



Universidade do Minho
Escola de Engenharia

MESTRADO EM ENGENHARIA DE TELECOMUNICAÇÕES E
INFORMÁTICA

INTELIGÊNCIA ARTIFICIAL

CONCEÇÃO E OTIMIZAÇÃO DE MODELOS DE MACHINE LEARNING

RELATÓRIO DE PROJETO INDIVIDUAL

Camila Pinto - PG53712

Conteúdo

1	Introdução	3
2	Dataset	4
2.1	Atributos	4
2.2	Exploração de Dados	6
2.2.1	Atributos region name e council area	6
2.2.2	Atributos car, yearbuilt e building area	7
2.2.3	Atributo propriety count, postcode e adress	7
2.2.4	Atributos latitude e longitude	8
2.2.5	Atributos price, landsize e builing area	8
2.2.6	Atributo price,rooms, distance	9
2.2.7	Atributo type e method	10
2.3	Preparação de dados e Modelação	11
2.3.1	Column Filter	11
2.3.2	Numeric Outliers	12
2.3.3	Missing Values	12
2.3.4	Group By	12
2.3.5	Normalizer	13
2.3.6	Auto-Binner	13
2.3.7	Rule Engine	13
2.4	Modelo e Resultados	14
2.4.1	Decision Tree	14
2.4.2	Random Florest	16
2.4.3	Linear Regression	18
3	Conclusão	19

Lista de Figuras

1	Resultados do módulo CrossTab entre councilarea e regioname.	6
2	Amostra da tabela de estatísticas.	7
3	Características do atributo postcode.	7
4	Características do atributo propertycount.	7
5	Características do atributo adress.	8
6	Características dos atributos latitude e longitude.	8
7	Resultado da visualização do diagrama de caixa.	8
8	Correlações Lineares entre os atributos.	9
9	Gráfico de barras do atributo type.	10
10	Gráfico de barras do atributo method.	10
11	Tratamento dos dados.	11
12	Atributos retirados.	11
13	Tratamento de outliers.	12
14	Configuração Auto-Binner.	13
15	Configuração Rule Engine.	13
16	Configuração Partitioning para modelo Decision Tree e Random Florest.	14
17	Modelo de Aprendizagem Decision Tree.	14
18	Configuração Decision Tree Learner.	15
19	Resultados do modelo baseado em Decision Tree.	15
20	Resultados do tuning ao modelo baseado em Decision Tree.	16
21	Modelo de Aprendizagem Random Florest.	16
22	Configuração Random Florest.	17
23	Resultados do modelo Random Florest.	17
24	Modelo de aprendizagem Linear Regression.	18
25	Resultados do modelo Linear Regression.	18

1 Introdução

O setor imobiliário, por natureza dinâmica reflete as condições socioeconômicas de uma região, refletindo não apenas as preferências dos compradores e vendedores, mas também as nuances econômicas e demográficas. Neste contexto, proponho um estudo aprofundado do mercado imobiliário de Melbourne, com o objetivo específico de analisar os dados e prever os preços de venda das propriedades.

Antes de mergulhar nos detalhes específicos do mercado imobiliário de Melbourne, é crucial observar lições aprendidas em outras regiões, especialmente em regiões do nosso próprio país, onde no geral os preços e rendas das residências têm vindo a aumentar significativamente, tornando-se uma preocupação crescente para a comunidade portuguesa. Podemos destacar essa situação ao observar o paralelo existente com o mercado em Lisboa, Portugal, onde fatores como localização, infraestrutura e demanda exercem influência substancial sobre os preços das propriedades.

Por esse motivo, a análise destes dados pode servir como a base fundamental para a construção de modelos preditivos capazes de antecipar e prever tendências no mercado imobiliário. Esses *insights* abrangem a identificação de padrões de comportamento do mercado, a previsão de demandas futuras, entre outros.

Para atingir o objetivo de previsão de preços, adotarei a metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Esta abordagem estruturada compreende etapas como: compreensão do negócio, compreensão dos dados e da sua qualidade, preparação dos dados (seleção dos atributos e limpeza), modelação, avaliação e, por fim, implementação.

O conjunto de dados (*dataset*) proposto contém informações abrangentes sobre o setor imobiliário em Melbourne, compreendendo cerca de 21 atributos, variando entre dados do tipo string, double e integer. Para a modelagem, propõe-se a utilização de dois modelos entre os seguintes: Decision Tree, Random Forest, Clustering (k-means) e Linear Regression.

2 Dataset

2.1 Atributos

O dataset apresenta os seguintes atributos:

- Suburb - nome do subúrbio,
- Address - morada,
- Rooms - Número de quartos,
- Type - br - quarto(s); h - casa, chalé, vila, geminada; u - apartamento, duplex; t - casa em condomínio; dev site - terreno para desenvolvimento; o res - outras residenciais,
- Price - Preço em dollars,
- Method - S - propriedade vendida; SP - propriedade vendida antecipadamente; PI - propriedade não vendida em leilão; PN - vendida antecipadamente sem divulgação; SN - vendida sem divulgação; NB - sem oferta; VB - oferta do vendedor; W - retirada antes do leilão; SA - vendida após o leilão; SS - vendida após o leilão sem divulgação de preço. N/A - preço ou lance mais alto não disponível,
- SellerG - Agente imobiliário,
- Date - Data de venda,
- Distance - Distância do CBD(central business district),
- Postcode - código postal,
- Bedroom - quantidade de quartos.
- Bathroom - número de casas de banho.
- Car - quantidade de vagas para os carros.
- Landsize - tamanho do terreno.
- BuildingArea - tamanho do edifício.
- YearBuilt - ano de construção,

- CouncilArea - concelho,
- latitude,
- longitude,
- Regionname - Nome da região,
- Propertycount - quantidade de propriedades existentes no subúrbio.

2.2 Exploração de Dados

De modo a perceber a qualidade dos dados, procedeu-se à exploração destes através da utilização de alguns componentes de análise estatística nomeadamente o Data Explorer, Statistics, Crosstab, Linear Correlation, bem como componentes de visualização gráfica, tal como Box Plot e Bar Chart.

2.2.1 Atributos region name e council area

A partir da análise do dataset verifica-se que os atributos region name e council area estavam relacionados. Assim sendo, utilizou-se o módulo Crosstable e foi possível analisar a distribuição por frequência entre estas duas colunas do dataset. Analisando a figura nota-se que os atributos correspondem à mesma informação, visto que as frequências são em muitos casos 100%, e não em todos devido à elevada quantidade de missing values.

Cross Tabulation of CouncilArea by Regionname								
Frequency Row Percent	Eastern Metropolitan	Eastern Victoria	Northern Metropolitan	Northern Victoria	South-Eastern Metropolitan	Southern Metropolitan	Western Metropolitan	W...
?	242 17,6771%	9 0,6574%	336 24,5435%	15 1,0957%	125 9,1308%	355 25,9313%	280 20,4529%	
Banyule	478 80,4714%		116 19,5286%					
Bayside						489 100%		
Boroondara	1 0,0862%					1 159 99,9138%		
Brimbank							424 100%	
Cardinia		8 100%						
Casey		11 28,9474%			27 71,0526%			
Darebin			934 100%					
Frankston		8			45			

Figura 1: Resultados do módulo CrossTab entre councilarea e regioname.

2.2.2 Atributos car, yearbuilt e building area

A partir da análise do módulo Statistics verificou-se que os atributos mencionados possuem um elevado número de missing values. Juntamente com a visualização dos resultados do Linear Correlation verifiquei que o atributo yearbuilt tem uma correlação negativa com a variável price, pelo que deste modo a sua inclusão não seria benéfica.

Statistics

Rows: 21 | Columns: 9

Name	Type	# Missing values	Minimum	Maximum	25% Quantile	50% Quantile (M...	75% Quantile	Mean
Date	String	0	⊖	⊖	⊖	⊖	⊖	⊖
Distance	Number (double)	0	0	48.1	6.1	9.2	13	10.138
Postcode	Number (double)	0	3,000	3,977	3,044	3,084	3,148	3,105.302
Bedroom2	Number (double)	0	0	20	2	3	3	2.915
Bathroom	Number (double)	0	0	8	1	1	2	1.534
Car	Number (double)	62	0	10	1	2	2	1.61
Landsize	Number (double)	0	0	433,014	177	440	651	558.416
BuildingArea	Number (double)	6450	0	44,515	93	126	174	151.968
YearBuilt	Number (double)	5375	1,196	2,018	1,940	1,970	1,999	1,964.684
CouncilArea	String	1369	⊖	⊖	⊖	⊖	⊖	⊖
Latitude	Number (double)	0	-38.183	-37.409	-37.857	-37.802	-37.756	-37.809
Longitude	Number (double)	0	144.432	145.526	144.93	145	145.058	144.995

Figura 2: Amostra da tabela de estatísticas.

2.2.3 Atributo propriety count, postcode e adress

Através do módulo Data Explorer, identificámos uma alta variância nas variáveis, propriety count e postcode, também notámos que a variável adress possui um unique value > 1000 . Desta forma, introduzir estes dados no modelo não tem grande importância, além de que pode levar a problemas de overfitting.


 Postcode	<input type="checkbox"/>	3000	3977	3105.302	90.677	8222.312	4.076
--	--------------------------	------	------	----------	--------	----------	-------

Figura 3: Características do atributo postcode.

 Propertycount	<input type="checkbox"/>	249	21650	7454.417	4378.582	19171978.332	1.069
---	--------------------------	-----	-------	----------	----------	--------------	-------

Figura 4: Características do atributo propertycount.

Address	<input type="checkbox"/>	0	>1000	5 Charles St, 21 Hatherley Gr, 7 Wallace St, 41 Helston St, 3 Clive Ct, [...], 166 Gipps St, 5/95 Balwyn Rd, 45 Bloomfield Rd, 6 Lovett Dr, 25 Kalimna St	Not all nominal values calculated.
---------	--------------------------	---	-------	---	---------------------------------------

Figura 5: Características do atributo address.

2.2.4 Atributos latitude e longitude

Em contrapartida estes atributos apresentam uma pequena variância e um baixo desvio padrão. Deste modo, como se trata de coordenadas pode-se assumir que sejam na mesma região.

<input checked="" type="radio"/> Latitude	<input type="checkbox"/>	-38.183	-37.409	-37.809	0.079	0.006	-0.427	1.573	-513448.973
<input checked="" type="radio"/> Longitude	<input type="checkbox"/>	144.432	145.526	144.995	0.104	0.011	-0.211	1.759	1969035.036

Figura 6: Características dos atributos latitude e longitude.

2.2.5 Atributos price, landsize e builing area

Com o auxílio do módulo Data Explorer, seguido do módulo para visualização Box Plot, foi possível identificar a presença de uma grande quantidade de outliers nestes atributos. Na figura 7 podemos ver os outliers no caso da variável price.

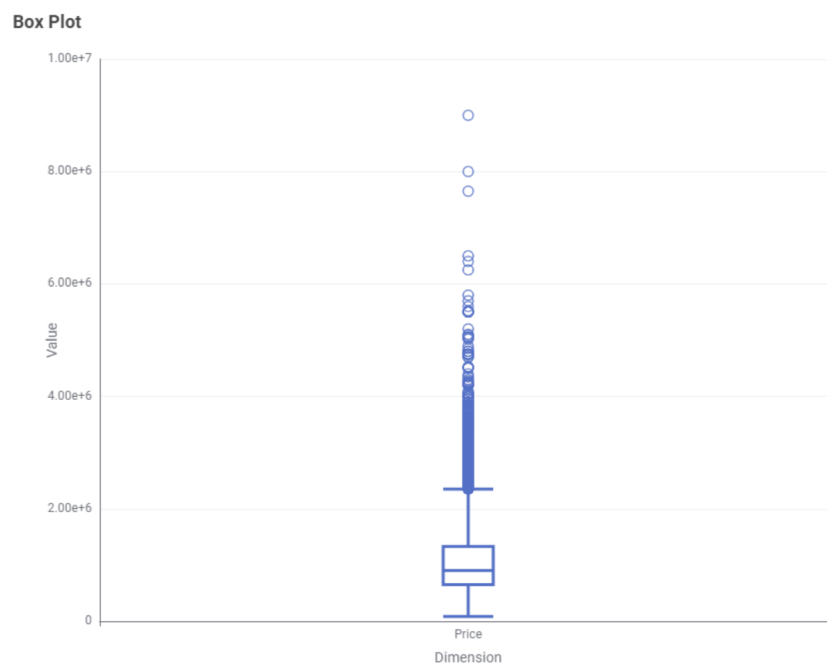


Figura 7: Resultado da visualização do diagrama de caixa.

2.2.6 Atributo price,rooms, distance

Um dos principais objetivos ao analisar estes dados é investigar possíveis correlações entre os atributos em questão. Ao utilizar o módulo Correlação Linear, pude identificar algumas associações notáveis, destacando a relação significativa entre o price e o rooms(0.4966). Isto faz sentido no contexto do mercado imobiliário, pois, à medida que o número de divisões aumenta, o preço da propriedade também tende a subir. Esta tendência é observada de forma semelhante para os atributos bedroom(0,476),bathrooms(0,467). Concluo, assim, que a infraestrutura de uma habitação está correlacionada positivamente com o preço.

Por outro lado, o atributo distance apresenta uma correlação negativa(-0.1625). Isto significa que, à medida que a distância até ao Central Business District aumenta, o preço da propriedade tende a diminuir. Esta correlação reflete adequadamente as dinâmicas do mercado imobiliário, onde o aumento da distância de pontos com recursos como supermercados, postos de trabalho, etc., contribui para a diminuição do interesse de compra.

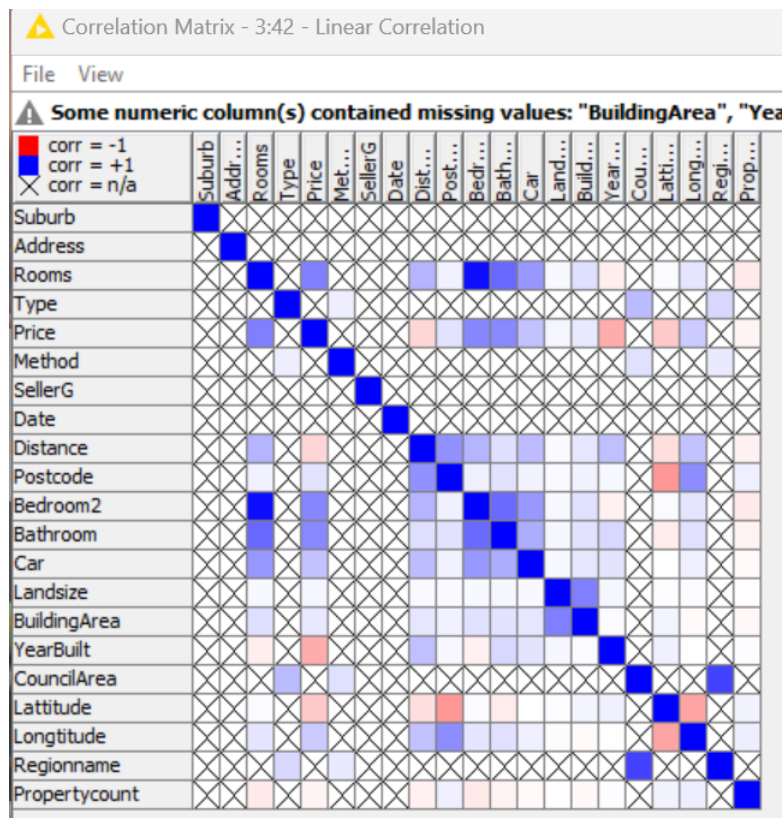


Figura 8: Correlações Lineares entre os atributos.

2.2.7 Atributo type e method

Cada propriedade apresenta um tipo, e conseqüentemente cada venda apresenta um método. Nas próximas figuras conseguimos ver os tipos e métodos que existem e a sua quantidade no dataset.

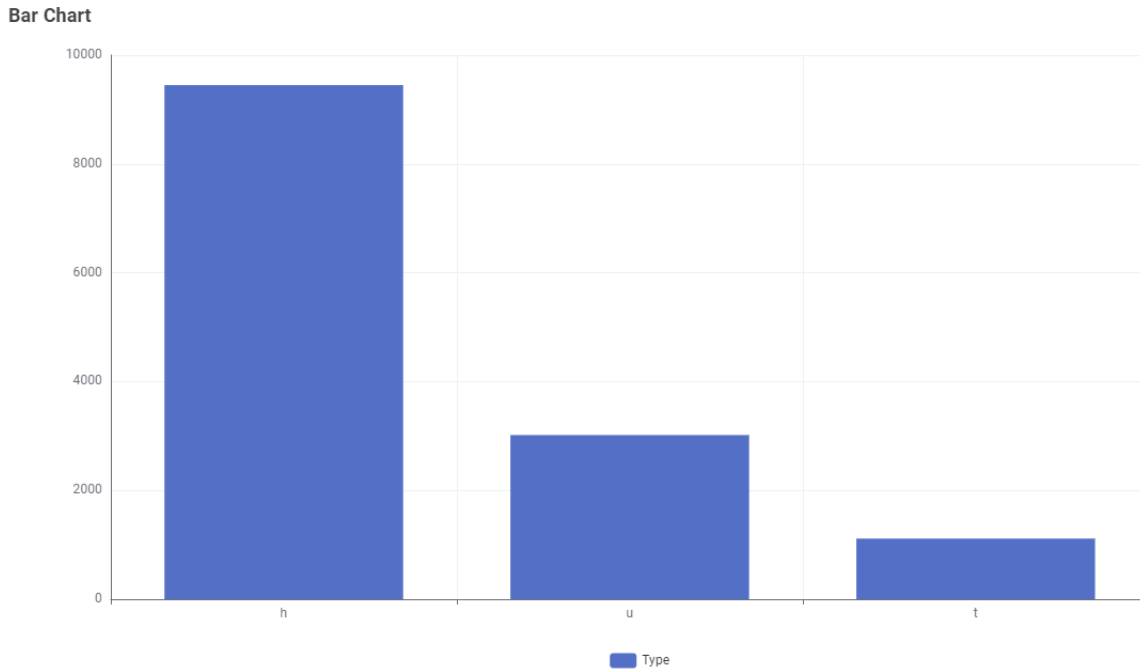


Figura 9: Gráfico de barras do atributo type.

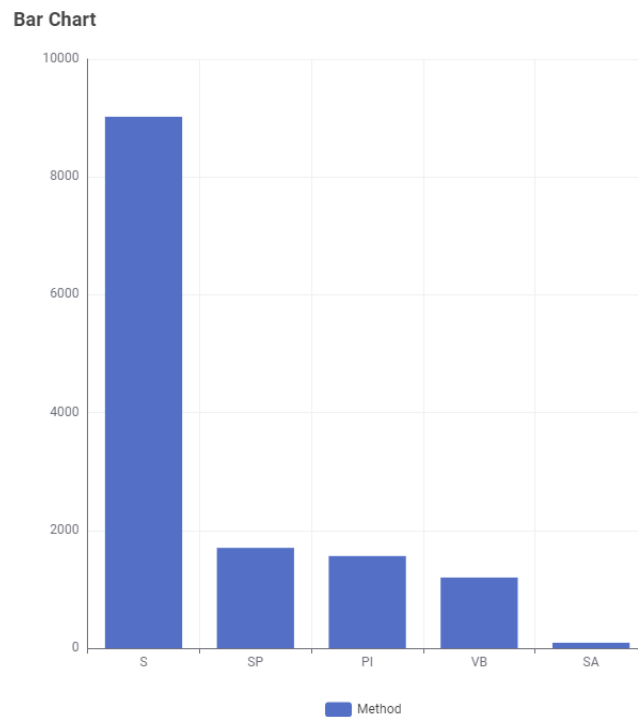


Figura 10: Gráfico de barras do atributo method.

2.3 Preparação de dados e Modelação

Com base na etapa da exploração dos dados observou-se que era necessário tratar os dados de modo a melhorar a qualidade do dataset.

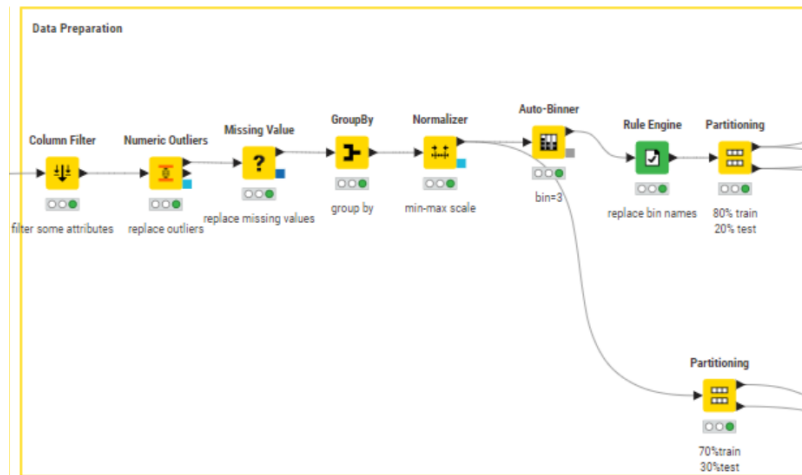


Figura 11: Tratamento dos dados.

2.3.1 Column Filter

De forma a colocar o dataset com melhor qualidade aplicou-se o módulo Column Filter. Este módulo permite retirar certos atributos à tabela. Assim, as colunas que foram retiradas podem ser observadas na figura 12. A decisão de remover esses atributos baseou-se em critérios específicos, incluindo a presença de um número considerável de valores ausentes, alta variabilidade e redundância de informações em relação a outros atributos. Estas decisões foram testadas de modo a perceber qual contribuía para o melhor desempenho do modelo.

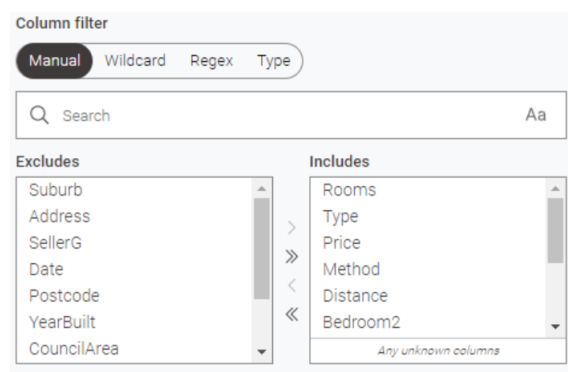


Figura 12: Atributos retirados.

Além dos representados na imagem, retirou-se Latitude, Longitude e PropertyCount.

2.3.2 Numeric Outliers

Como vimos anteriormente, as variáveis apresentam outliers, assim temos de tratar os mesmos. Este nó deteta e trata os valores atípicos para cada uma das colunas selecionadas individualmente através do intervalo interquartil (IQR). Optei por tratar os atributos que apresentam valores mais extremos e contribuem para uma melhoria do modelo.

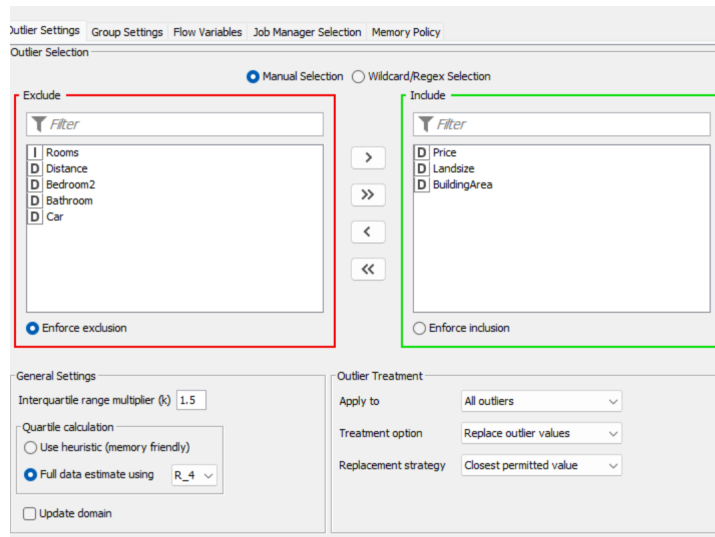


Figura 13: Tratamento de outliers.

2.3.3 Missing Values

Como observado no módulo Statistics, identificámos a presença de missing values nos dados, optamos por excluir o atributo "YearBuilt", sendo um dos principais motivos a alta incidência de valores ausentes. Além disso, notamos a presença de valores ausentes nas variáveis "Building Area" (tipo double) e "Car" (tipo integer). Para lidar com esses valores ausentes, aplicamos o módulo Missing Values, tal que:

Number (integer): Mediana

Number (double): Mediana

2.3.4 Group By

Tendo em conta que no setor imobiliário, as características da casa têm uma influência significativa no preço, agrupei todas as características relevantes com a média do atributo preço. Para tal utilizei o módulo group by.

2.3.5 Normalizer

Como os atributos apresentam escalas muito diferentes, e para treinar o modelo de Regressão Linear é benéfico colocar todas as variáveis na mesma escala, utilizei o Normalizer, para colocar na escala [0,1].

2.3.6 Auto-Binner

No módulo Auto-Binner procedeu-se à uniformização dos dados. Deste modo, foi possível agrupar dados numéricos em intervalos, denominados de bins. O número de bins a utilizar foi sendo mudado ao longo do desenvolvimento do projeto até encontrar um valor, no qual se atingiu o melhor resultado, neste caso, 3 bins.

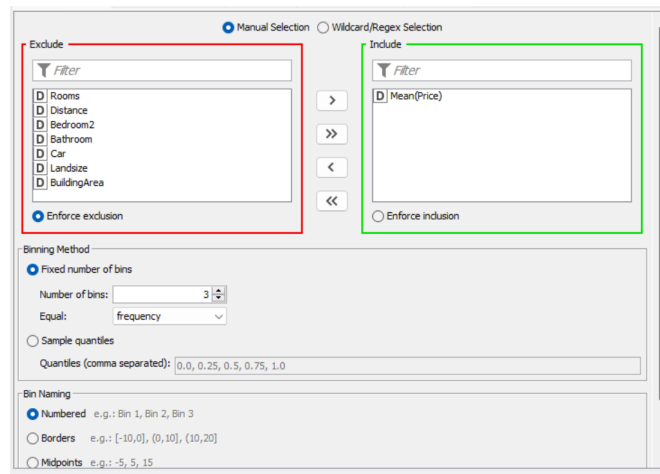


Figura 14: Configuração Auto-Binner.

2.3.7 Rule Engine

Neste módulo procedeu-se a alteração da denominação dos bins do atributo *Mean(price)* para uma certa categoria. A denominação das categorias usadas foram “Low, Medium e High”. Também foi selecionada a opção de substituir a coluna do price.

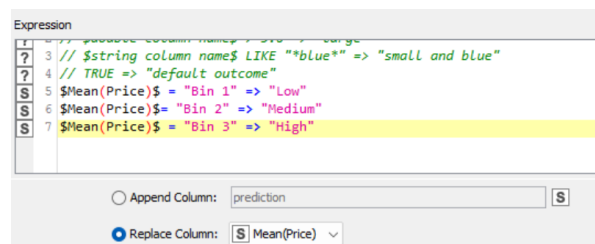


Figura 15: Configuração Rule Engine.

2.4 Modelo e Resultados

Para o desenvolvimento de modelos de aprendizagem foram utilizados os modelos: Decision Tree, Random Florest e o Linear Regression.

Para alimentar o learner, utilizou-se o módulo **Partitioning** de forma a dividir o dataset em dois. No caso dos modelos Linear Decision Trees, 80% train e 20% test, no caso do modelo Linear Regression 70% train e 30% test.

De modo a homogeneizar a amostra foi usado o modo *stratified sampling*, assim foram testados vários atributos de modo a selecionar o que iria fornecer melhor desempenho ao modelo. Optou-se pelo *type*, cuja diferença do segundo melhor classificado é 1,94%.

Para o modelo Regression Linear, optei por usar o modo *Draw Randomly*.

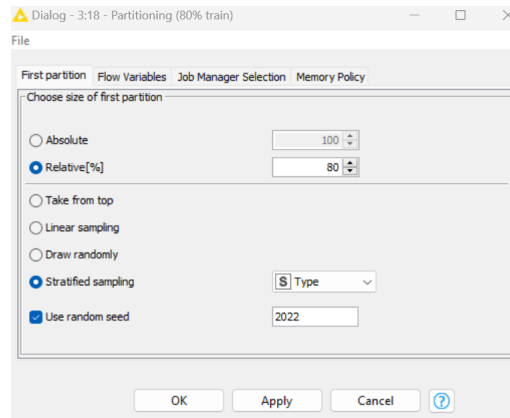


Figura 16: Configuração Partitioning para modelo Decision Tree e Random Florest.

2.4.1 Decision Tree

Na figura 17 pode-se observar a implementação do modelo de aprendizagem Decision Tree.

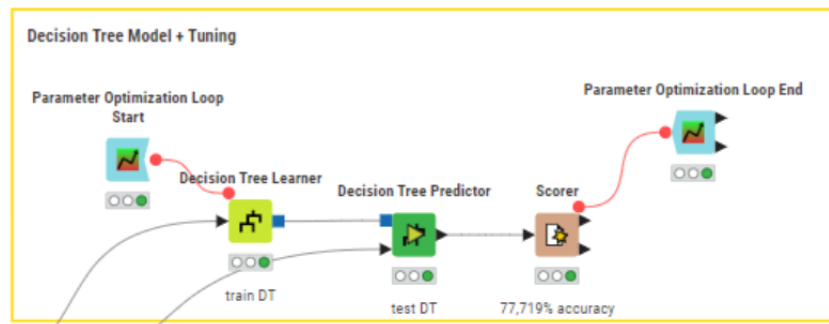


Figura 17: Modelo de Aprendizagem Decision Tree.

Após vários testes, concluiu-se que a melhor configuração do node coincide com a utilização do Gini index e do MDL.

Options PMMLSettings Flow Variables Job Manager Selection

General

Class column Mean(Price) ▾

Quality measure ▾

Pruning method ▾

☒ Reduced Error Pruning

Min number records per node ▴ ▾

Number records to store for view ▴ ▾

☒ Average split point

Number threads ▴ ▾

☒ Skip nominal columns without domain information

Root split

☐ Force root split column

Root split column BuildingArea ▾

Binary nominal splits

☐ Binary nominal splits

Max #nominal ▴ ▾

Figura 18: Configuração Decision Tree Learner.

Assim, obteve-se uma *accuracy* de 77,19%.

Confusion Matrix - 3:39 - Scorer (75,2% accuracy)

File Hilite

Mean(Price...	Medium	Low	High
Medium	548	155	159
Low	122	749	9
High	152	7	747

Correct classified: 2 044 Wrong classified: 604

Accuracy: 77,19% Error: 22,81%

Cohen's kappa (κ): 0,658%

Figura 19: Resultados do modelo baseado em Decision Tree.

Tuning

Para otimizar o modelo realizei o processo de ajustar os hiperparâmetros para melhorar seu desempenho.

Primeiro testei aumentando o número de *min number records per node* alcançando o valor ótimo demonstrado na figura 18.

Posteriormente, decidi implementar um loop com o objetivo de otimizar os parâmetros respeitando a função de preço, para tal segui o que foi aprendido nas aulas. Deste modo verifiquei que o melhor valor é 0,777.

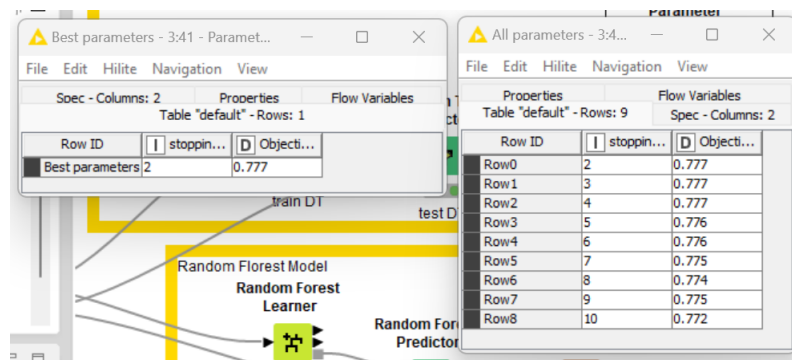


Figura 20: Resultados do tuning ao modelo baseado em Decision Tree.

2.4.2 Random Florest

Utilizei também o modelo de aprendizagem Random Florest para verificar se a utilização de múltiplas árvores de decisão traria alguma vantagem em termos de resultados, assim verifiquei que o valor de accuracy aumentou.

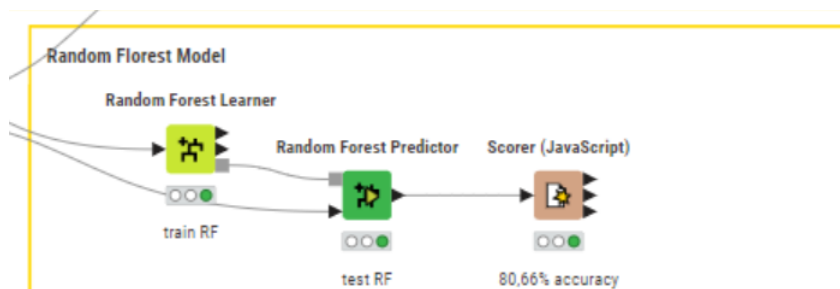


Figura 21: Modelo de Aprendizagem Random Florest.

Após vários testes, conclui que a melhor configuração do modelo é com base no *Information Gain* com uma diferença de 0,68% em relação ao *Gini Index*.

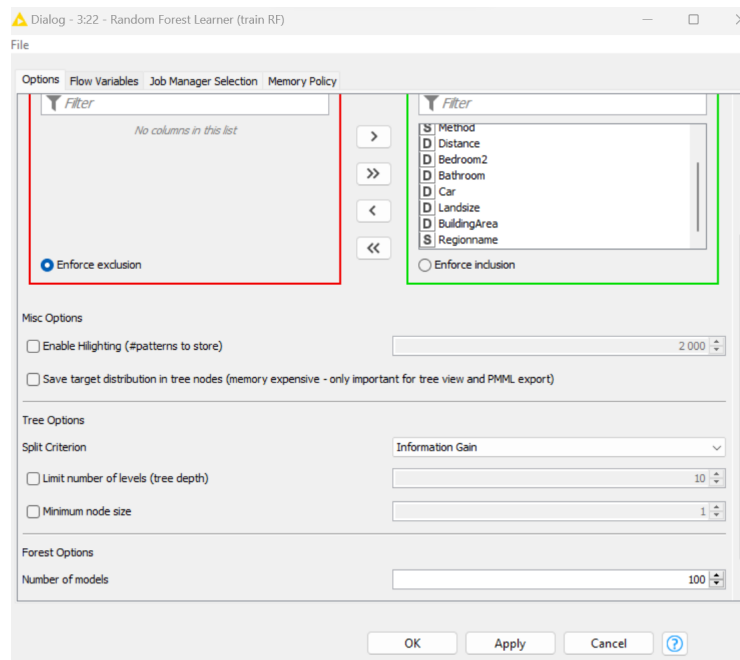


Figura 22: Configuração Random Florest.

Assim, obtive uma accuracy de 80,66%.

Scorer View

Confusion Matrix

	High (Predicted)	Low (Predicted)	Medium (Predicted)	
High (Actual)	755	5	146	83.33%
Low (Actual)	5	767	108	87.16%
Medium (Actual)	127	121	614	71.23%
	85.12%	85.89%	70.74%	

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa ($\delta_{\square\square'}$)	Correctly Classified	Incorrectly Classified
80.66%	19.34%	0.710	2136	512

Figura 23: Resultados do modelo Random Florest.

2.4.3 Linear Regression

Para conhecimento e avaliação adicional implementei também o modelo Linear Regression.

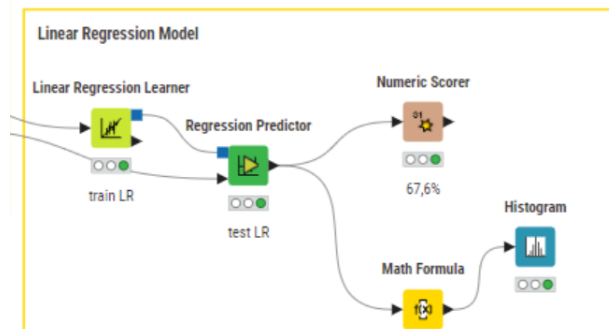


Figura 24: Modelo de aprendizagem Linear Regression.

Desta forma, obtivemos os seguintes resultados:

Row...	Prediction (Mean(Price)) Number (double)
R^2	0.676
mea...	0.102
mea...	0.017
root ...	0.131
mea...	0.004
mea...	0.307
adju...	0.676

Figura 25: Resultados do modelo Linear Regression.

Tendo em conta estes resultados verificamos que o valor de 0.676 significa que aproximadamente 67,6% da variabilidade nos dados de saída é explicada pelo modelo. O valor *mean signed difference* de 0.004 sugere que, em média, o modelo tem uma ligeira tendência de superestimar as previsões em relação aos valores reais. Por fim, o valor R^2 ajustado ,0.676, é o mesmo que o R^2 padrão, indicando que o ajuste do modelo permanece consistente após ajustar para o número de variáveis.

3 Conclusão

Em suma, com a realização deste projeto foi possível conceber e otimizar modelos de machine learning, nomeadamente os modelos de Decision Tree, Random Florest e Linear Regression.

Além disso,verificou-se um desempenho ligeiramente melhor por parte do modelo baseado em Random Forest.