

# UNIVERSIDADE DO MINHO

## LICENCIATURA EM ENGENHARIA INFORMÁTICA

---

Aprendizagem e Decisão Inteligentes

**Grupo 5**

---

## Conceção de Modelos de Aprendizagem

Joana Alves (A93290)

Maria Cunha (A93264)

Tânia Teixeira (A89613)

Vicente Moreira (A93296)

Maio 2022

# Conteúdo

<b>1</b>	<b>Introdução</b>	<b>2</b>
<b>2</b>	<b>Housing Regression</b>	<b>3</b>
2.1	Contexto e Introdução . . . . .	3
2.1.1	Objetivo do modelo . . . . .	3
2.2	Atributos do <i>Data Set</i> . . . . .	3
2.3	Exploração e Tratamento do <i>Data Set</i> . . . . .	4
2.3.1	Exploração . . . . .	4
2.3.2	Tratamento de Dados . . . . .	6
2.3.3	Visualização dos Dados . . . . .	8
2.4	Modelos Utilizados . . . . .	10
2.4.1	Simple Linear Regression - K-Holding . . . . .	10
2.4.2	Simple Linear Regression - K-Folds . . . . .	11
2.4.3	Polynomial Regression - K-Folds . . . . .	11
2.4.4	Multi Layer Perceptor . . . . .	12
2.5	Resultados e Conclusões . . . . .	13
<b>3</b>	<b>Collision DataBase</b>	<b>14</b>
3.1	Contexto e Introdução . . . . .	14
3.1.1	Objetivo do modelo . . . . .	14
3.1.2	Pré-Tratamento . . . . .	14
3.2	Atributos do <i>Data Set</i> . . . . .	15
3.3	Exploração e Tratamento de Dados . . . . .	16
3.3.1	Exploração de um Caso de Estudo . . . . .	16
3.3.2	Exploração de Correlação . . . . .	17
3.3.3	Tratamento do <i>Data Set</i> . . . . .	18
3.3.4	Visualização dos Dados . . . . .	19
3.4	Modelos Utilizados . . . . .	22
3.4.1	Decision Tree - K-Holding . . . . .	22
3.4.2	Decision Tree K-Folds . . . . .	23
3.4.3	Random Forest- K-Folds . . . . .	23
3.5	Resultados e Conclusões . . . . .	24
<b>4</b>	<b>Conclusão</b>	<b>25</b>

# 1 Introdução

Este trabalho prático foi realizado no âmbito da unidade curricular de Aprendizagem e Decisão Inteligentes e tem como objetivo aprofundar o conhecimento na matéria lecionada ao longo do semestre, através do desenvolvimento de dois modelos inteligentes capazes de prever e gerar conhecimento através da informação disponibilizada.

Este relatório contém todos os objetivos, decisões e problemas encontrados na exploração e tratamento de dois *data sets* distintos, assim como o desenvolvimento dos modelos de previsão. Para estes modelos, utilizamos uma abordagem de aprendizagem por supervisão visto que ambos os *data sets* de teste utilizados contêm informação sobre os resultados pretendidos.

O primeiro *data set*, fornecido pelos docentes, revolve-se na área da imobiliária e contém uma listagem de dados acerca destes imóveis, sendo o objetivo deste desenvolver um modelo de aprendizagem utilizando técnicas de **regressão**. Já o segundo *data set*, escolhido pelo grupo, trata sobre acidentes rodoviários, contendo informação sobre os intervenientes, assim como as várias condições do acidente. Para a construção do modelo deste problema, pretendemos utilizar e explorar várias técnicas de **classificação**.

## 2 Housing Regression

### 2.1 Contexto e Introdução

O *data set* "Housing Regression" foi o *data set* atribuído ao grupo de acordo com as regras impostas pelos docentes. Este oferece informação sobre o preço de casas no território dos Estados Unidos da América assim como a informação acerca da sua zona, nomeadamente, o número médio de divisões e quartos das casas da zona, o rendimento médio da população e o número de habitantes, assim como informação relativa à morada da casa em questão.

Este *data set* é constituído por 6 *features*, sendo a maior parte do tipo *Double*, com a exceção do endereço da casa, sendo este um atributo nominal. O endereço, visto que contém valores únicos para cada casa e, devido à sua dimensão, pode ser interpretado como um "pequeno texto".

#### 2.1.1 Objetivo do modelo

O nosso principal objetivo, definido pelo enunciado, será criar um modelo capaz de prever o valor da habitação numa região dos Estados Unidos da América, utilizando os restantes atributos como auxílio.

### 2.2 Atributos do *Data Set*

Atributos <i>Data Set</i> "Housing Regression"		
Atributo	Tipo de Dado	Descrição
Avg. Area Income	Double	Rendimento Médio dos Moradores na área
Avg. Area House Age	Double	Idade Média das Casas na área
Avg. Area Number of Rooms	Double	Nº Médio de Divisões das Casas na área
Avg. Area Number of Bedrooms	Double	Nº Médio de Quartos das Casas na área
Avg. Area Population	Double	População na área
Price	Double	Valor da Casa
Adress	String	Morada da Casa

## 2.3 Exploração e Tratamento do *Data Set*

### 2.3.1 Exploração

O primeiro passo tomado na avaliação do *data set* foi a exploração deste, ou seja, desenvolveu-se uma exploração inicial dos dados recebidos para que depois se conseguisse analisar corretamente estes. Sendo assim, preparamos os dados, através de alguns nodos, para que fosse possível obter um bom tratamento dos dados fornecidos a serem aplicados nos modelos de regressão.

- **Box Plot:** Para a avaliação de *outliers* relativos aos dados das casas, da área onde estas são localizadas e o preço estabelecido, recorremos a *Box Plots*.
- **Statistics:** Recolhemos uma tabela de estatísticas onde reparamos na ausência de *missing values* e portanto concluímos que não teremos de fazer qualquer tratamento específico para este problema.
- **Ranking Correlation - Correlation Matrix:** Por fim, utilizamos este nodo para obter uma matriz de correlação para uma melhor compreensão das relações entre os dados.

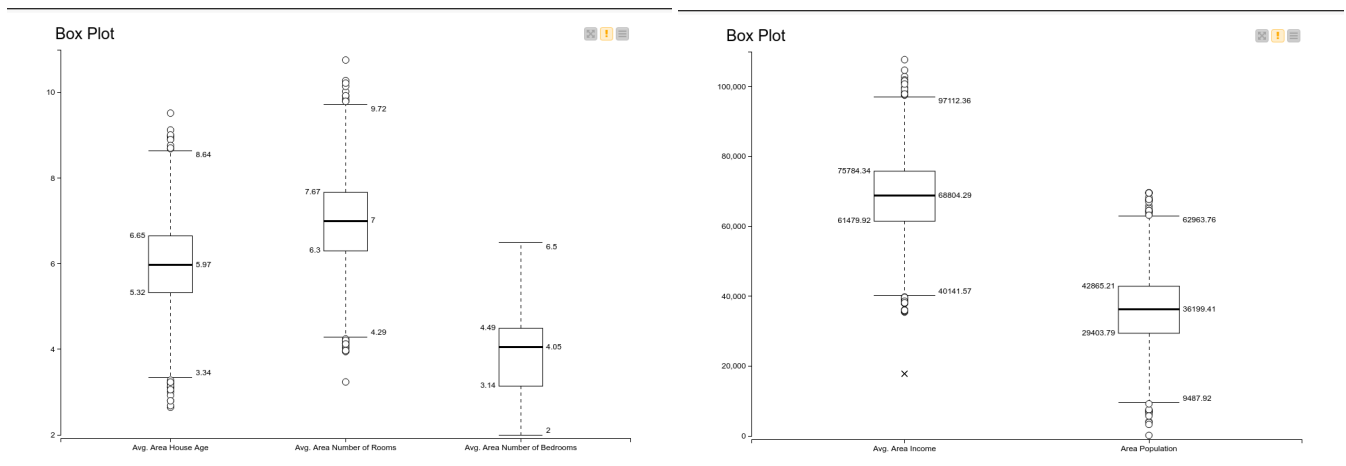


Figura 1: House BoxPlot

Figura 2: Area BoxPlot

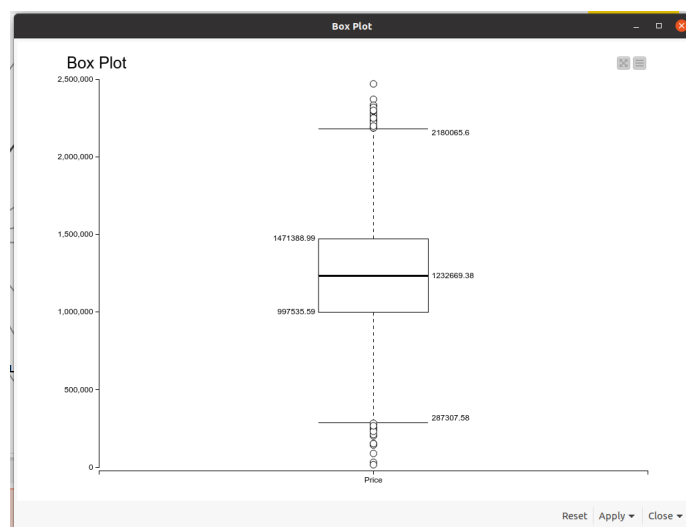


Figura 3: Price Box Plot

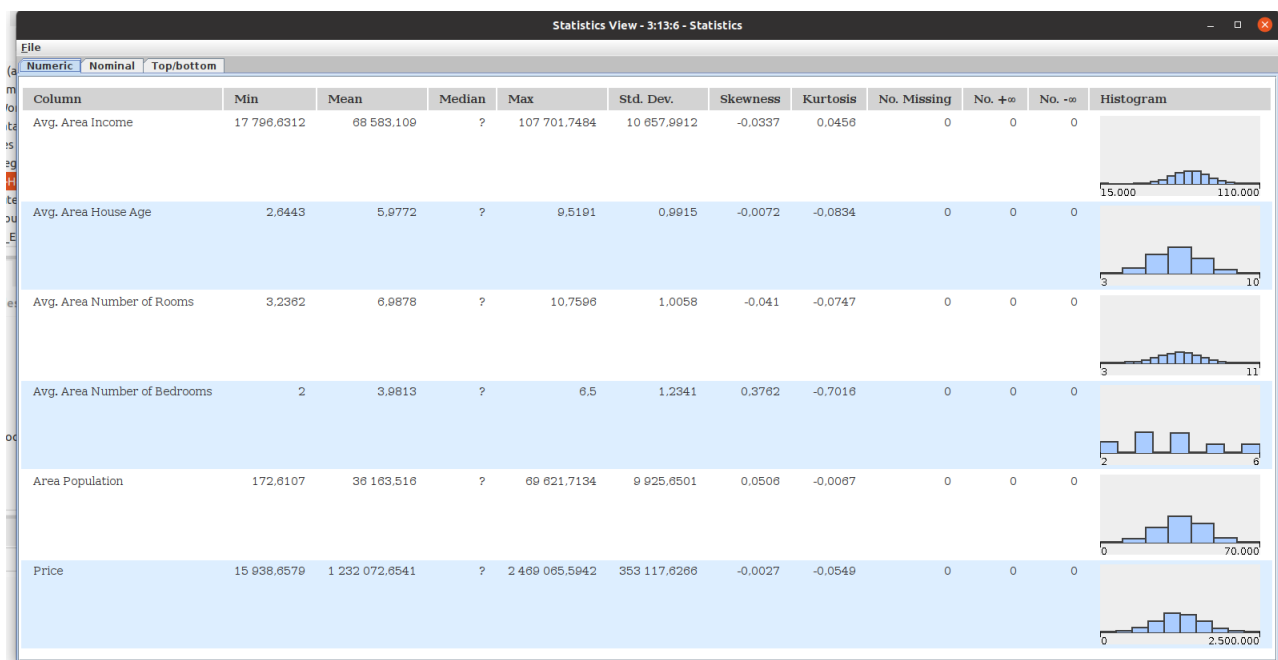


Figura 5: Estatísticas

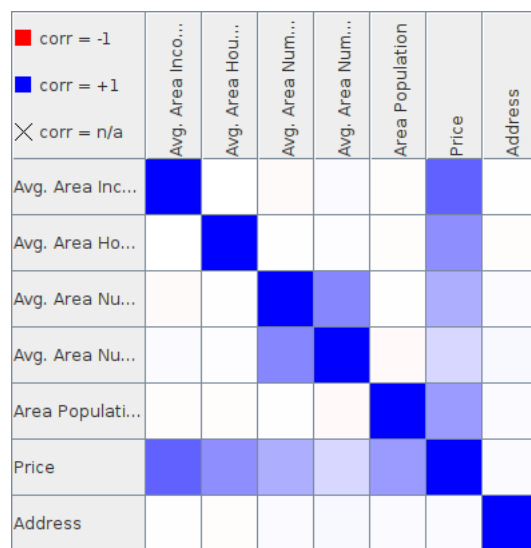


Figura 4: Matriz de Correlação

Pela observação da Figura 4, podemos encontrar algumas correlações pertinentes entre os dados, como, por exemplo, a correlação entre o número médio de divisões e o número médio de quartos, mostrando que estes estão interligados. Observamos, também, que o preço da casa está fortemente relacionado com as várias condições da sua zona, como o rendimento médio da população ou a dimensão desta.

Esta exploração deu-nos confiança no potencial da capacidade de previsão do modelo a ser desenvolvido para este problema.

### 2.3.2 Tratamento de Dados

Para além do tratamento a efetuar nos atributos numéricos, observamos que, como os endereços presentes no *data set* seguem a norma de endereços dos Estados Unidos da América, estes contêm informação implícita acerca da sua localização, como por exemplo, a sua cidade, o seu estado e a sua região global. Para além disto, também poderemos conseguir extrair informação mais granular de zonas a partir do *zip-code* (código postal).

Devido a isto, decidimos dividir o tratamento de dados em duas secções, uma primeira fase de tratamento e extração de informação a partir dos endereços e, de seguida, o tratamento dos atributos numéricos.

#### Tratamento de Endereços

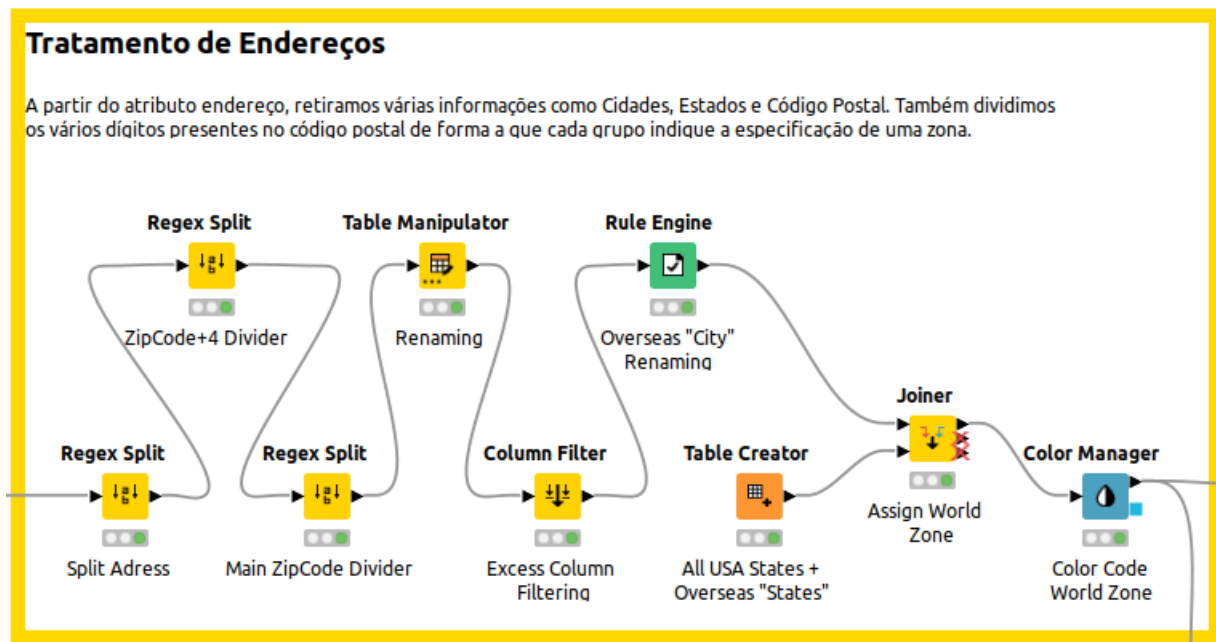


Figura 6: Tratamento de Endereços

- **Regex Split:** Para o processamento de endereços, recorremos a vários nodos *Regex Split*. A função destes nodos é encontrar os padrões no endereço e dividi-los de acordo com as várias "categorias" que definirmos. Apresentamos algumas imagens como exemplos para explicitar o processo de divisão.

0057 Jacob Coves Apt. 932 → Rua/Casa  
Lake Emily; VA 39465-6041  
↓                      ↓                      ↓  
Cidade                      Estado                      Código Postal



Também é de notar que, visto que existem endereços especiais, reservados a instituições diplomáticas e militares fora do território nacional, fomos obrigados a ter um cuidado extra na recolha desta informação.

- **Table Manipulator:** Renomeação das novas colunas geradas, para fácil interpretação.
- **Column Filter:** Remoção de colunas "lixo" geradas na fase de *Regex Split*.
- **Rule Engine:** Alteração das classificações do atributo "Cidade", em específico, os valores que correspondem a endereços especiais.
- **Table Creator & Joiner:** Criamos uma tabela com todos os estados pertencentes aos Estados Unidos da América, incluindo os estados *Overseas*, com as suas respectivas siglas e zona geográfica. De seguida, juntamos as tabelas de forma a acrescentar o atributo geográfico à casa através do nodo *Joiner*.

## Tratamento de Valores Numéricos

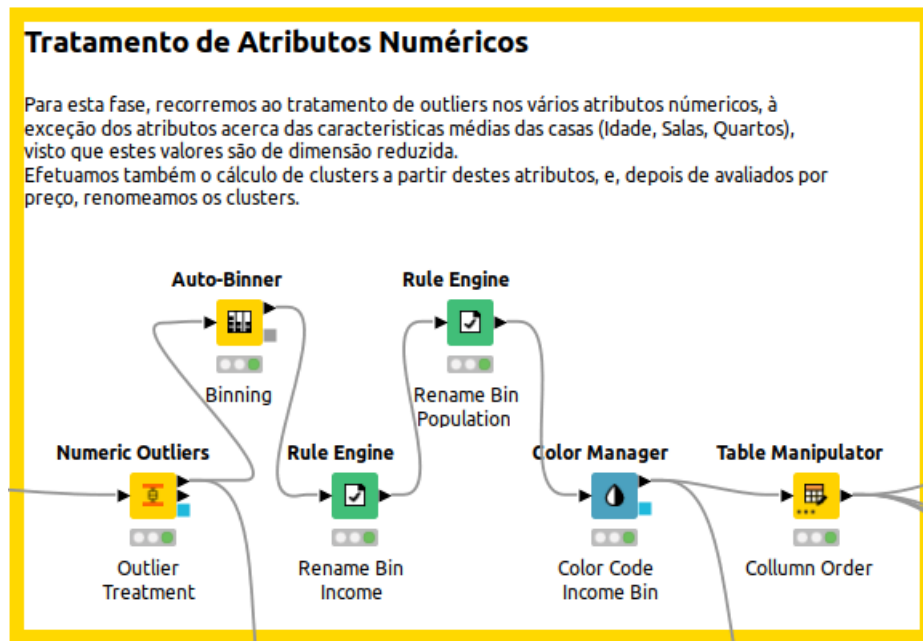


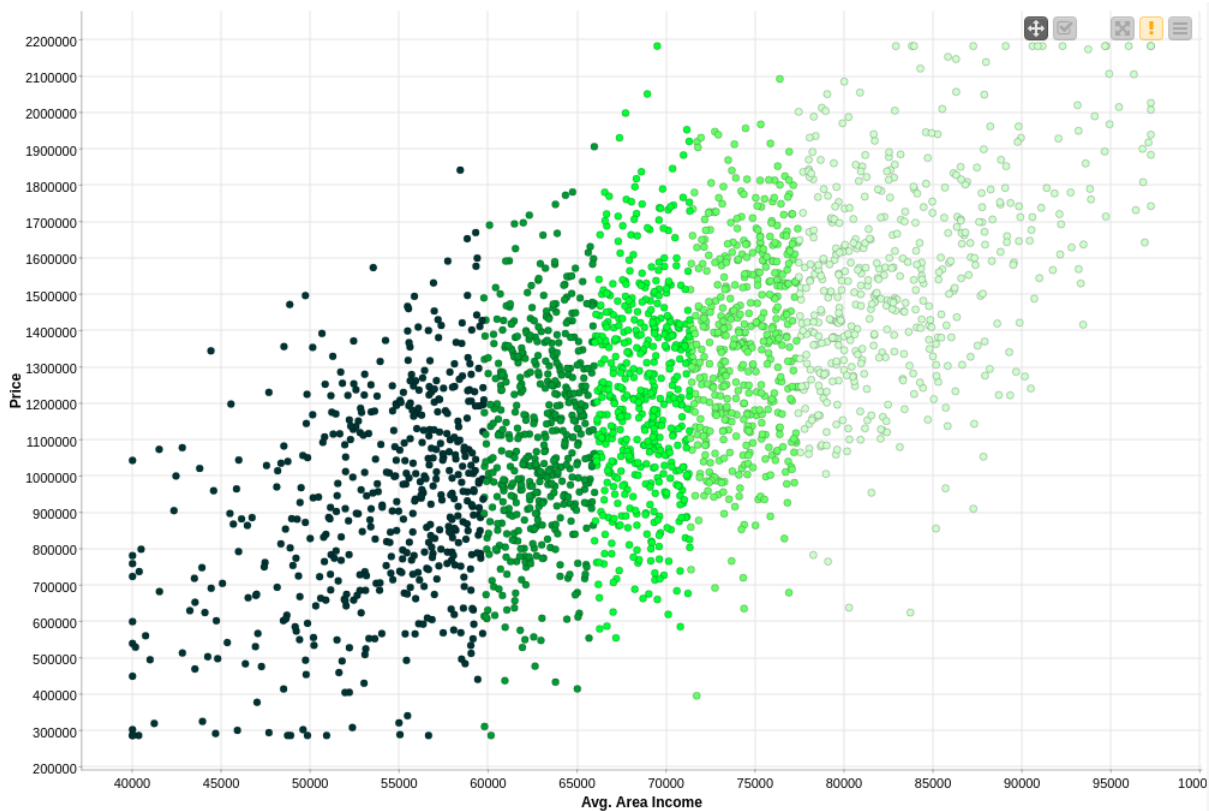
Figura 7: Tratamento de atributos numéricos

- **Numeric Outlier:** Devido ao grande número de *outliers* encontrado na fase de exploração, decidimos proceder ao tratamento destes, mas apenas nos atributos *Price*, *Avg. Area Income* e *Area Population* visto que os restantes atributos, sendo características das casas na área, possuem valores de dimensão reduzida.
- **Auto-Binner:** Para categorizar os valores de forma a facilitar a aprendizagem do modelo a ser desenvolvido, decidimos fazer o *binning* dos atributos numéricos *Avg. Area Income* e *Area Population*. Recorremos à técnica de *binning* por frequência, obtendo 5 intervalos distintos.
- **Rule Engine:** Renomeamos os *binnings* obtidos para valores mais legíveis, categorizando por classes os vários intervalos.
- **Color Manager:** Colorimos as várias categorias geradas pelo binning, de forma a criar gráficos para visualização



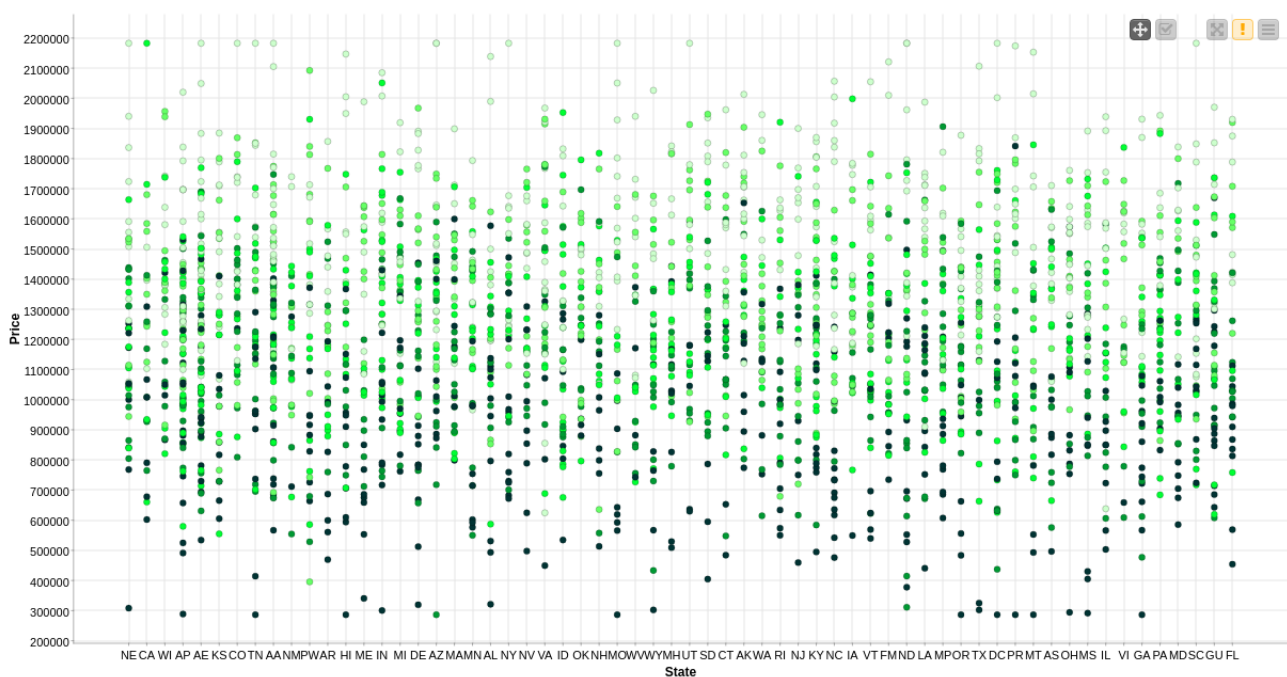
### 2.3.3 Visualização dos Dados

De modo a visualizar os nossos dados utilizamos um nodo *Scatter Plot* que relaciona o atributo *Avg. Area Income* e o atributo objetivo *Price*, obtendo os seguinte gráfico:



Como podemos observar, é possível verificar o resultado do *binning* que foi efetuado por frequência assim como a observação direta da correlação entre *Avg. Area Income* e *Price*.

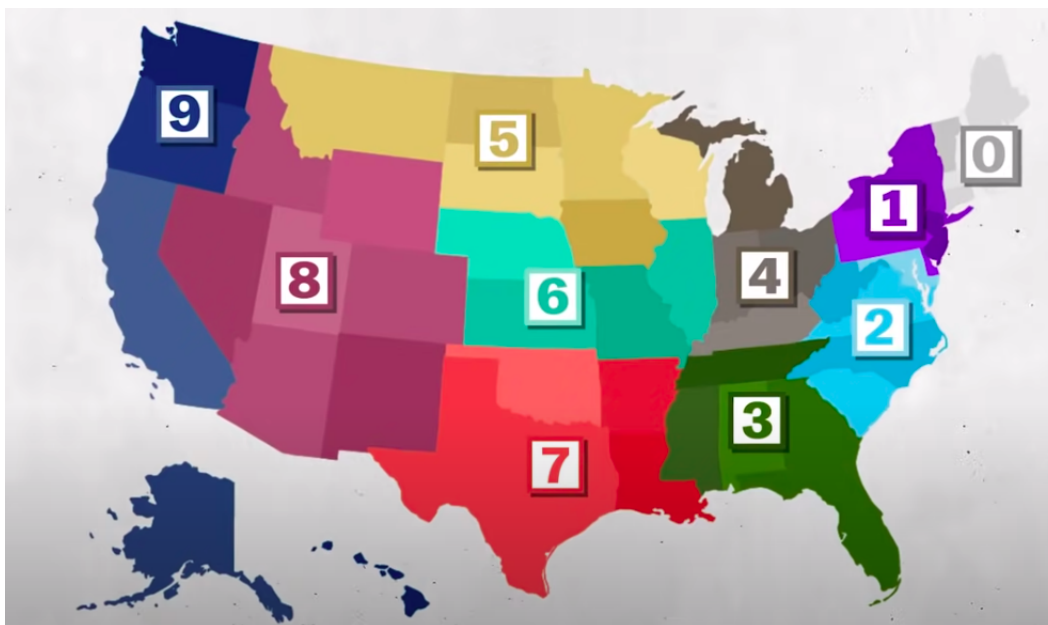
No entanto, na criação de um outro *Scatter Plot*, que pretendia relacionar os vários estados com os seus preços, encontramos algo inesperado.



Os valores representados no gráfico não contêm nenhum padrão identificável, sendo impossível obter qualquer informação a partir do estado da casa em avaliação. Face a estes resultados, decidimos explorar melhor o resultado final do tratamento de endereços obtido, analisando os zip-codes e a sua morada correspondente. Para isso, criamos um nodo *Group By* onde, agrupando pelo atributo *Zip-Code National Area* e realizando uma concatenação única dos estados, encontramos os seguintes resultados.

Row ID	5	zipCode National Area	5	Unique concatenate(State)
Row0	0			WI, AE, AP, MI, PW, IN, MN, NY, PA, AS, TX, NC, RI, HI, MH, OR, DC, GA, CO, AL, NJ, KY, ID, WY, KS, IL, MO, TN, ME, NV, AR, CT, AK, NH, WV, OH, IA, MD, NE, FM, MP, OK, AA, VA, DE, GU, AZ, VT, MS, SD, NM, ND, SC, VA, MT, FL, UT, WA, MA, CA, LA, PR
Row1	1			KS, TN, CT, WA, AA, PW, AE, NH, DC, ID, WV, CO, NM, AK, ME, MO, AZ, OK, MD, NV, ND, AS, DE, NJ, VT, NC, SD, KY, MH, AL, UT, MI, NY, VA, WI, IA, TX, OR, GA, MT, WY, FM, IL, AP, PR, MA, MS, CA, IN, MN, FL, NE, GU, AR, PA, HI, VI, LA, OH, MP, RI, SC
Row2	2			CO, AA, NM, AR, TN, MN, ID, UT, VT, KY, AK, KS, MH, DE, MI, RI, AE, HI, AP, MA, MT, AS, ME, PA, WY, WI, CA, MS, PW, NE, PR, GU, WA, IN, SC, NC, FL, OK, NY, MP, MD, OR, DC, WV, IL, AZ, AL, NH, FM, NV, GA, CT, TX, IA, VA, SD, LA, ND, MO, OH, VI, NJ
Row3	3			NE, ME, VA, WY, ID, IA, KS, CA, DC, NH, AA, MT, NV, CO, NY, PR, AZ, PA, MI, MA, AR, GA, WI, MN, RI, SC, MH, OR, AP, TX, NC, NM, NJ, MO, TN, AE, MS, HI, VI, FM, KY, AK, CT, GU, ND, LA, IL, AL, OK, WV, OH, SD, MP, PW, DE, WA, MD, AS, IN, VT, FL, UT
Row4	4			CA, AP, HI, SD, NC, OR, PR, AE, MS, ME, NE, MN, MP, CT, RI, GU, FM, LA, AA, AZ, GA, TX, NH, AS, ND, OH, VT, AR, IN, KS, AL, SC, MA, MD, PA, VA, MI, AK, DE, KY, NY, DC, IL, TN, UT, NV, NM, MO, WY, MH, NJ, ID, VI, MT, FL, WV, IA, PW, WI, WA, OK, CO
Row5	5			NE, PW, NJ, AA, DE, KY, OR, KS, IL, MA, WA, AP, WY, AK, AE, SC, GA, MS, TN, ID, MP, UT, TX, VI, OK, MD, OH, MN, ND, AR, NM, IA, PR, GU, PA, NH, VT, DC, CO, AS, NV, AZ, IN, PA, MI, HI, CT, WV, LA, AL, ME, NY, RI, CA, VA, MD, WI, FL, SD, MT, NC
Row6	6			DE, AL, NV, NH, MD, PW, NM, WY, ND, LA, AS, SD, OH, VA, AP, VI, VT, PA, DC, OR, MI, NY, WA, CO, MP, ME, OK, WV, TN, NE, MH, MA, KS, TX, NC, MD, SC, IL, AE, HI, KY, UT, CT, GU, GA, AZ, PR, CA, AR, AA, WI, FL, NJ, AK, ID, MS, FM, RI, MT, IN, IA, MN
Row7	7			PW, AE, AZ, WV, AP, MH, RI, KY, FM, KS, PR, MT, NC, ID, OK, NJ, VT, MN, GU, NY, CO, SC, AA, VI, CA, IA, MD, TN, LA, OR, HI, MO, WA, AK, UT, TX, AL, CT, MI, MP, SD, WY, NH, NE, MA, OH, DE, IL, DC, NV, IN, PA, VA, GA, FL, WI, AS, ME, ND, MS, NM, AR
Row8	8			NH, HI, AP, MA, OK, WI, MP, OR, PR, WY, SC, MI, MO, CA, IN, DC, IA, KS, NE, LA, AZ, NY, ME, DE, NJ, FL, CO, UT, AE, TN, PA, NH, KY, MS, PW, AL, ND, AA, RI, GU, OH, AR, TX, CT, MD, MT, AK, WV, VI, GA, VT, VA, IL, NC, AS, WA, ID, NV, SD, FM, MN, MH
Row9	9			AE, IN, NY, WV, AK, NH, CT, TX, PW, GA, MD, SC, PA, VI, AA, KY, MP, FL, RI, OK, VT, CO, OR, OH, NV, AP, SD, NC, FM, ID, MS, MT, MO, LA, VA, AZ, ND, TN, WY, UT, PR, NM, GU, MH, AR, WA, KS, AL, IL, DE, HI, NE, IA, MN, ME, DC, MI, CA, WI, AS, NJ, MA

Este resultados demonstram que algo de errado ocorreu no processamento de endereços, visto que, cada *Zip-code National Area* deveria representar apenas um pequeno grupo único de estados, como a seguinte imagem representa: (Video de referência)



Depois de uma análise do processamento de endereços e análise de exemplos, observamos que, na maior parte dos endereços, a conexão entre a morada da casa e o seu Zip-Code não correspondem como necessário. Logo, chegamos à conclusão que o data set contém endereços não representativos do mundo real. Apesar desta conclusão, decidimos manter o tratamento de endereços, visto que este poderá ser aplicado a endereços reais no futuro.

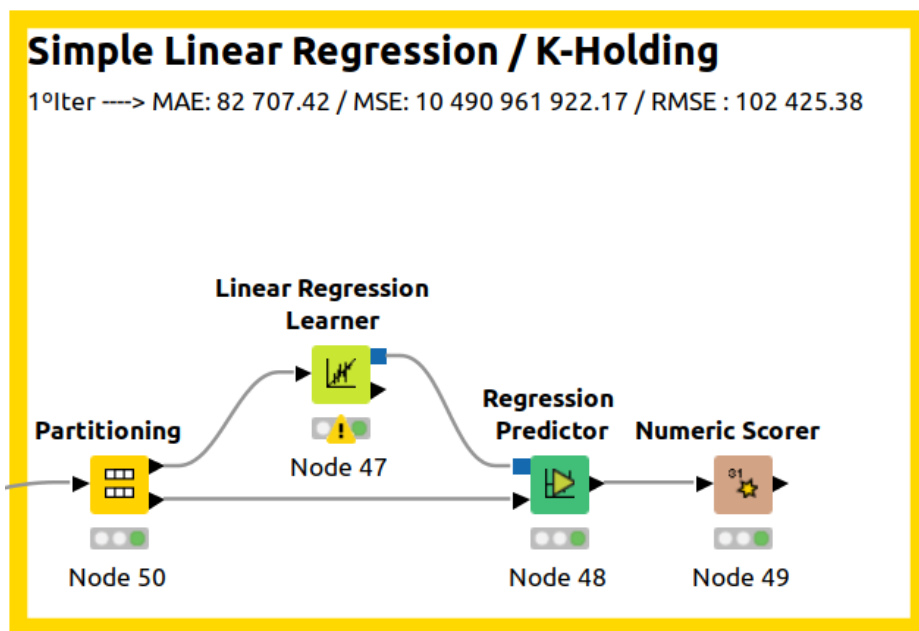
## 2.4 Modelos Utilizados

Para a criação de modelos, decidimos criar vários modelos de regressão utilizando vários algoritmos de aprendizagem e validação com o objetivo de encontrar um modelo com melhor precisão, ou seja, com métricas de erro menores. Também decidimos aplicar técnicas de redes neurais, para avaliar o seu desempenho geral na previsão deste problema.

### 2.4.1 Simple Linear Regression - K-Holding

Começamos por criar um modelo mais básico, utilizando um *partitioning* simples (*Hold-Out Validation*) com um *ratio* 80/20 entre o *data set* de treino e o *data set* de teste, utilizando o método de "Stratified Sampling" aplicado ao atributo *Avg Area Income [Binned]*, com a *random seed* estática "2022". Escolhemos este atributo como alvo para o "Stratified Sampling" pois, baseado no *Scatter Plot* anterior, concluímos que este *sampling* irá dividir de forma igual os vários preços presentes no *data set*.

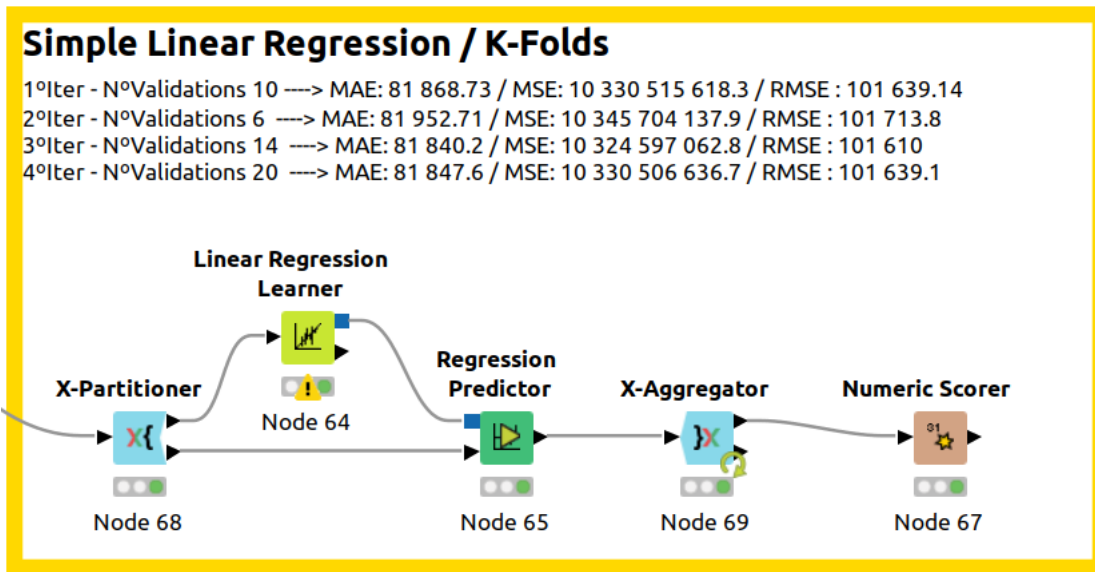
Este modelo, obteve resultados inicialmente satisfatórios, como se pode verificar na seguinte imagem, no entanto, acreditamos que será possível obter modelos mais precisos com a utilização de técnicas de validação/aprendizagem diferentes.



### 2.4.2 Simple Linear Regression - K-Folds

Depois, decidimos explorar a validação *K-folds* (*k-fold Cross Validation*). Logo, aplicamos um nodo "X-Partitioner" e um nodo "X-Aggregator", aplicando estes ao mesmo algoritmo de aprendizagem. Começamos por utilizar um número de validações igual a 10, fazendo o *partitioning* com o método de "Stratified Sampling" aplicado ao atributo alvo *Avg Area Income [Binned]*, com a *random seed* estática "2022", como no modelo anterior

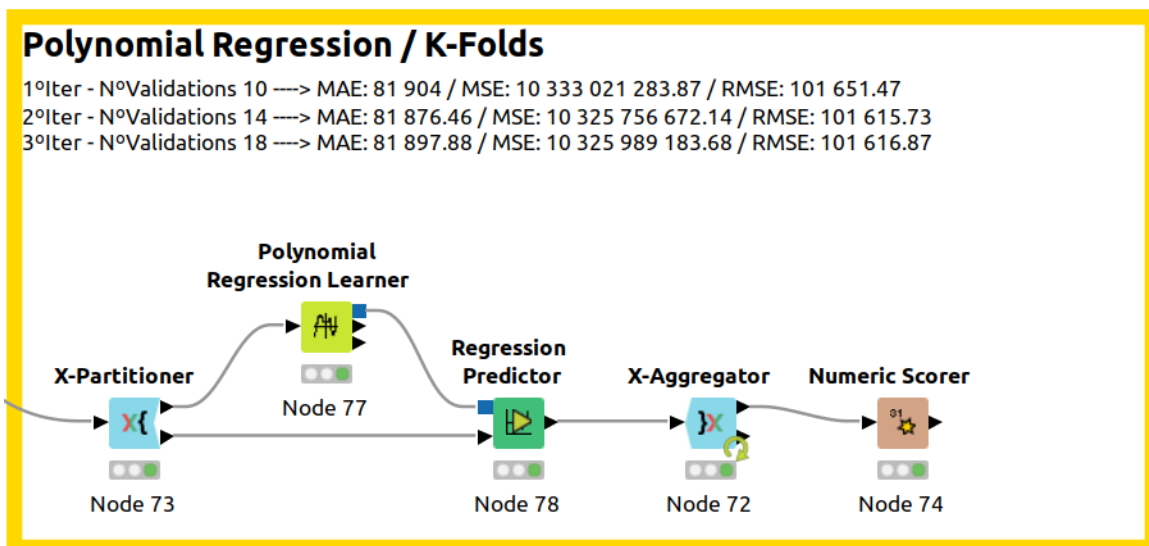
Depois de alguns ajustes no número de validações, conseguimos alcançar melhores resultados em relação ao modelo anterior, tal como se pode verificar na seguinte imagem:



### 2.4.3 Polynomial Regression - K-Folds

Com o objetivo de obter melhores resultados, decidimos modificar apenas o algoritmo de aprendizagem utilizado, escolhendo o algoritmo "Polynomial Regression".

No entanto, mesmo modificando o número de validações, não verificamos melhorias significativas em relação ao modelo anterior, como podemos verificar na seguinte imagem:



## 2.4.4 Multi Layer Perceptor

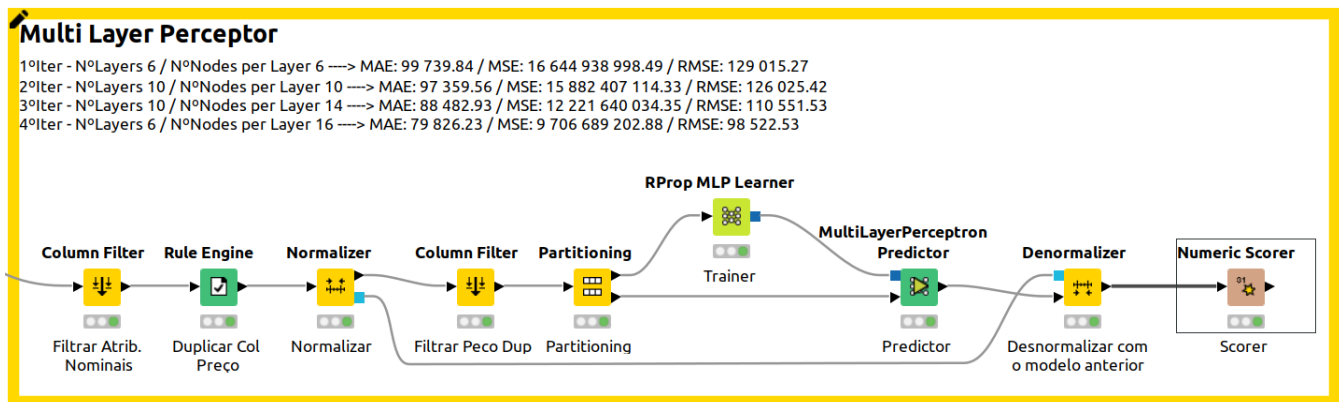
Por último, decidimos aplicar um modelo de aprendizagem baseado em redes neurais, optando por aplicar o algoritmo *Multi-layer Perceptor*. A aplicação deste apresentou alguns desafios nomeadamente a remoção de todos os atributos nominais presentes no *data set*, visto que o *MLP learner* não é compatível com estes e, sendo este o mais desafiante, a normalização dos valores e a sua posterior leitura.

Inicialmente, normalizamos os valores dos atributos numéricos de forma à rede neuronal operar normalmente, no entanto, quando obtivemos as previsões finais e avaliamos as suas métricas de erro, visto que tanto o valor do preço real das casas como o valor da previsão do seu preço estavam normalizados, os valores destes não eram aplicáveis a um contexto real.

Para resolver este problema aplicamos um nodo *Denormalizer* depois dos resultados das previsões serem calculados, de forma a obter valores "reais". Porém esta estratégia falhou pois, reparamos que, no *data set* após a desnormalização, todos os atributos numéricos à exceção da previsão do preço foram desnormalizados, resultando num *mismatch* de escalas entre os valores reais e a sua previsão.

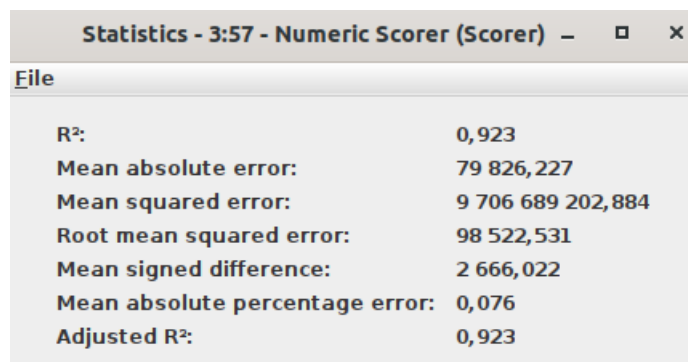
De forma a resolver esta situação, decidimos duplicar o atributo *Price*, criando uma nova coluna com o mesmo nome da coluna de previsão *Prediction (Price)*, tendo o cuidado de remover esta duplicação antes de treinar a rede neuronal, desta forma, o modelo criado pelo nodo *normalizer* irá conter a "desnormalização" da coluna de previsão. Esta solução funcionou corretamente, obtendo valores reais e de previsão não só na mesma escala, como também representativos do problema.

Com a leitura de erros corrigida, começamos por afinar a rede neuronal, alterando sequencialmente o número de camadas da rede neuronal, assim como o número de nodos por camada, obtendo, no fim, um modelo com maior precisão que os seus precedentes, como verificado na seguinte imagem, estando o grupo satisfeito com estes resultados:



## 2.5 Resultados e Conclusões

Apresentamos de seguida a tabela de estatísticas dos resultados obtidos no melhor modelo, na sua melhor iteração: (*Multi-Layer Perceptron* - 6 *Layers* - 16 *Nodes per Layer*):



File	
<b>R²:</b>	<b>0,923</b>
<b>Mean absolute error:</b>	<b>79 826,227</b>
<b>Mean squared error:</b>	<b>9 706 689 202,884</b>
<b>Root mean squared error:</b>	<b>98 522,531</b>
<b>Mean signed difference:</b>	<b>2 666,022</b>
<b>Mean absolute percentage error:</b>	<b>0,076</b>
<b>Adjusted R²:</b>	<b>0,923</b>

Estes resultados demonstram que este modelo apresenta um erro médio absoluto de 79 826 euros e um erro médio quadrado de 98 522 euros quando prevê o preço de uma casa, baseando-se apenas nos fatores da sua região. O grupo encontra-se satisfeito com este valor de erro visto que, neste *data set*, os preços das casas possuem uma variação elevada.

Acreditamos também que, caso os endereços presentes no *data set* fossem representativos no mundo real, seria possível identificar novos padrões nos quais seria possível extrair informação, permitindo previsões mais precisas.

## 3 Collision DataBase

### 3.1 Contexto e Introdução

Visto que o *data set* anterior (fornecido pelos docentes) tem como objetivo desenvolver um modelo de regressão, o grupo decidiu procurar um *data set* que tivesse como objetivo um modelo de classificação, de forma a diversificar os casos de estudo.

Assim, o *data set* extra escolhido pelo grupo tem como título "National Collision Database". Este *data set*, pertencente ao Governo Canadano, detalha todos as colisões rodoviárias ocorridas no ano 2019, contendo informação acerca dos vários veículos e entidades envolvidas num acidente assim como as condições acerca do ambiente do acidente. Mais informação no seguinte link: [National Collision Database](#)

Este *data set* é constituído por 23 *features*, sendo todas estas do tipo nominal nos quais os seus valores são principalmente números (com a exceção de alguns casos) que, inicialmente, não parecem carregar informação útil. No entanto, no mesmo website onde o *data set* é encontrado, é possível descarregar um PDF adicional (*Data Dictionary*) o qual contém as descrições de cada coluna presente no *data set*, assim como os vários significados de cada valor.

#### 3.1.1 Objetivo do modelo

Para este *data set*, decidimos que seria relevante criar um modelo no qual, através da informação de uma colisão, fosse possível prever com um bom nível de precisão a severidade de ferimentos sustidos pela pessoa envolvida, sendo este um dos atributos presentes no *data set*.

#### 3.1.2 Pré-Tratamento

Visto que os nomes originais utilizados no *data set* não são claros no seu significado, um dos primeiros pré-tratamentos que efetuamos foi a renomeação das várias colunas de forma a simplificar a sua leitura.

Outro problema encontrado com o *data set* foi a elevada dimensão deste, contendo 272 mil casos de colisões, o que leva a um elevado tempo de processamento de cada fase. Devido a isso, decidimos começar por reduzir significativamente o *data set*, recorrendo a um "**Partitioning**" inicial. Este reduz o número de casos para 80 mil, utilizando a estratégia de "**Stratified Sampling**" aplicada ao atributo objetivo (P\_ISEV).

### 3.2 Atributos do *Data Set*

Atributos <i>Data Set</i> "Collision DataBase"		
Atributo (Antes)	Atributo (Depois)	Descrição
C_CASE	Case ID	Identificador do caso da colisão
C_YEAR	Case Year	Ano da colisão
C_MNTH	Case Month	Mês da colisão
C_WDAY	Case WeekDay	Dia da semana da colisão
C_HOUR	Case Hour	Hora da colisão
C_SEV	Case Severity	Severidade da colisão
C_VEHS	Case Num_Vehicles	Número de veículos na colisão
C_CONF	Case Collision Config	Configuração da colisão
C_RCFCG	Case Road Config	Configuração da estrada
C_WTHR	Case Weather	Metereologia no dia da colisão
C_RSUR	Case Road Condition	Condições da estrada na colisão
C_RALN	Case Road Inclination	Inclinação da estrada na colisão
C_TRAF	Case Traffic	Identificador do caso da colisão
V_ID	Vehicle ID	Identificador do veículo
V_TYPE	Vehicle Type	Tipo de veículo
V_YEAR	Vehicle Year	Ano do veículo
P_ID	Person ID	Identificador da pessoa
P_SEX	Person Sex	Sexo da pessoa
P_AGE	Person Age	Idade da pessoa
P_PSN	Person Vehicle Position	Posição da pessoa no veículo
P_ISEV	Person Injury Severity	Ferimentos da pessoa
P_SAFE	Person Safety Gear	Equipamento de segurança utilizado
P_USER	Person Role	Papel da pessoa no veículo



### 3.3 Exploração e Tratamento de Dados

#### 3.3.1 Exploração de um Caso de Estudo

Para melhor entender o *data set* em questão, decidimos investigar e interpretar um caso de estudo. Escolhemos estudar o caso número 2717946 visto que este contém 3 entradas na nossa tabela, correspondendo a um acidente que envolveu 3 pessoas.

Figura 8: Caso de estudo

Row ID	I Case...	I Case ...	S Case...	S Case ...	S Case ...	S Case ...	S Case ...	S Case ...	S Case ...	S Case ...	S Case ...	S Case ...	S Case ...
Row11492	2717946	2019	2	5	6	1	3	35	2	1	1	1	3
Row11493	2717946	2019	2	5	6	1	3	35	2	1	1	1	3
Row11494	2717946	2019	2	5	6	1	3	35	2	1	1	1	3

Figura 9: Caso de estudo

S Vehicl...	S Vehicl...	S Vehicl...	S Perso...	S Perso...	S Perso...	S Perso...	S Perso...	S Perso...	S Perso...
1	1	2008	1	M	50	11	3	1	1
2	8	2016	1	M	38	11	1	2	1
3	8	2007	1	M	37	11	1	2	1

Com o auxílio do "Data Dictionary" fornecido, alcançamos a seguinte interpretação do acidente:

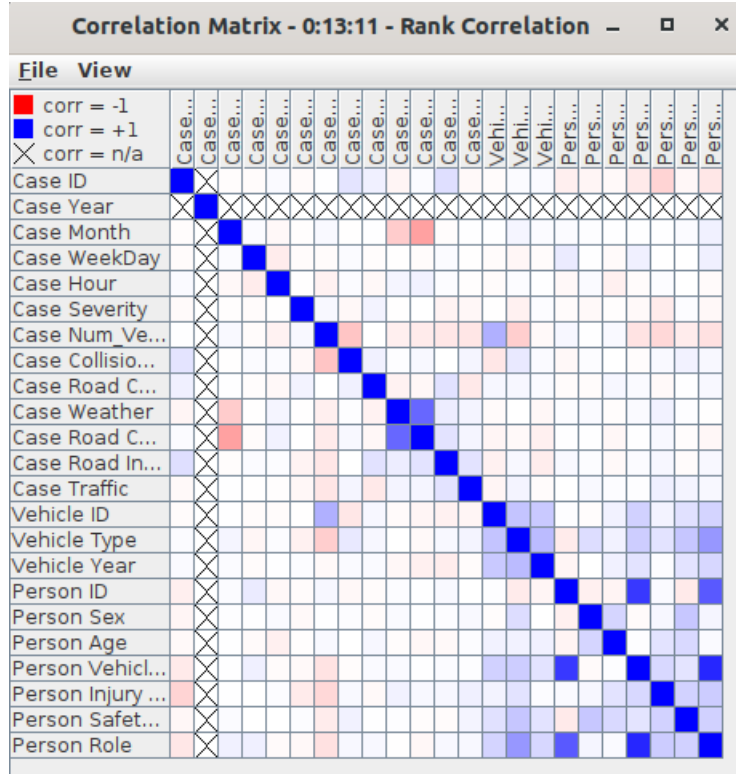
Numa sexta feira de fevereiro, o céu estava limpo e a estrada em boas condições. Por volta das 6 da manhã, um condutor de um veículo ligeiro, assim como dois condutores de tratores chocaram numa interseção que continha um sinal de STOP, havendo um choque lateral direito. Ambos os condutores dos tratores (idades de 38 e 37 anos, masculinos) não sofreram quaisquer ferimentos, no entanto, o condutor do veículo ligeiro (idade 50 anos, masculino) faleceu no acidente.

A análise deste caso de estudo demonstra que este *data set* contém bastante informação que pode ser extraída e interpretada entre cada colisão/acidente, estando assim o grupo confiante que se possa obter um modelo de previsão real e útil ao problema em questão.

### 3.3.2 Exploração de Correlação

Antes de efetuar o tratamento dos dados, começamos por um breve exploração destes, analisando especificamente a matriz de correlação entre os vários atributos:

Figura 10: Matriz de Correlação (Rank Correlation)



Apresentamos de seguida algumas das correlações mais relevantes por ordem decrescente de correlação ( $\text{Corr} \geq 0.3 \parallel \text{Corr} \leq -0.3$ ):

- Person Role — Person Vehicle Position (0.85)
- Person Vehicle Position — Person ID (0.78)
- Person Role — Person ID (0.65)
- Case Road Condition — Case Weather (0.58)
- Person Role — Vehicle (0.41)
- Case Road Condition — Case Month (-0.37)
- Vehicle ID — Case Num\_Vehicles (-0.37)

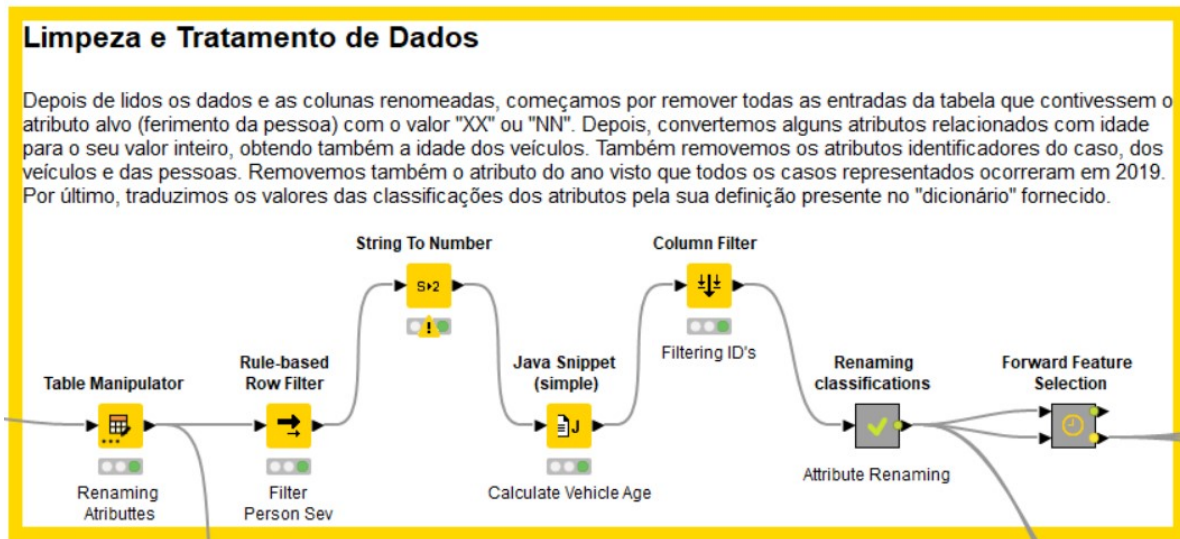
Algumas correlações como a correlação entre o papel da pessoa no veículo e a sua posição neste, assim como a correlação entre as condições de estrada e a meteorologia no momento do acidente comprovam que o *data set* contém informação válida e lógica, sendo possível realizar modelos de previsão capazes de extrair informação acerca dos acidentes

No entanto, também vimos algumas correlações de menos interesse, envolvendo os identificadores das pessoas ou do seus veículos. Este fenómeno ocorre visto que, no caso dos identificadores de pessoas, a primeira pessoa identificada num veículo é o seu condutor, havendo, por isso, esta forte correlação.

### 3.3.3 Tratamento do *Data Set*

Depois de analisado, começamos por tratar o *data set*, com o objetivo de extrair informação adicional, assim como apurar quaisquer problemas presentes nos dados. Apresentamos o tratamento efetuado:

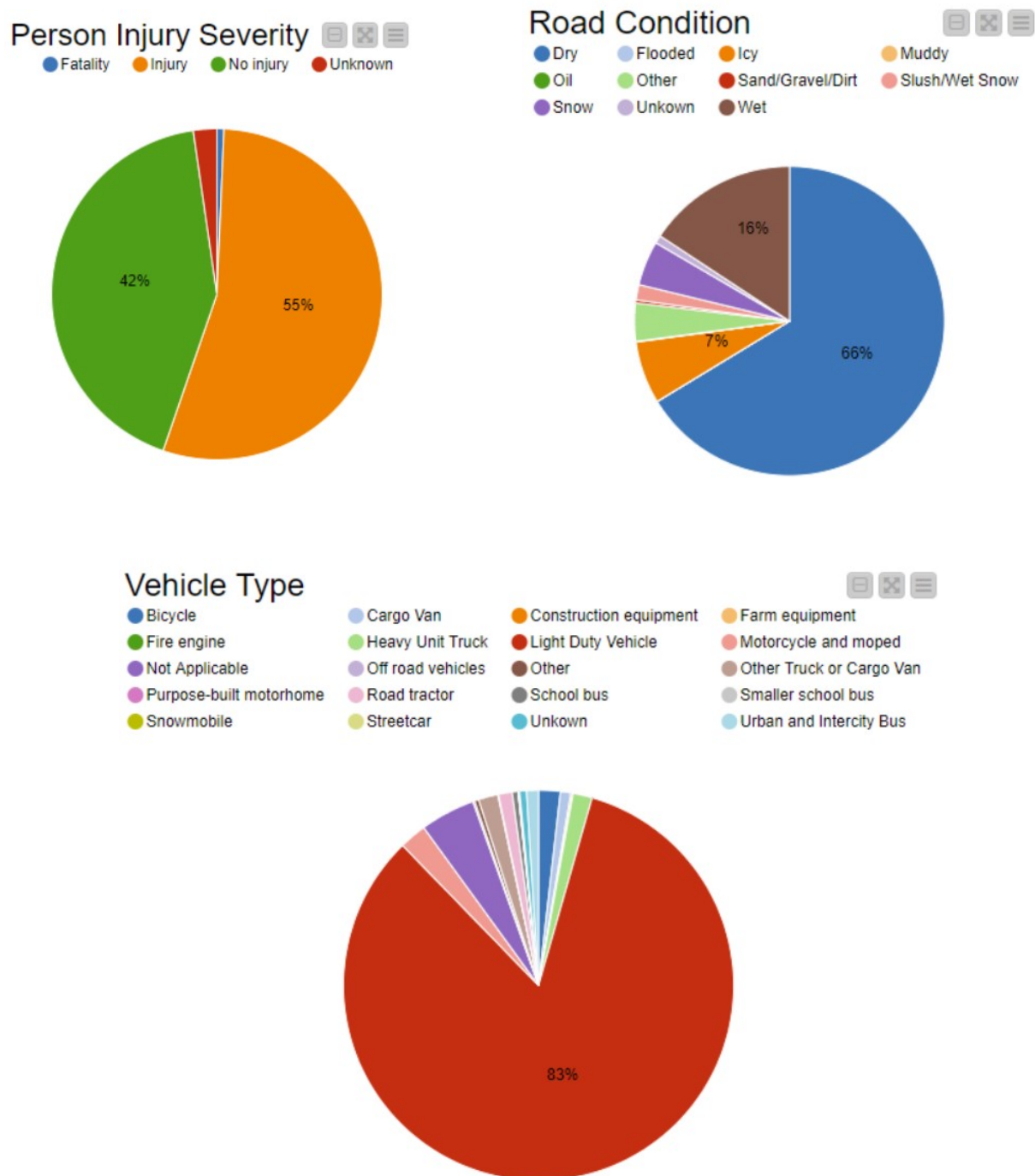
Figura 11: Tratamento de Dados



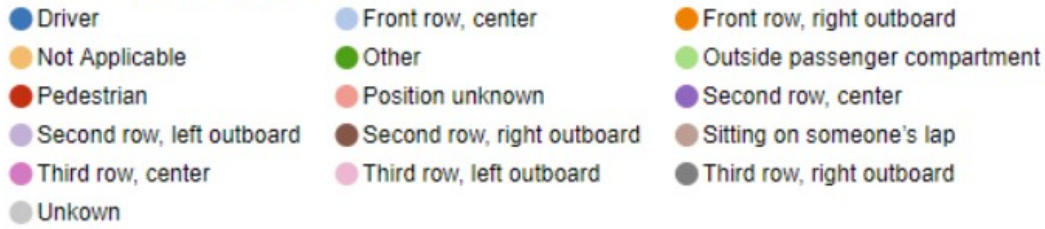
- **Table Manipulator:** Renomeação dos nomes das colunas para simplificar a leitura da tabela e futura visualização/interpretação.
- **Rule-based Row Filter:** Remoção de todos os casos da tabela que contenham os valores "XX" e "NN" no atributo objetivo. Recorremos a essa remoção pois estes valores indicam, respetivamente, a informação restrita/ofuscada e a casos que correspondem a veículos estacionados (*Dummy Entry*), não sendo relevantes para o modelo a ser desenvolvido.
- **String to Number:** Conversão dos atributos "Case Num\_Vehicles", "Vehicle Year" e "Person Age" para valores inteiros. Esta conversão é útil pois estes atributos são numéricos pela sua natureza, podendo ser possível extrair informação do número "em si", ao invés de valores de classificação independentes. Em alguns casos, podem ser gerados *Missing Values* visto que nestes o ano do veículo/idade da pessoa não foram recolhidos. Decidimos que estes não seriam tratados pois consideramos esta "falta de informação" a informação em si, que poderá ser utilizada pelo modelo.
- **Java Snippet:** Criação de uma nova coluna "Vehicle Age", calculada à partir da idade do carro e o ano do *data set*.
- **Column Filter:** Filtragem de atributos Identificadores (Case ID, Vehicle ID, Person ID), assim como remoção da coluna (Case Year), visto que todos os casos ocorrem no ano de 2019 e remoção da coluna "Vehicle Year", dado que esta informação já está contida na coluna "Vehicle Age".
- **Renaming Classifications:** Renomeação dos valores nos atributos nominais utilizando o *Data Dictionary* fornecido, visto que os valores utilizados são de difícil leitura e interpretação.

- **Foward Feature Selection:** Como este *data set* contém um número elevado de atributos, decidimos aplicar a técnica de *featuring Foward Feature Selection* para selecionar os atributos mais relevantes. Para este nodo, decidimos focar o algoritmo de seleção na métrica de precisão, escolhendo o *set* de atributos com uma elevada precisão mas também um número de atributos satisfatório.

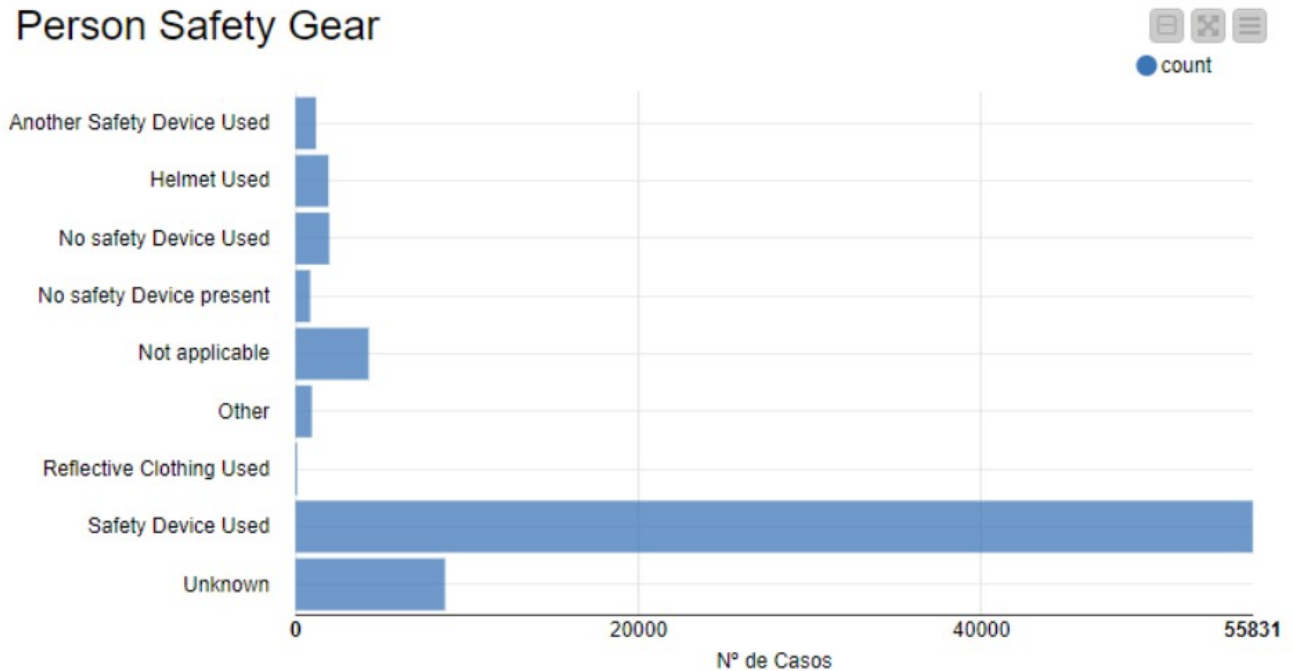
### 3.3.4 Visualização dos Dados



## Person Vehicle Position



## Person Safety Gear



A partir das figuras acima apresentadas podemos, de maneira mais perceptível, visualizar o conteúdo do *data set*, obtendo estatísticas sobre o mesmo.

Assim, pela figura relativa ao ***Person Injury Severity***, podemos concluir que os ferimentos resultantes das colisões se distribuem, maioritariamente, com o resultado de *Injury* (55% - ferimento) e *No Injury* (42% - sem ferimentos), possuindo um número muito diminuído de fatalidades.

Na figura relativa ao ***Road Condition***, podemos notar que, surpreendentemente, no momento da maior parte das colisões (66%) a estrada encontrava-se seca (*Dry*), estando molhada (*Wet*) em apenas 16% dos casos.

No que toca ao ***Vehicle Type*** dos vários veículos presentes nas colisões, a maior parte (83%) pertence à categoria *Light Duty Vehicle* que corresponde a veículos ligeiros de passageiros.

Relativamente à ***Person Vehicle Position***, que corresponde à posição da pessoa relativamente ao veículo no momento do acidente, em 68% dos casos a vítima encontrava-se na posição do condutor (*Driver*), estando na primeira fila no lugar do passageiro direito em 13% das vezes.

Por último, relativamente à ***Person Safety Gear***, ou seja, o tipo de equipamento de segurança que a pessoa estaria a utilizar no momento da colisão, a maior parte utilizava algum dispositivo de segurança (*Safety Device Used*).

## 3.4 Modelos Utilizados

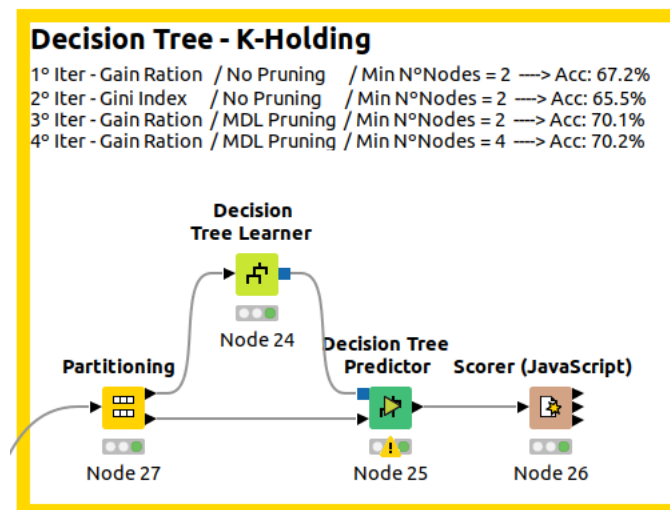
Para o desenvolvimento de modelos, decidimos criar 3 modelos de classificação (Aprendizagem por Supervisão), utilizando várias técnicas de algoritmos de aprendizagem e partição do *data set*, de modo a encontrar o modelo com melhor desempenho para o caso em estudo.

### 3.4.1 Decision Tree - K-Holding

Este modelo, sendo um dos mais simples, obteve resultados iniciais abaixo do que pretendíamos, mas, com o ajuste dos parâmetros do algoritmo, conseguimos alcançar uma métrica de *accuracy* mais satisfatória.

Efetuamos um *partitioning* 80/20 utilizando o método de "Stratified Sampling aplicado" ao atributo alvo, com a *random seed* estática "2022", de forma a dividir o *data set* de treino e de teste. (*Hold-Out Validation*)

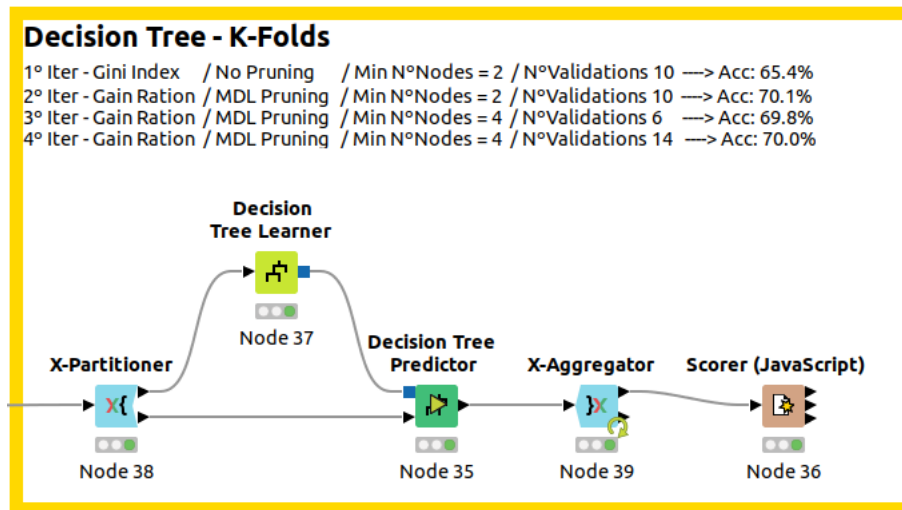
Através da análise da seguinte imagem verificamos o resultado dos testes para este modelo de acordo com os diferentes parâmetros:



### 3.4.2 Decision Tree K-Folds

De seguida, decidimos explorar a validação de *K-folds* (*k-fold Cross Validation*). Para isso aplicamos um nodo "X-Partitioner" e um nodo "X-Aggregator" aplicado ao mesmo algoritmo de aprendizagem. Começamos com os mesmos parâmetros iniciais do modelo anterior, utilizando um número de validações igual a 10, fazendo o *partitioning* com o método de "Stratified Sampling" aplicado ao atributo alvo, com a *random seed* estática "2022".

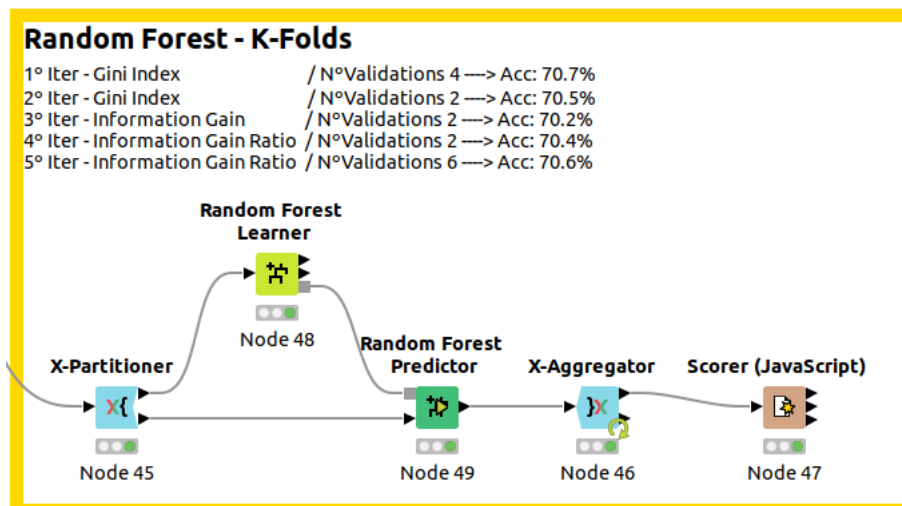
Depois de alguns ajustes quer no algoritmo, quer no número de validações, conseguimos alcançar resultados semelhantes ao modelo anterior, tal como se pode verificar na seguinte imagem:



### 3.4.3 Random Forest- K-Folds

Por último, decidimos utilizar o mesmo raciocínio anteriormente descrito, diferenciando apenas no algoritmo de aprendizagem, escolhendo o "Random Forest Learner", alterando apenas o critério de avaliação.

Devido ao elevado tempo de processamento deste algoritmo, optamos por reduzir significativamente o número de validações a ser efetuada. Obtivemos, assim, um modelo com resultados de precisão satisfatórios, sendo estes os mais elevados para a resolução deste problema, como se pode verificar:





### 3.5 Resultados e Conclusões

Apresentamos, de seguida, a tabela de confusão dos resultados obtidos no melhor modelo, na sua melhor iteração: (*Random Forest K-Folds: Gini Index* - 4 Validações)

#### Scorer View

Confusion Matrix



	Fatality (Predicted)	Injury (Predicted)	No injury (Predict...	Unknown (Predic...	
Fatality (Actual)	328	153	37	0	63.32%
Injury (Actual)	83	32145	9130	8	77.71%
No injury (Actual)	37	12617	19479	4	60.61%
Unknown (Actual)	1	43	56	1637	94.24%
	73.05%	71.50%	67.87%	99.27%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa ( $\kappa$ )	Correctly Classified	Incorrectly Classified
70.74%	29.26%	0.431	53589	22169

Como podemos ver, é possível observar alguns padrões presentes no nosso modelo, em particular, nas previsões de resultados "Unknown" e nos erros entre as previsões "Injury" e "No Injury".

Observando a coluna de previsões da classificação "Unknown" podemos concluir que o modelo desenvolvido tem uma elevada precisão no que toca às previsões "Unknown" (com uma taxa de precisão de 99.27%). Concluimos que este fenómeno ocorre devido aos vários padrões encontrados nos atributos, visto que quando o ferimento da pessoa na colisão é desconhecido, muitas vezes representa uma pessoa que fugiu do local do acidente, não sendo outras informações como o seu veículo e o seu papel recolhidos.

Também observamos uma fraqueza do modelo nas previsões entre pessoas que sofreram ferimentos e aquelas que escaparam ilesas, tendo o nosso modelo "trocado" várias vezes estes dois resultados. (Por exemplo, apenas acertou 67.87% das previsões "No Injury").

Apesar destes resultados, a nossa equipa está confiante de ter desenvolvido um modelo adequado ao problema, com a capacidade de prever de forma coerente os resultados de ferimentos de pessoas envolvidas num acidente.

## 4 Conclusão

Após a realização deste projeto, o grupo encontra-se satisfeito com o trabalho desenvolvido, tendo sido alcançados todos os objetivos estabelecidos quer pelos docentes, como pelo próprio grupo. No entanto, acreditamos que alguns dos modelos apresentados poderão ser melhorados no futuro, através de um tratamento mais aprofundado e metódico.

Relativamente aos *data sets* trabalhados:

- No *data set* **Housing Regression**, o grupo desenvolveu modelos de regressão a partir dos dados inicialmente tratados e analisados, obtendo um modelo capaz de prever o preço de uma dada casa. Consideramos ter obtido modelos com uma métrica de erro razoável.
- No *data set* **Collision Database**, desenvolvemos modelos de classificação a partir das várias condições presentes num acidente, obtendo um modelo capaz de prever os ferimentos de uma pessoa envolvida nesse acidente. No fim deste desenvolvimento, acreditamos ter obtido uma série de modelos com um bom nível de precisão.

Em suma, aprofundamos o conhecimento lecionado nas aulas teóricas e práticas da unidade curricular em relação às várias fases de exploração, tratamento e criação de modelos de aprendizagem através de algoritmos de *machine learning*, aumentando também a nossa capacidade de utilização e exploração do ambiente de desenvolvimento *Knime*, que permitiu a realização deste projeto.