



Escola de Engenharia
Universidade do Minho

Mestrado Integrado em Engenharia de
Gestão e Sistemas de Informação
2023/2024

4º Ano

1º Semestre

Aprendizagem Automática em Sistemas de Informação



Francisco Miguel Pinheiro Cardoso

A79570

Índice

1.	Modeling	1
1.1	Select Modeling Technique.....	1
1.1.1	Modeling Techniques.....	1
1.1.2	Modeling Assumptions.....	2
1.2	Generate Test Design.....	2
1.2.1	Test Design.....	2
1.3	Build Model	3
1.3.1	Parameter settings	3
1.3.2	Models.....	7
1.4	Assess Model.....	11
1.4.1	Model Assessment	11
2.	Evaluation	15
2.1	Evaluation Results.....	15
2.1.1	Assessment of Data Mining Results	15
2.1.2	Approved Models	16
2.2	Review Process	17
2.2.1	Review Process	17
2.3	Determine Next Steps	17
2.3.1	List of possible actions.....	17
2.3.2	Decision	17

1. Modeling

1.1 Select Modeling Technique

1.1.1 Modeling Techniques

Nesta fase selecionei diferentes técnicas de modelação de regressão a aplicar aos quatro cenários criados na fase anterior. As técnicas escolhidas estão na seguinte tabela.

Técnicas	Descrição
Decision Tree	Uma árvore de decisão é um mapa dos possíveis resultados de uma série de escolhas relacionadas.
Random Forest	Cria várias árvores de decisão durante o treinamento e as combina para obter uma decisão mais robusta e geralmente mais precisa do que as decisões individuais das árvores.
Support Vector Machine	Uma SVM constrói Hiper planos num espaço n-dimensional para classificar ou regredir dados.
Linear Regression	O objetivo principal da regressão linear é a análise de duas variáveis e seus respectivos resultados. Essa análise parte sempre de uma variável dependente com outras chamadas de independentes. O objetivo geral é encontrar relações entre essas variáveis de análise.
Deep Learning	O mecanismo Deep Learning é caracterizado pela inserção de dados em um computador para avaliar as respostas, a fim de confirmar previsões precisas e corrige as erradas, sendo modelos quase impossíveis de explicar e compreender, mesmo por especialistas que podem ver as suas estruturas

1.1.2 Modeling Assumptions

Muitas técnicas de modelação fazem suposições específicas sobre um conjunto de dados. Por exemplo, que todos os atributos têm distribuições uniformes, que o atributo de classe deve ser simbólico, etc.

Suposições:

- Decision Tree: Se os valores forem contínuos, eles vão sofrer uma discretização antes da construção do modelo.
- Random Forest: Se os valores forem contínuos, eles vão sofrer uma discretização antes da construção do modelo.
- Support Vector Machine: Assume que os dados são independentes e distribuídos de forma idêntica
- Linear Regression: Pressuposto de que existe uma relação linear entre as variáveis independentes e dependentes e que não deveria existir multicolinearidade (problema comum em regressões, onde as variáveis independentes possuem relações lineares exatas ou aproximadamente exatas).
- Deep Learning: Pressuposto de que há mínima ou nenhuma multicolinearidade entre as variáveis independentes.

1.2 Generate Test Design

1.2.1 Test Design

Nesta etapa vou descrever o plano de conceção dos testes. Inclui a definição dos diferentes cenários de testes a escolha das técnicas e métodos de avaliação a aplicar a cada um deles, como se pode verificar na seguinte tabelas.

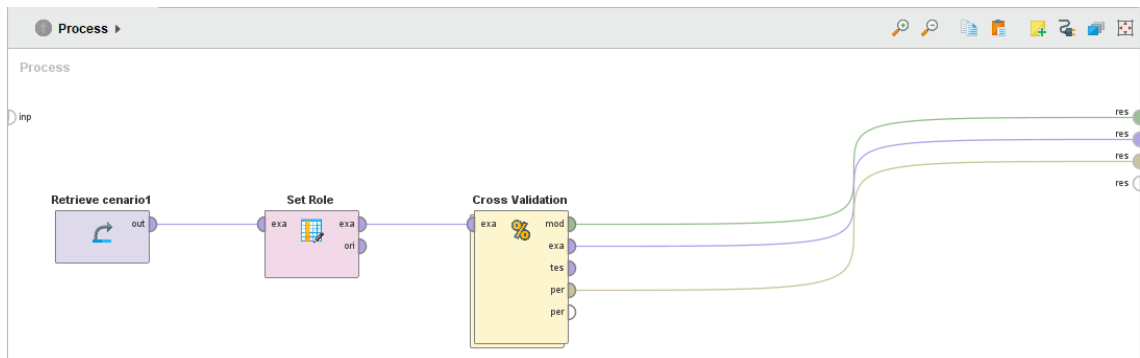
Cenário	Técnicas a utilizar	Método de teste e validação
Cenário 1	Todas	Cross Validation
Cenário 2	Todas	Cross Validation
Cenário 3	Todas	Cross Validation
Cenário 4	Todas	Cross Validation

1.3 Build Model

1.3.1 Parameter settings

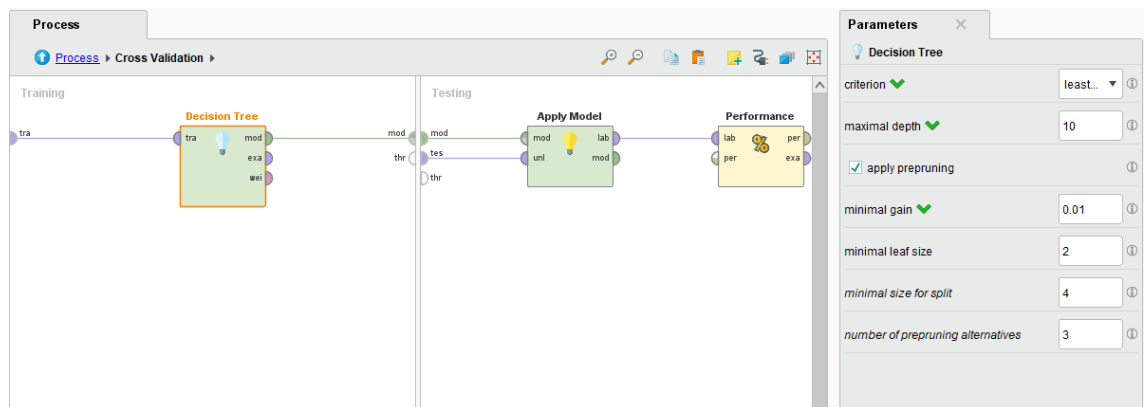
Depois de concluída a definição de cenários, bem como as técnicas a utilizar para cada um, de forma a obter os resultados mais eficazes e eficientes, foram atualizadas as configurações da ferramenta utilizada, o Rapid Miner.

Nas figuras abaixo estão representados os processos utilizados na construção dos diferentes modelos definidos. Na seguinte imagem está representado o processo utilizado nos diferentes modelos definidos.

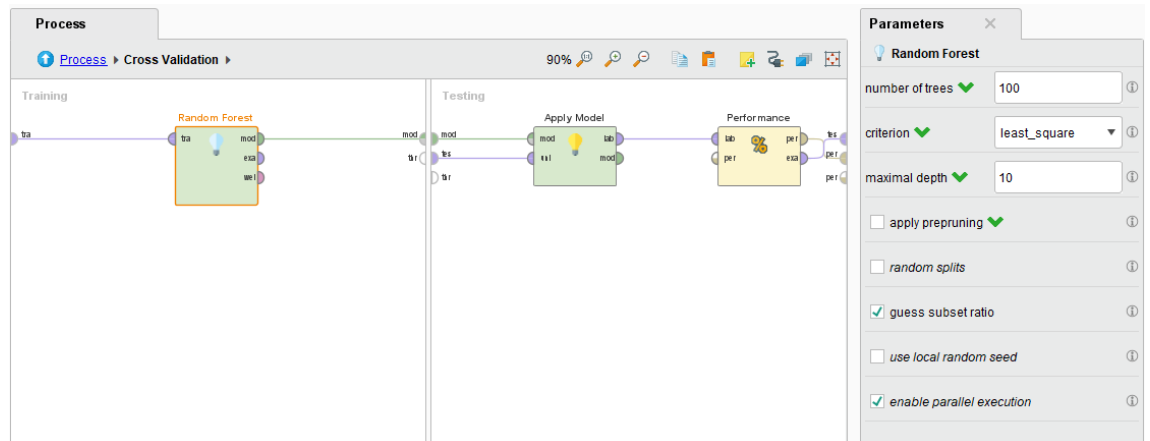


Em seguida, são apresentados os parâmetros escolhidos para as definições de cada técnica, bem como o subprocesso utilizado para cada um no operador cross validation.

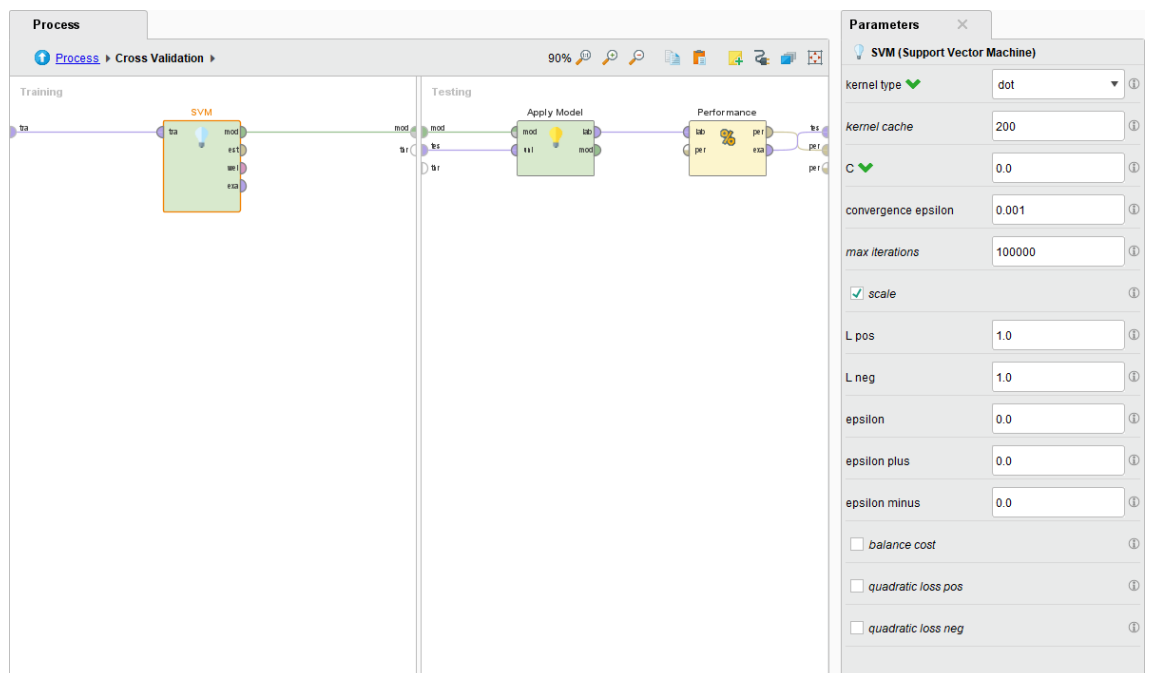
Decision Tree



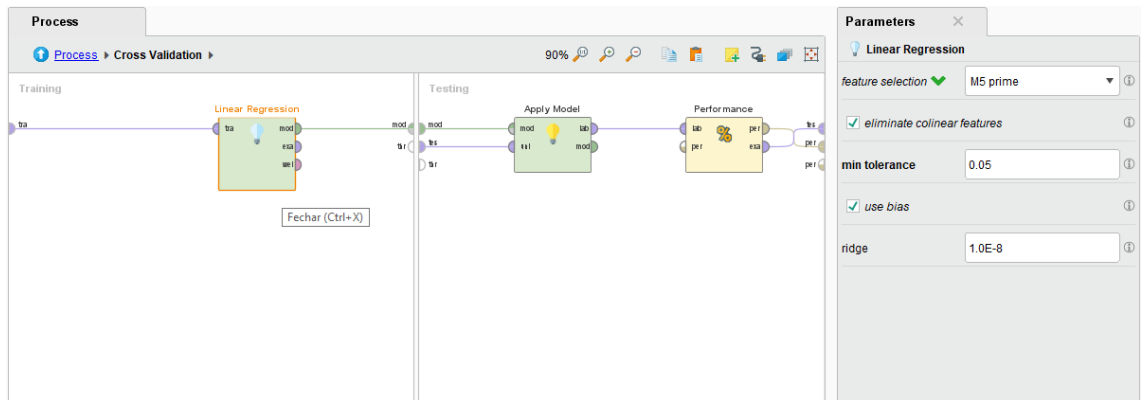
Random Forest



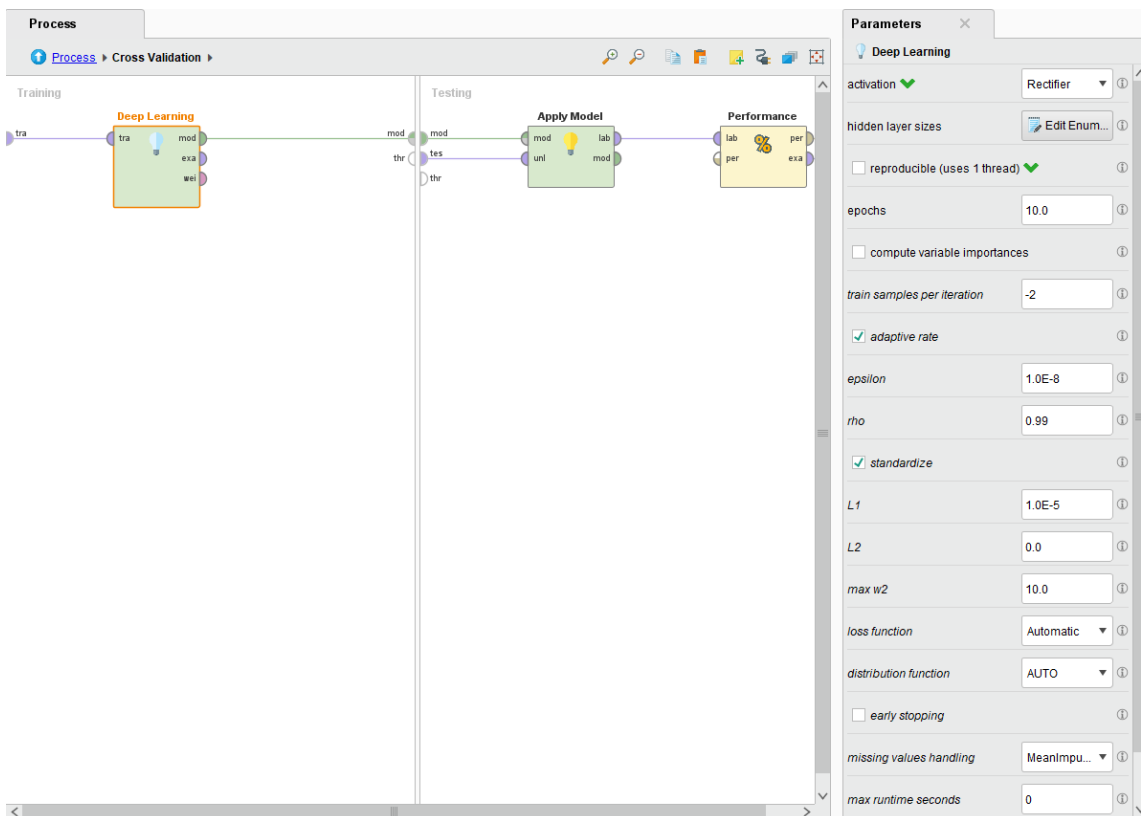
SVM



Linear Regression



Deep Learning



Performance

Parameters

% Performance (Performance (Regression))

main criterion

first

☒ root mean squared error

☒ absolute error

☒ relative error

☐ relative error lenient

☐ relative error strict

☐ normalized absolute error

☒ root relative squared error

☐ squared error

☒ correlation

☐ squared correlation

☐ prediction average


☐ spearman rho

☐ kendall tau

☒ skip undefined labels

comparator class

☒ use example weights

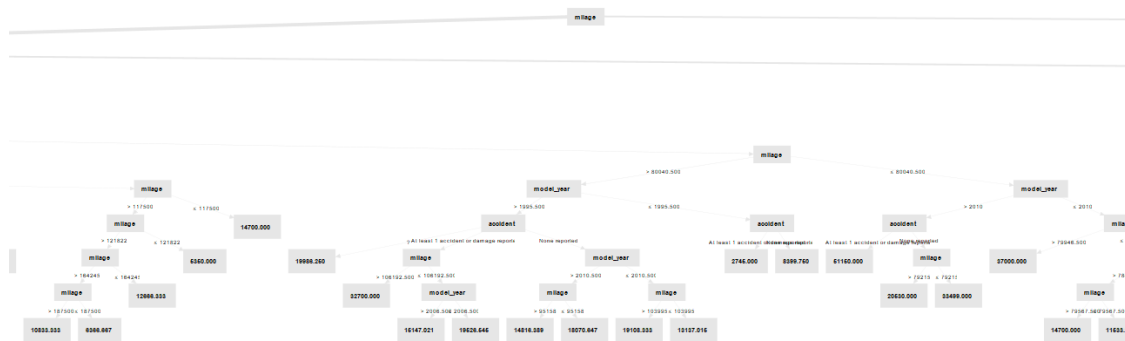
 [Hide advanced parameters](#)

1.3.2 Models

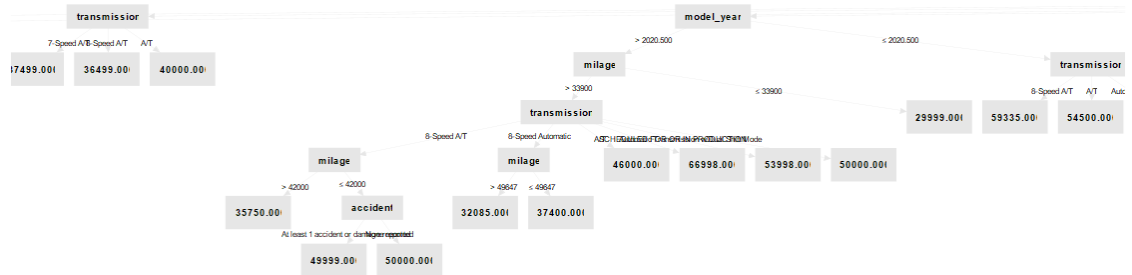
De seguida, apresentam-se os diversos modelos obtidos para as diferentes técnicas respetivos a cada cenário.

Cenário 1

Amostra do modelo de Árvore de Decisão

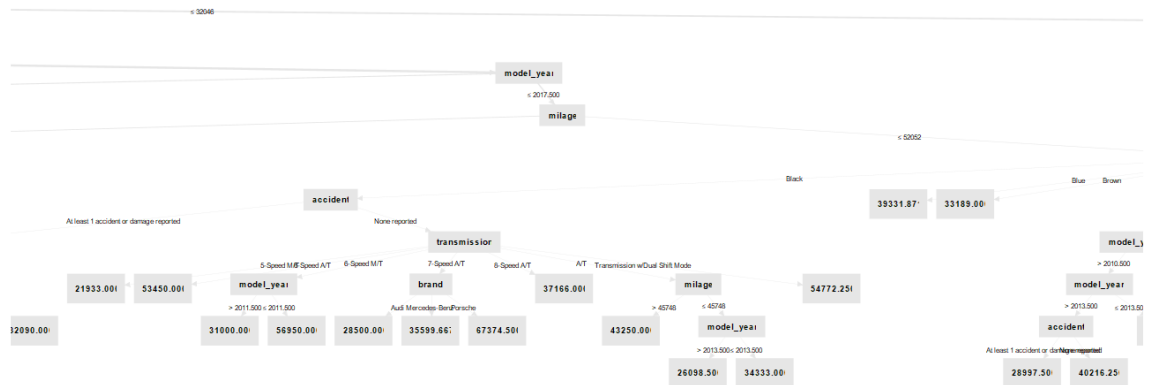


Amostra do modelo de Random Forest

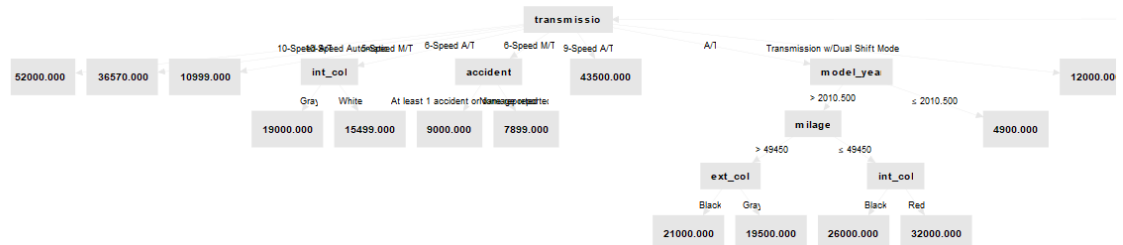


Cenário 2

Amostra do modelo de Árvore de Decisão

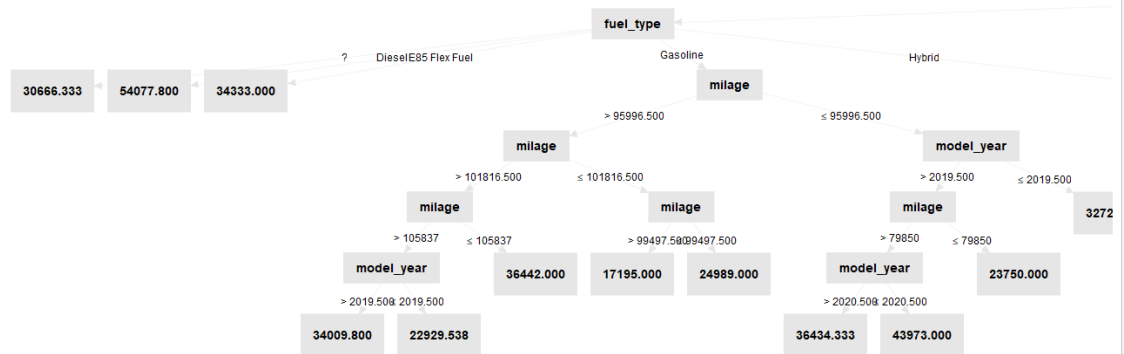


Amostra do modelo de Random Forest

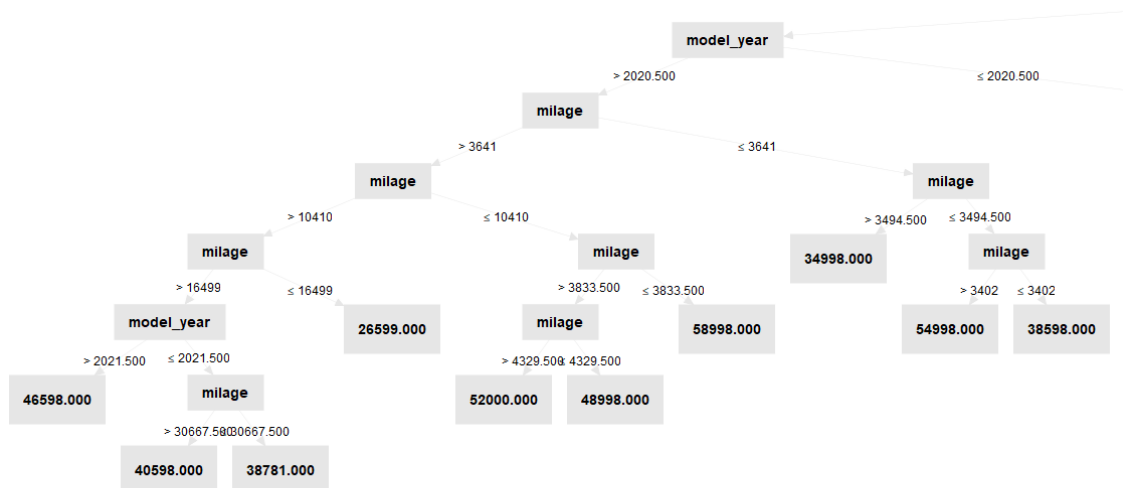


Cenário 3

Amostra do modelo de Árvore de Decisão

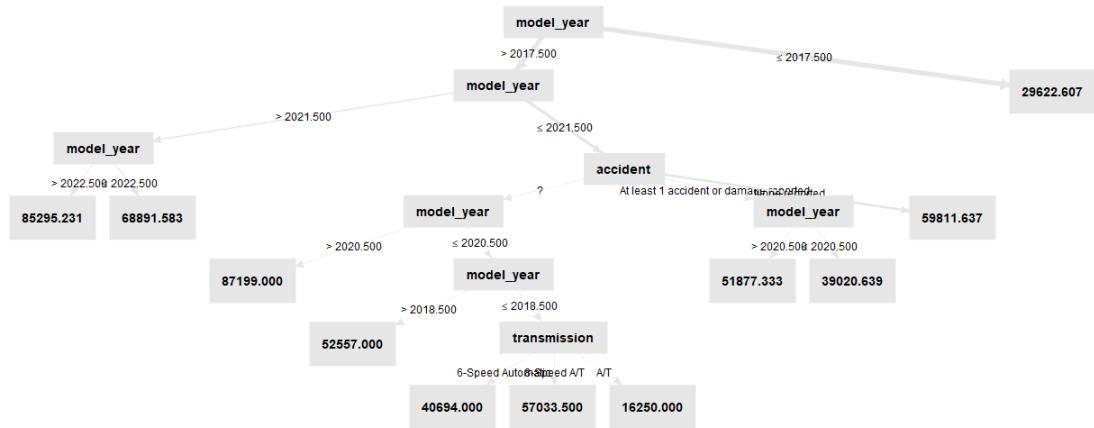


Amostra do modelo de Random Forest



Cenário 4

Modelo de Árvore de Decisão



Amostra do modelo de Random Forest



1.4 Assess Model

1.4.1 Model Assessment

Cenário 1

	Técnicas				
	Decision Tree	Random Forest	SVM	Linear Regression (500)	Deep Learning
Root mean squared error	75711	61488	75105	53275	56995
Absolute error	23866	16370	28291	37414	18583
Relative error	59.35%	48.51%	108.38%	177.13%	59.34%
Root relative squared error	1.041	0.762	1.014	1.162	0.696
Correlation	0.417	0.622	0.443	0.431	0.666

Analisando os resultados da tabela anterior referente ao cenário 1, consigo concluir que a técnica Random Forest é a que apresenta melhores resultados. No entanto, é possível verificar que os valores não são muito favoráveis para a previsão de valores pois há uma margem muito grande erro. É de notar também que para a técnica Linear Regression foram utilizados apenas 500 casos como amostra devido à capacidade do computador em analisar os dados.

Cenário 2

	Técnicas				
	Decision Tree	Random Forest	SVM	Linear Regression	Deep Learning
Root mean squared error	74983	56195	67873	50719	51292
Absolute error	23747	17209	26431	30750	19440
Relative error	65.11%	57.14%	94.15%	107.79%	70.10%
Root relative squared error	1.201	0.786	1.022	1.323	0.704
Correlation	0.354	0.674	0.504	0.527	0.748

Analisando os resultados da tabela anterior referente ao cenário 2, consigo concluir que a técnica Random Forest é a que apresenta melhores resultados. No entanto, é possível verificar que os valores não são muito favoráveis para a previsão de valores pois há uma margem muito grande erro. É de notar também que para a técnica Linear Regression foram utilizados apenas 100 casos como amostra devido à capacidade do computador em analisar os dados.

Cenário 3

	Técnicas				
	Decision Tree	Random Forest	SVM	Linear Regression (100)	Deep Learning
Root mean squared error	75482	60637	75105	81034	55899
Absolute error	23646	16147	28291	55507	16730
Relative error	58.30%	43.40%	108.38%	164.28%	49.67%
Root relative squared error	1.016	0.747	1.014	3.121	0.664
Correlation	0.403	0.653	0.445	0.416	0.689

Analisando os resultados da tabela anterior referente ao cenário 3, consigo concluir que a técnica Random Forest é a que apresenta melhores resultados. No entanto, é possível verificar que os valores não são muito favoráveis para a previsão de valores pois há uma margem muito grande erro. É de notar também que para a técnica Linear Regression foram utilizados apenas 100 casos como amostra devido à capacidade do computador em analisar os dados.

Cenário 4

	Técnicas				
	Decision Tree	Random Forest	SVM	Linear Regression (100)	Deep Learning
Root mean squared error	71570	66891	75106	51192	57794
Absolute error	25765	18433	28291	32220	19086
Relative error	92.60%	56.76%	108.38%	114.52%	65.78%
Root relative squared error	0.946	0.863	1.014	1.455	0.702
Correlation	0.299	0.564	0.427	0.412	0.663

Analisando os resultados da tabela anterior referente ao cenário 4, consigo concluir que a técnica Random Forest é a que apresenta melhores resultados. No entanto, é possível verificar que os valores não são muito favoráveis para a previsão de valores pois há uma margem muito grande erro. É de notar também que para a técnica Linear Regression foram utilizados apenas 100 casos como amostra devido à capacidade do computador em analisar os dados.

2. Evaluation

2.1 Evaluation Results

2.1.1 Assessment of Data Mining Results

Depois de criados todos os modelos, é possível verificar que o erro relativo surge na ordem dos 43.40% e 57.14% o que são valores não satisfatórios, pois indica que há um erro relativo de metade das previsões efetuadas serem falhadas.

É possível verificar melhor esta condição, na prática, através do erro absoluto médio (MAE) onde na melhor previsão existe um erro de previsão de 16147\$, em relação aos preços reais dos carros.

A métrica root relative squared error, combina elementos do erro relativo (RMSE) e erro quadrático. Os valores demonstram que o erro relativo do RMSE é menor do que 1. O que indica que as previsões desempenhadas pelo modelo são relativamente boas, no que diz respeito aos valores reais.

Por fim, na correlação podemos verificar que um valor de 0.653, por exemplo, sugere que há uma tendência para as previsões e os valores reais dos preços dos carros aumentem e diminuam juntos, ou seja, quando o modelo prevê um preço maior, os valores reais tendem a ser mais altos e o mesmo acontece de forma inversa.

2.1.2 Approved Models

Na tabela seguinte, encontram-se apresentados para cada cenário, o seu melhor modelo.

Cenário	Modelo	Root mean squared error	Absolute error	Relative error	Root relative squared error	Correlation
Cenário 1	Random Forest	61488	16370	48.51%	0.762	0.622
Cenário 2	Random Forest	56195	17209	57.14%	0.786	0.674
Cenário 3	Random Forest	60637	16147	43.40%	0.747	0.653
Cenário 4	Random Forest	66891	18433	56.76%	0.863	0.564

Analisando a tabela anterior é possível perceber que, de forma geral, todos os modelos apresentam valores semelhantes, uns melhores que outros, no entanto, consoante os valores dos dados fazem com que não sejam muito divergentes.

Relativamente ao absolute error, o cenário que apresenta melhores resultados é o Cenário 3, bem como relativamente ao root relative squared error. Já em relação à correlation os valores são bastantes semelhantes e positivos. Ou seja, o melhor cenário seria o 3.

Assim sendo, posso afirmar que os critérios de sucesso, não foram atingidos pois as previsões possuem valores demasiados altos do que o esperado.

2.2 Review Process

2.2.1 Review Process

A principal limitação do projeto sentida, foi a qualidade dos dados. Os mesmos não apresentavam uma estrutura ideal para avaliação, tendo de ser realizadas várias transformações e talvez seria necessário a realização de mais para uma melhor obtenção de previsões. Os modelos foram insatisfatórios na generalidade dos cenários avaliados, não conseguindo cumprir critérios de sucesso pretendidos.

2.3 Determine Next Steps

2.3.1 List of possible actions

Devido ao contexto académico em que este projeto se enquadra, o projeto será dado como terminado. Num caso real, os próximos passos estariam relacionados ou com o deployment, ou com a iniciação de uma nova iteração. Neste caso, o grupo optaria pela segunda opção, caso houvesse recursos, orçamento e tempo para tal.

2.3.2 Decision

Para ir ao encontro do objetivo de negócio e reduzir o erro absoluto, considera-se necessário continuar a recolher dados das transações realizadas. Assim, com uma maior quantidade de dados, seria possível reduzir a taxa de erro através da geração de um modelo mais assertivo e verificar os carros mais antigos e perceber se os valores são mais assertivos para tentar promover a venda dos mesmos.