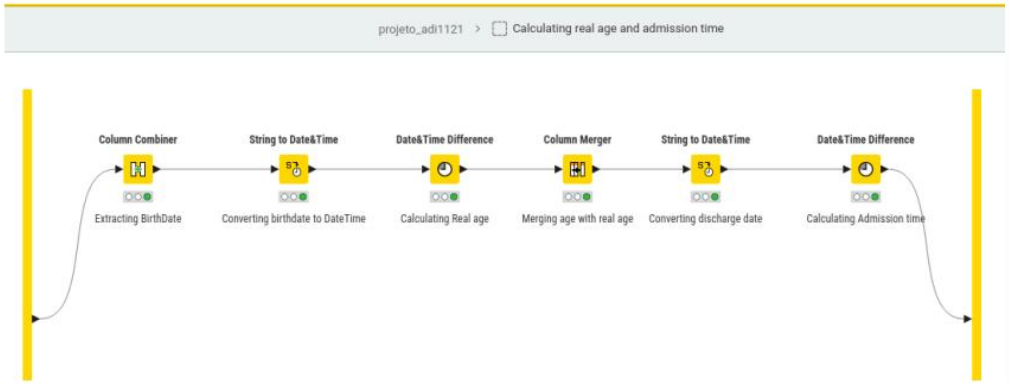
Observa-se uma **correlação negativa** entre:

- Age e Year



Exploração, Preparação & Tratamento dos Dados

- 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** que e uma nova coluna **Calculated Age** que resulta da idade calculada desde a **Birth Date** até à **Date of Admission**
- Criámos a coluna **Admission Time** calculada pela diferença entre a **Date of Admission** e a **Discharge Date**, representando esta a duração da estadia,

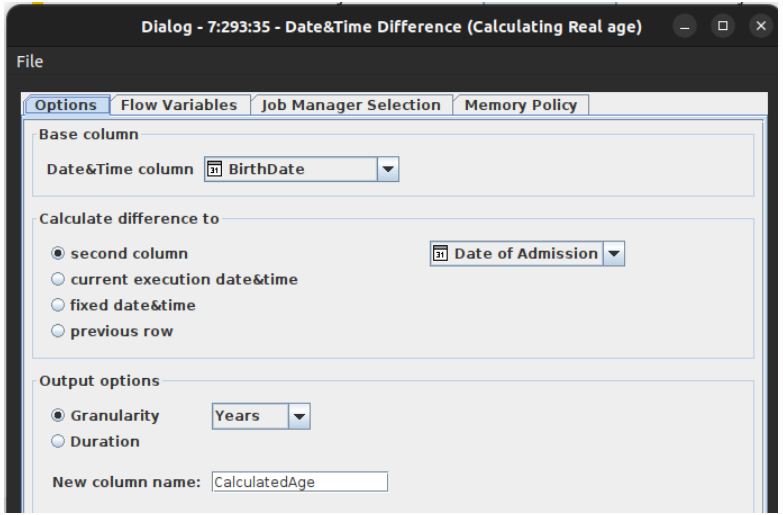


Exploração, Preparação & Tratamento dos Dados

Tratamento da idade e datas

Tratamento do atributo Age:

- Idade calculada erradamente com base na data do sistema.
- **Feature Engineering -** Criação da feature BirthDate e conversão para DateTime.
- Cálculo da idade real resultada pela diferença entre a data de admissão e data de nascimento.



Exploração, Preparação & Tratamento dos Dados

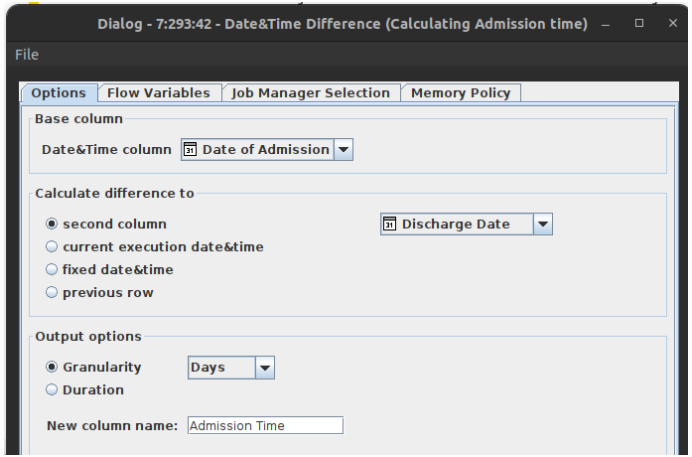
Tratamento de datas

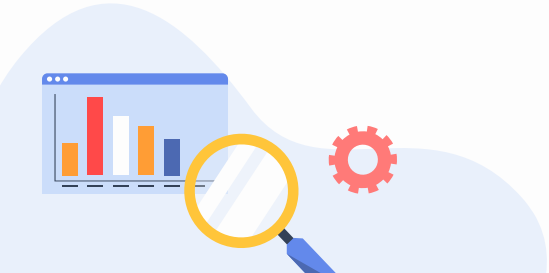
Tratamento do atributo Discharge Date:

- Converter para DateTime, útil para o posterior cálculo do tempo de admissão.

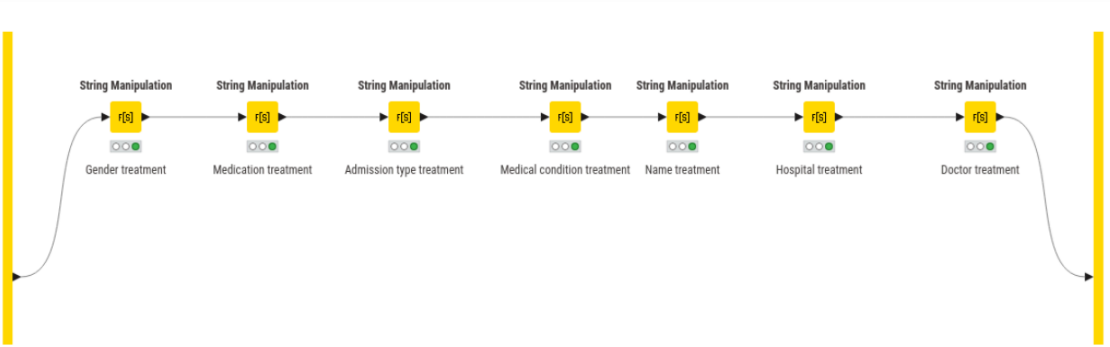
Criação da feature Admission Time:

- **Cálculo da duração da estadia hospitalar,** em dias**.**





Exploração, Preparação & Tratamento dos Dados



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.

Exploração, Preparação & Tratamento dos Dados

Cuidados lógicos, criação de novas features e tratamento de missing values

- Calculámos 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.

- Criámos uma coluna chamada **Degree** com os graus académicos dos doutores extraídos do nome e, embora esta não tenha trazido informação relevante que proporcionasse melhores resultados.

- Para o tratamento dos missing values da feature **Billing Amount** decidimos colocar como 0 o valor nos casos em que este era negativo.

  - Usado Rule Engine:

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

TRUE => $Billing
Amount$
```

Exploração, Preparação & Tratamento dos Dados

- Tratamos dos *missing values* removendo as linhas onde os valores de **Doctor, Blood Type, Medication, Insurance Provider e Billing Amount** 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.

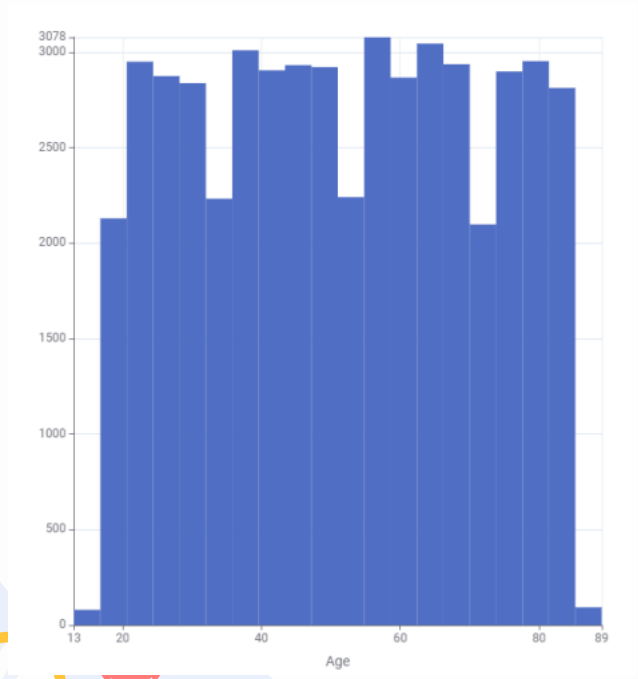
- Por fim, normalizámos o conjunto de dados da target **Test Results** e uma vez que os valores obtidos pelos modelos preditivos estavam a ser influenciados por um desbalanceamento dos dados que originava resultados fracos, recorremos aos nodos **Equal Size Sampling** e **SMOTE**, tendo obtido melhores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

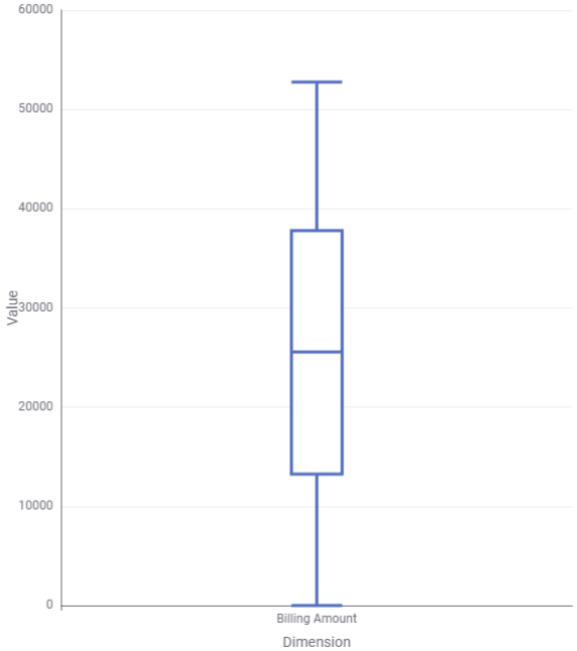
Exploração, Preparação & Tratamento dos Dados

Decisões Inteligentes

- Padronização dos valores das colunas **Hospital**, **Doctor** e **Name** para lowerCase e, só no fim, tratar de remover linhas duplicadas.
- Remoção de linhas duplicadas apenas após todo o tratamento ao garantir que qualquer tratamento posterior que leve a que duas linhas fiquem iguais, não aconteça.
- **Cuidados lógicos** ao verificar datas inválidas, idades inválidas e compatibilidade entre atributos**.**
- Balanceamento dos dados da target **Test Results** para melhor aprendizagem mais equilibrada dos modelos preditivos.
- Diferentes formas de **tratamento dos dados**, de modo a obter os resultados mais fiáveis.

Exploração, Preparação & Tratamento dos Dados





Modelação e Análise de Resultados

- Após testes com vários modelos, a melhor precisão obtida foi de **41,375%**, com uma **Neural Network simples** (1 hidden layer, 10 neurónios). **Random Forest** e **Tree Ensemble** também se destacaram (~40,5%). **Gradient Boosted Tree** (37,25%) e **Decision Tree** (37,33%) mostraram-se sensíveis a ajustes.
- Apesar da aplicação de técnicas como **SMOTE** e **Equal Size Sampling**, o impacto foi limitado. Os resultados sugerem **dados desbalanceados**, **valores em falta** e **baixa relevância de atributos**. Melhorias dependem de **tratamento de dados mais profundo** ou da **inclusão de novas variáveis**.

Test Resul...	Normal	Abnormal	Inconclusive
Normal	1396	649	1
Abnormal	1185	607	4
Inconclusive	727	375	1

TOP 6

Modelo	Tentativa	Accuracy
Neural Network	0	41,375%
Random Forest	0	40,526%
Tree Ensemble	0	40,526%
Decision Tree	3	37,329%
Gradient Boosted Tree	3	37,250%
SVM	0	32,992%

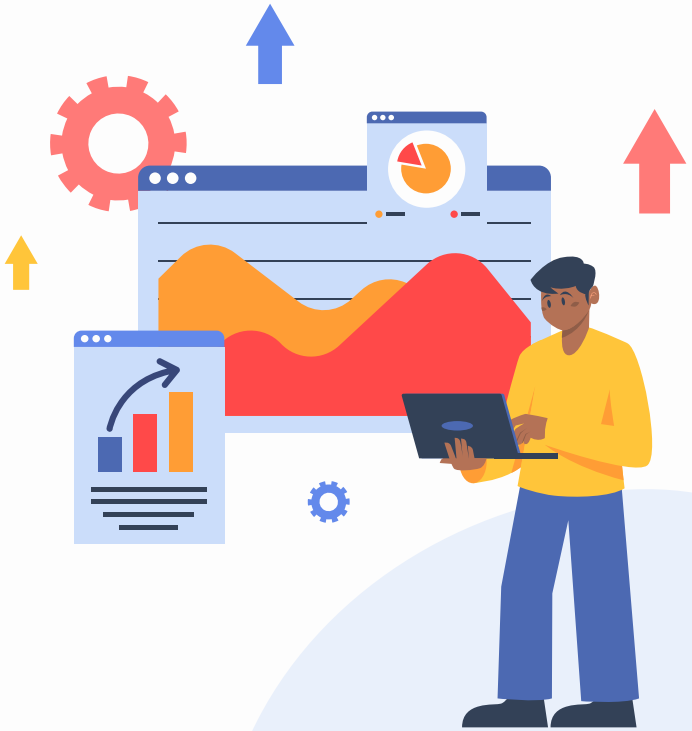
04
Conclusão

- Neste projeto aplicamos diversas técnicas de aprendizagem automática, combinando conteúdos lecionados com abordagens exploradas autonomamente. O trabalho envolveu **análise e tratamento de dados**, fundamentais para definir estratégias eficazes.

- Enfrentamos desafios na escolha inicial do dataset, o que exigiu **reavaliação da estratégia**. Durante os dois workflows, surgiram **questões técnicas e metodológicas**, resolvidas com pesquisa e colaboração entre os membros do grupo.

- Superamos as dificuldades e o resultado final superou as nossas expectativas. Destacamos a **documentação rigorosa** no KNIME e relatório detalhado como um dos principais pontos fortes do projeto.

Conceção de modelos de aprendizagem e decisão



Aprendizagem e Decisão Inteligentes
3º Ano, 2º Semestre
Ano letivo 2024/2025

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