Inteligência Artificial

Docente: Dalila Durães

Conceção e otimização de modelos de Machine Learning

Camila Pinto PG53712

Ano Letivo 2023/2024

Conteúdo

01 Introdução

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.
- Para atingir o objetivo de previsão de preços, adotarei a metodologia CRISP-DM (CrossIndustry 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.

02 Dataset

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 edíficio.
• YearBuilt - ano de construção,
• CouncilArea - concelho,
• latitude,
• longitude,
• Regionname - Nome da região,
• Propertycount - quantidade de propriedades existentes no subúrbio

03 Exploração dos dados

Atributos region name e council area

A partir da análise do dataset verifica-se que os atributos region name e council área estavam relacionados. Analisando a figura nota-se que os atributos correspondem a 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	We
?	242	9	336	15	125	355	280	
	17,6771%	0,6574%	24,5435%	1,0957%	9,1308%	25,9313%	20,4529%	
Banyule	478		116					
	80,4714%		19,5286%					
Bayside						489		
						100%		
Boroondara	1					1 159		
	0,0862%					99,9138%		
Brimbank							424	
							100%	
Cardinia		8						
		100%						
Casey		11			27			
		28,9474%			71,0526%			
Darebin			934					
			100%					
Frankston		8			45			

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.

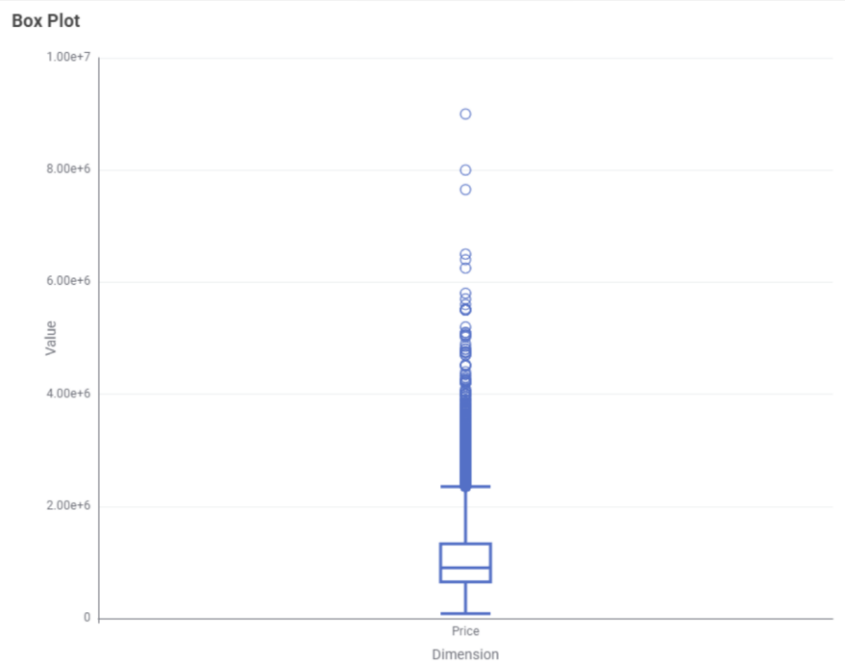
Statistics

Rows: 21 | Columns: 9

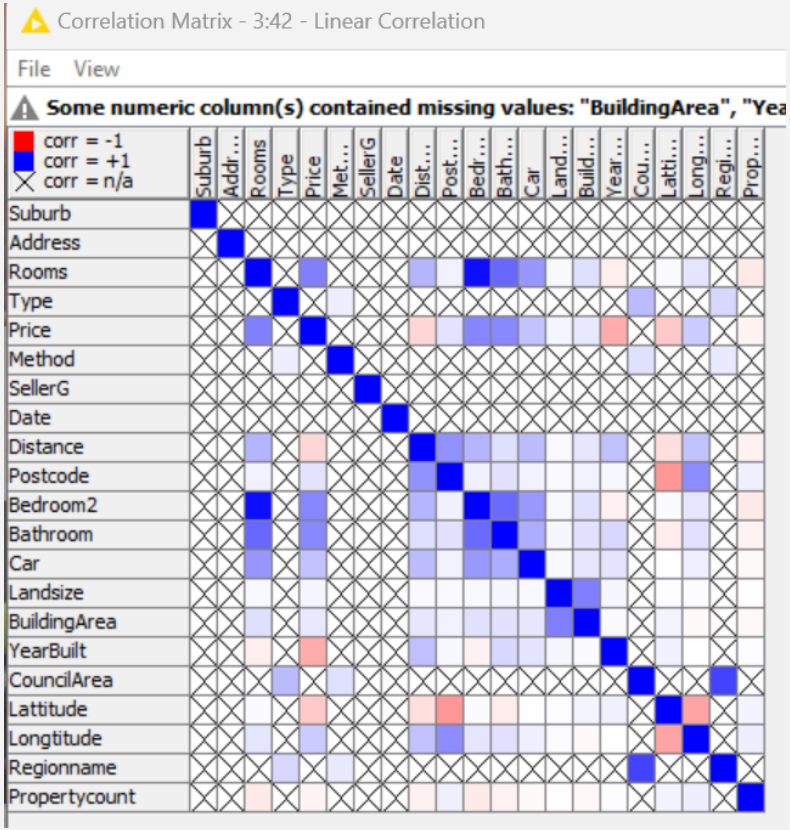
Name	Type	# Missing values	Minimum	Maximum	25% Quantile	50% Quantile (M...	75% Quantile	Mean
Date	String	0						
Distance	Number (double)	0	0	48.1	6.1	9.2	13	10.138
Postcode	Number (double)	0	3,000	3,977	3,044	3,084	3,148	3,105.302
Bedroom2	Number (double)	0	0	20	2	3	3	2.915
Bathroom	Number (double)	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						
Lattitude	Number (double)	0	-38.183	-37.409	-37.857	-37.802	-37.756	-37.809
Longtitude	Number (double)	0	144.432	145.526	144.93	145	145.058	144.995

Atributos price, landsize e builing area

- Com o auxilio do módulo Data Explorer, seguido do módulo para visualização Box Plot, foi possível identificar a presença de uma grande quantidade de outliers nestes atributos.

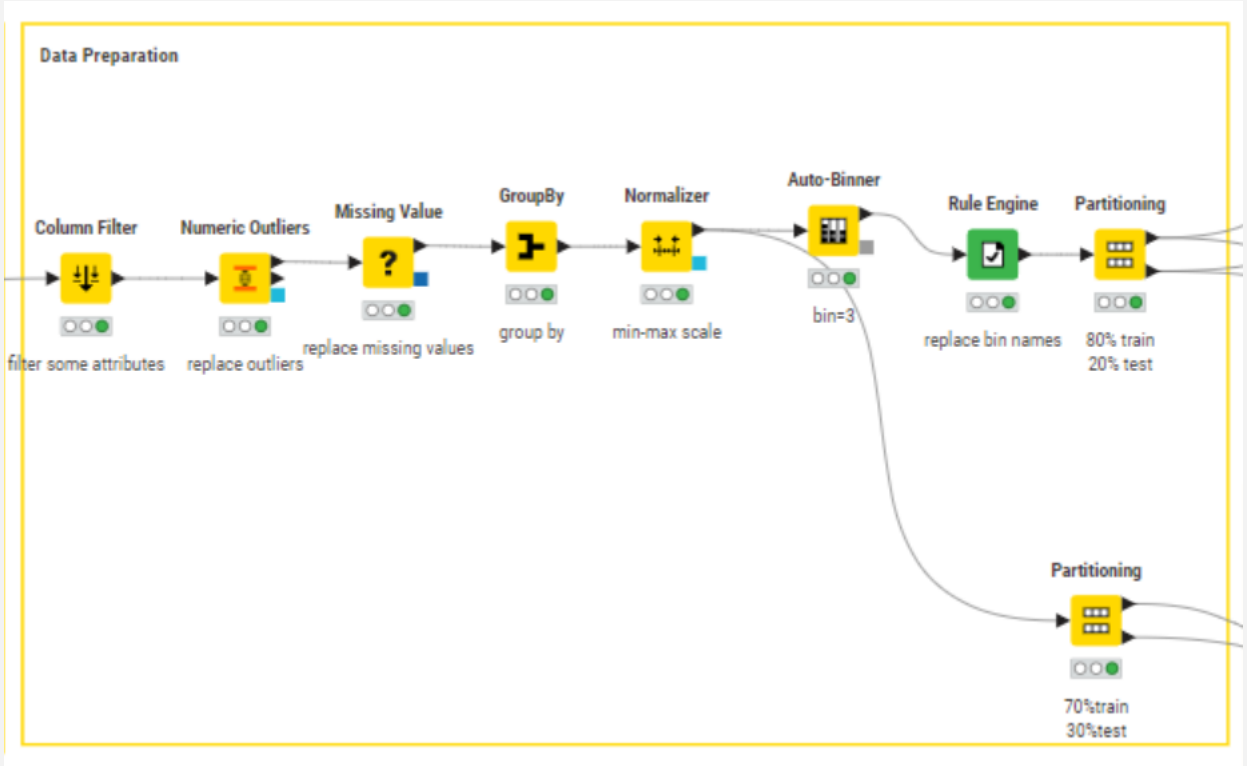


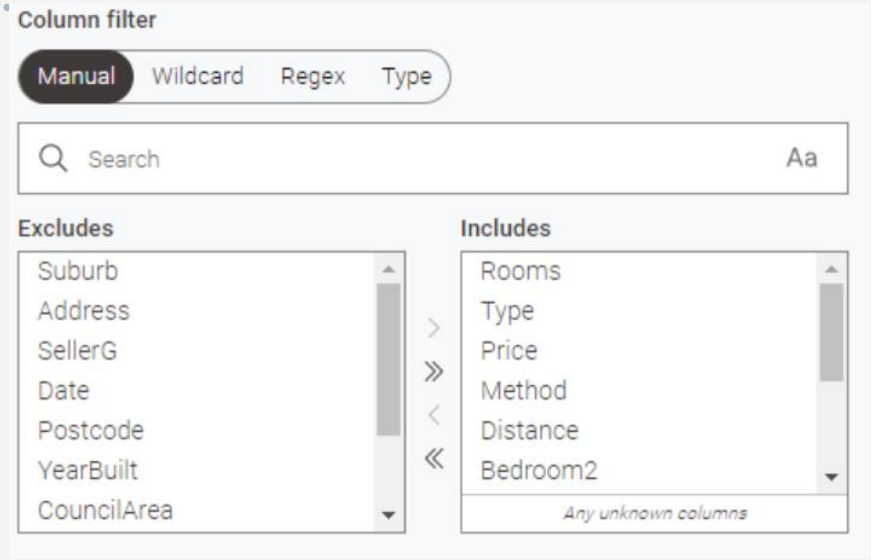
Atributos price,rooms, distance



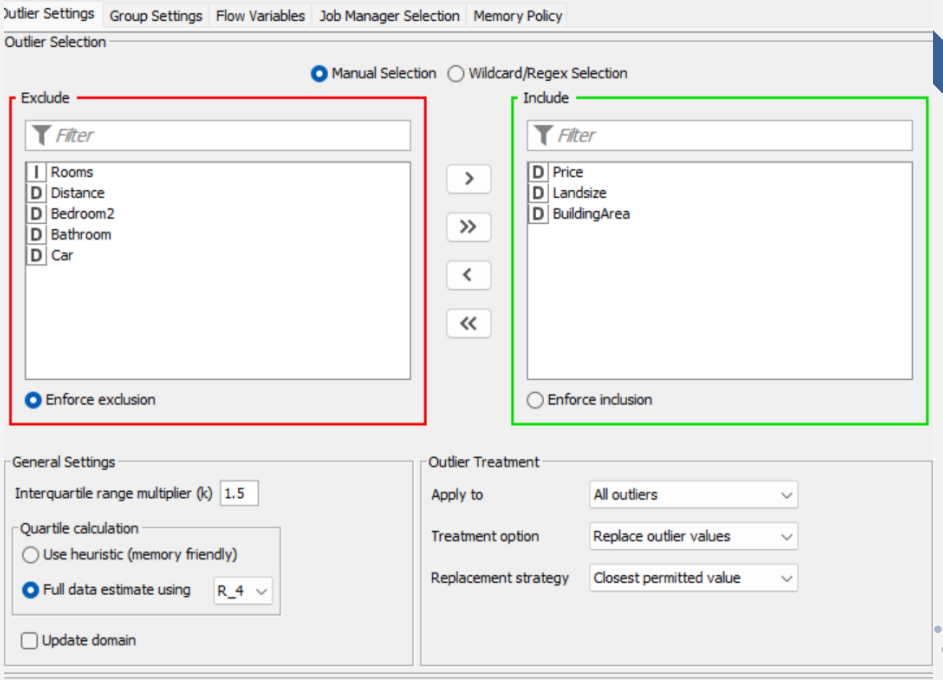
04

Preparação de dados e modelação

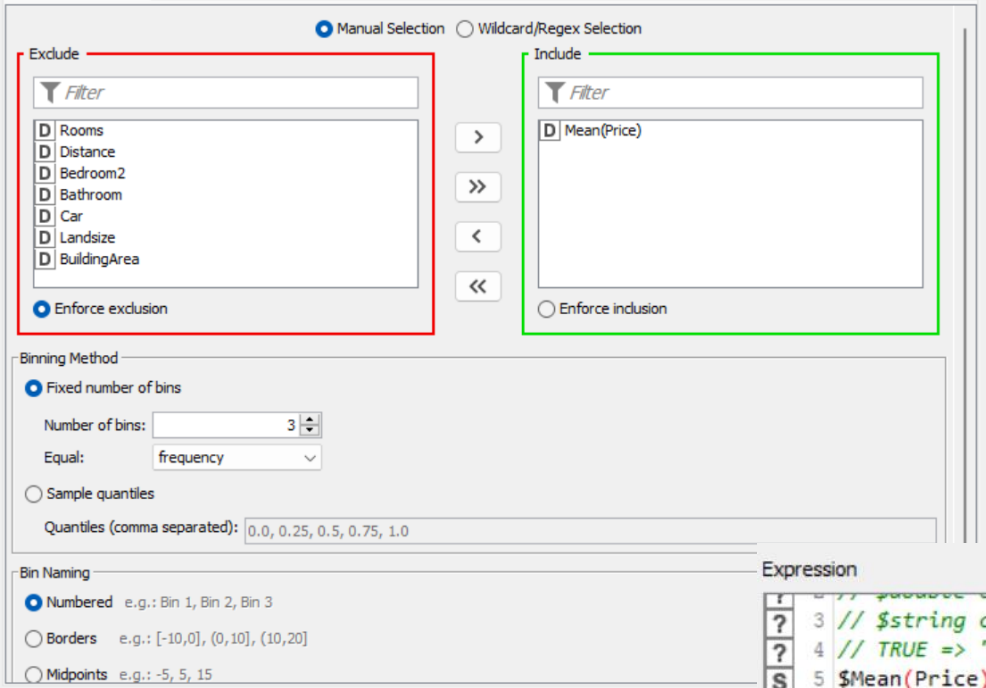

Data Preparation
Column Filter
filter some attributes
Numeric Outliers
replace outliers
Missing Value
replace missing values
GroupBy
group by
Normalizer
min-max scale
Auto-Binner
bin=3
Rule Engine
replace bin names
Partitioning
80% train
20% test
Partitioning
70%train
30%test



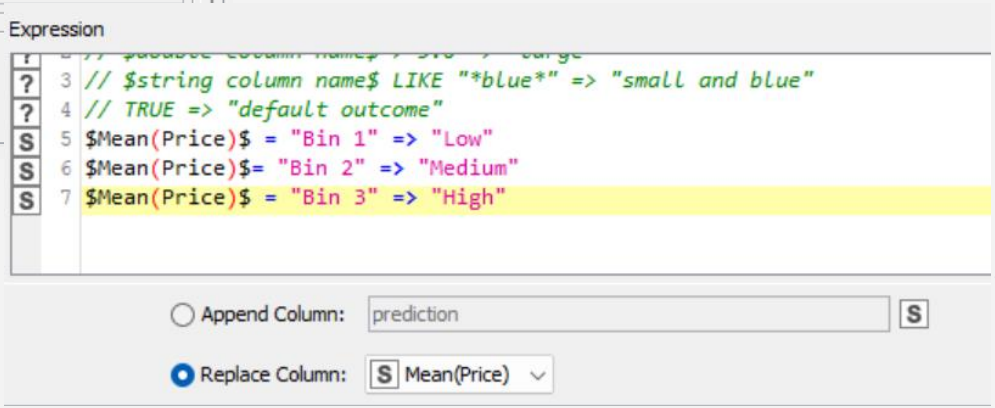
Column Filter



Numeric Outliers

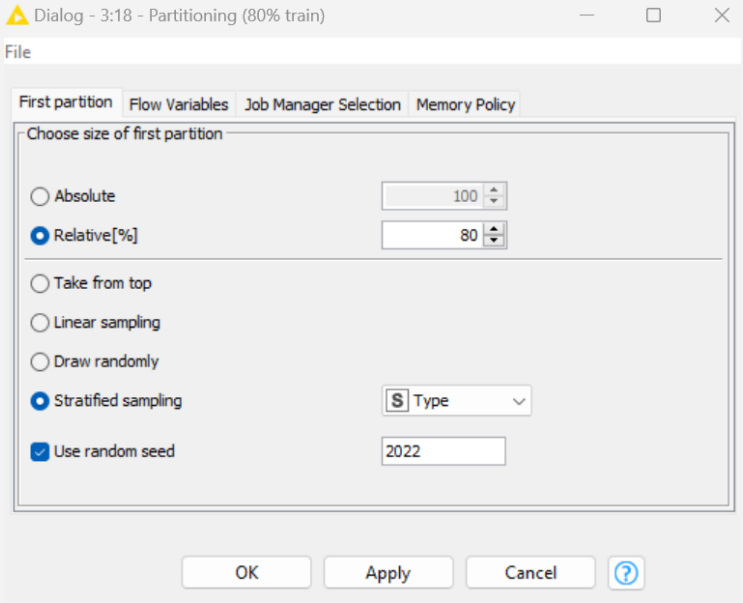


Group By



Rule Engine

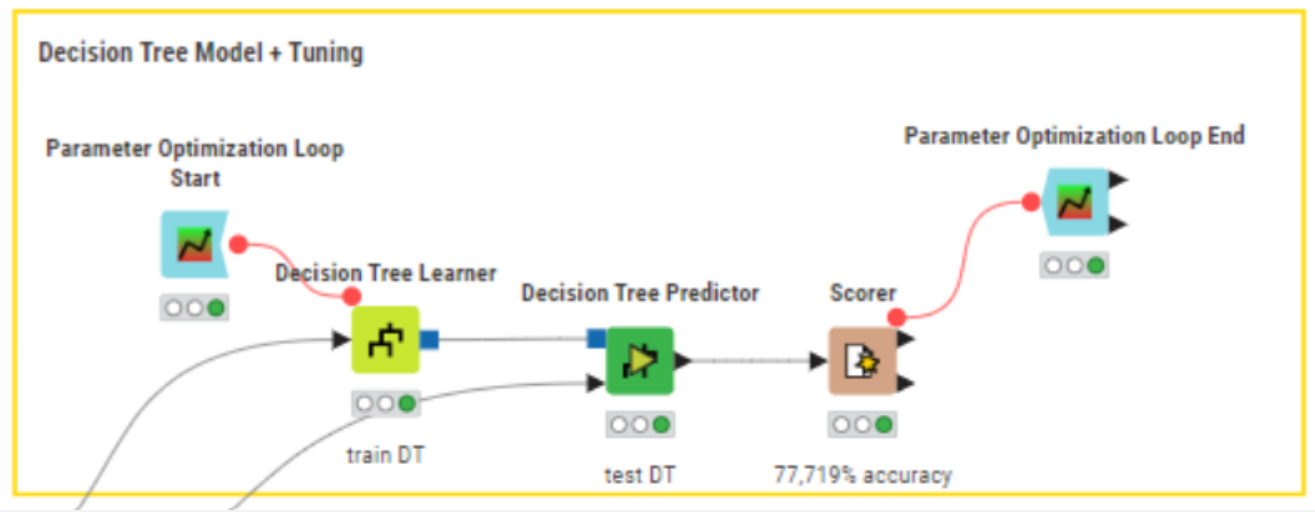
De modo a homogeneizar a amostra foi usado o modo stratifed sampling, assim foram testados vários atributos de modo a selecionar o que iria fornecer melhor desempenho ao modelo.Optouse pelo type, cuja diferença do segundo melhor classifado é 1,94%. Para o modelo Regression Linear, optei por usar o modo Draw Randomly.

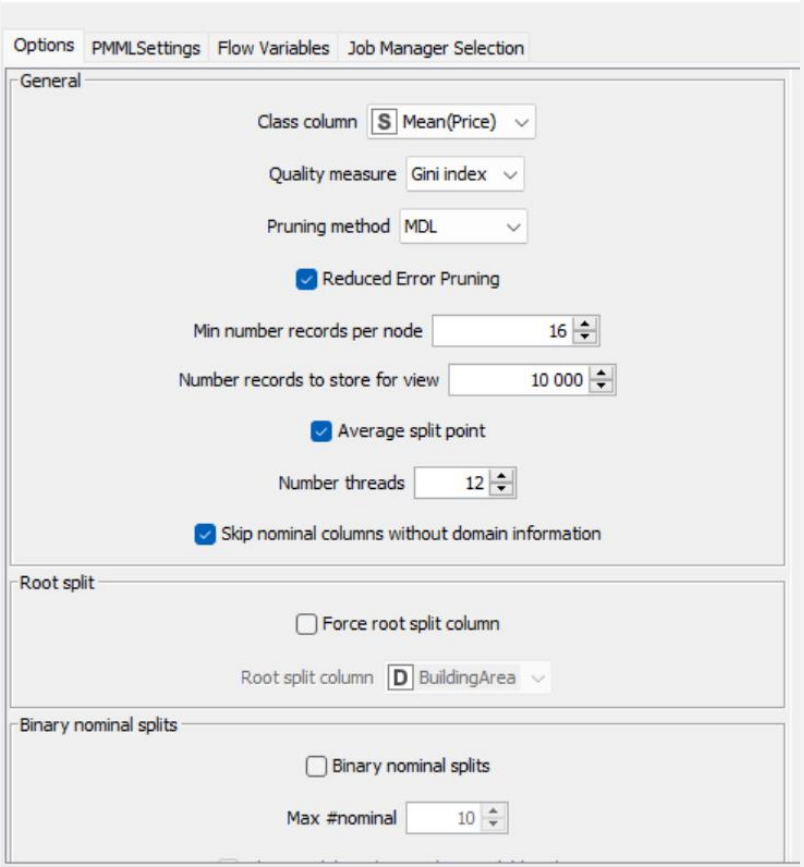


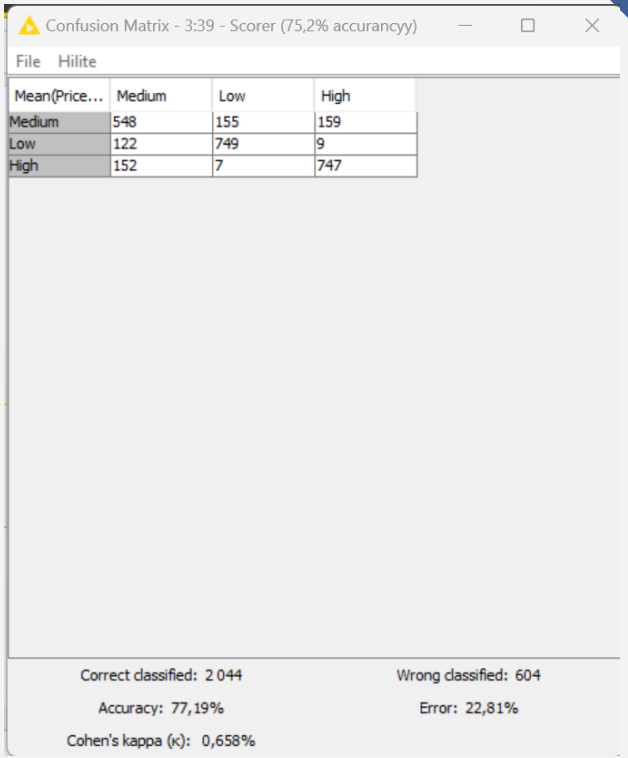
05

Modelo e Resultados

Decision Tree

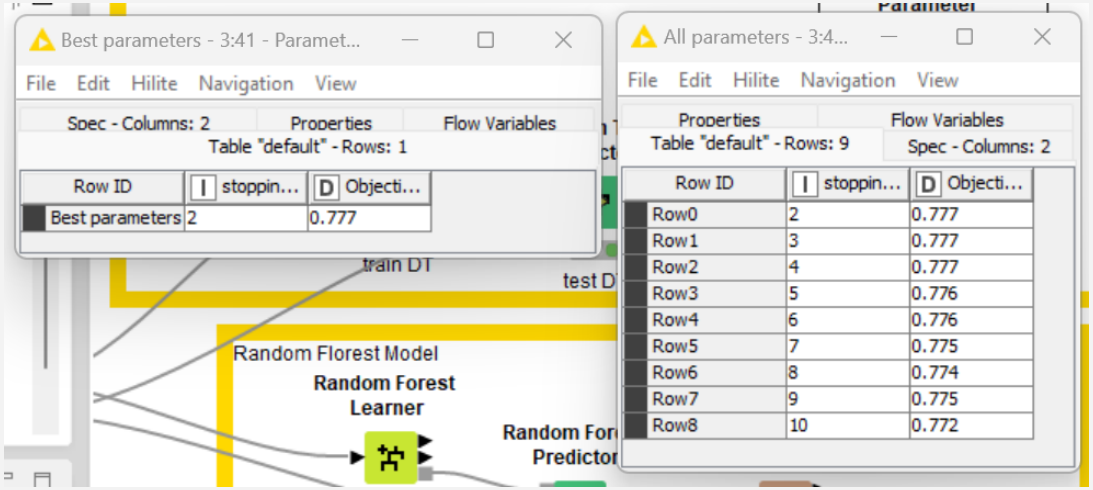






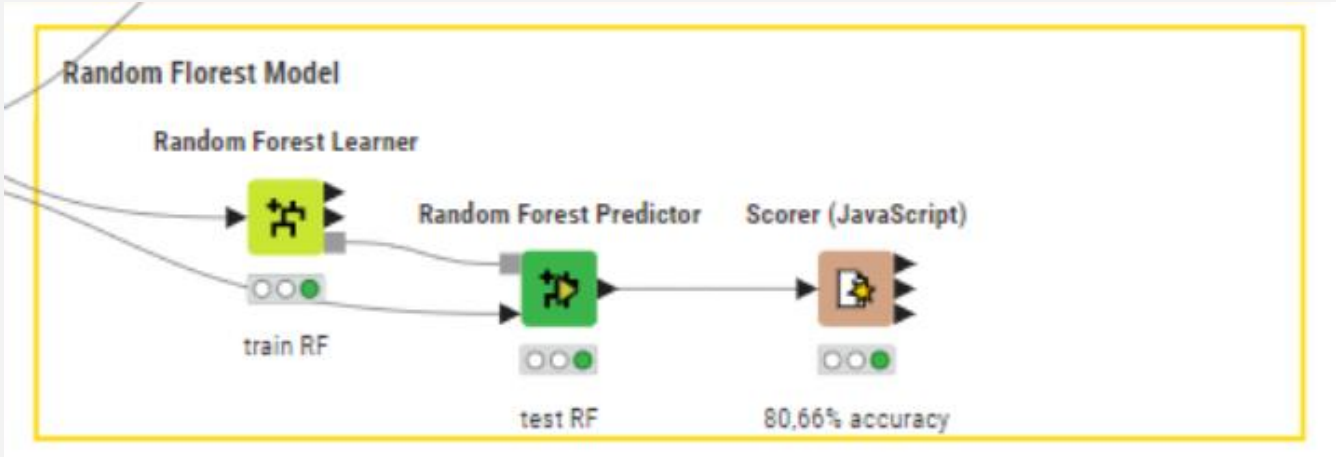
Após vários testes, concluiu-se que a melhor configuração do node coincide com a utilização do Gini index e do MDL. Assim, obteu-se uma accuracy de 77,19%.

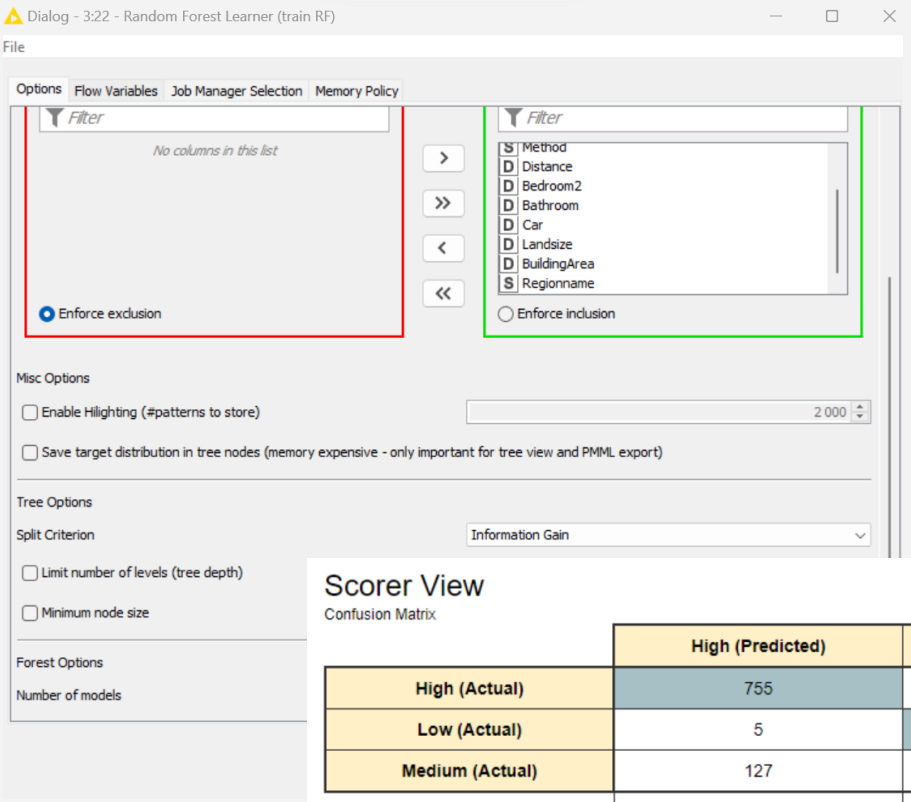
Tuning



Primeiro testei aumentando o número de min number records per node alncançando o valor ótimo. Posteriormente, decidi implementar um loop com o objetivo de otimizar os parâmentros respeitando a função de preço, para tal segui o que foi aprendido nas aulas. Deste modo verifiquei que o melhor valor é 0.777.

Random Florest





A melhor configuração do modelo é com base no Information Gain com uma diferença de 0,68% em relação ao Gini Index.

Scorer View

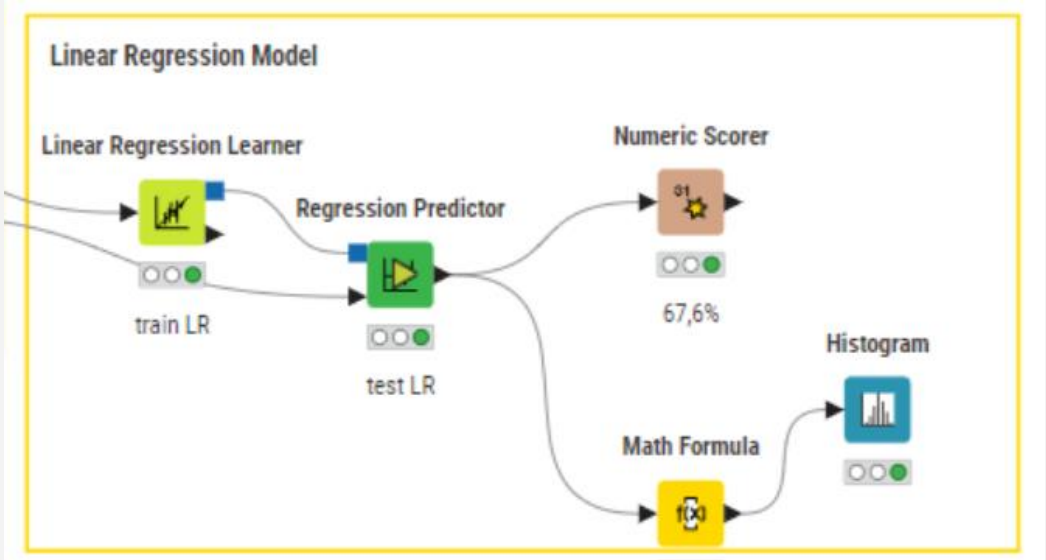
Confusion Matrix

	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 (ð□□¹)	Correctly Classified	Incorrectly Classified
80.66%	19.34%	0.710	2136	512

Linear Regression



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

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.

06 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.

Obrigado!