Chatbot - Árvore de decisão Programação Estruturada

GEOVANNE DA SILVA GUIMARÃES
MATHEUS MOREIRA FRANCISCO
PEDRO HENRIQUE DE MORAES LUI

Análise dos dados – Base geral

- Organização dos dados a partir do ZIPCODE: 98056;
- Formatação da planilha através do Google Sheets:

Quartos	Banheiros	Área construída (m^2)	Área do lote (m^2)	Andares	Estado de conservação	Nível de construção
4	2.5	2640	25038	2	4	8
3	1	1600	7324	1	4	7
4	2.5	3270	24750	1	4	8
3	1.75	1400	8364	1	4	7
4	2	1440	9477	1	3	5
4	2.5	2510	5258	2	3	7
5	3.5	3060	8862	2	3	8
2	1	1240	42247	1	4	7
2	1	880	6900	1	3	6

Análise dos dados - Colunas

- Exclusão das colunas:

-id;
-date;
-price - (Excluída posteriormente);
-waterfront;
-view;
-sqft_above;
-sqft_basement – (Modificado o critério);
-yr_built - (Excluída posteriormente);
-yr_renovated;
-zipcode:
-lat e long;
-sqft_living15 e sqft_lot15.

Análise dos dados - Colunas

- Adicionamos:

- "Preço por área construída" - Relacionando as colunas ***price*** e ***sqft_living.*** Após isto, realizamos a média para cada casa referente a esta coluna.

- "Caro?" - Verifica se "Preço por área construída" é maior ou menor que a média.

Preço por Área Construída	Caro?
248,8636364	VERDADEIRO
212,5	VERDADEIRO
171,2538226	FALSO

Análise dos dados - Binarização

- Binarização dos dados para criação do Chatbot;
- Média para cada critério escolhido;
- Binarização em **True** and **False**, conforme a média.

Quartos	Quartos > 3,43?	Banheiros	Banheiros > 1,44?	Área construída (m^2)	AC > 2017?
4	VERDADEIRO	2.5	VERDADEIRO	2620	VERDADEIRO
5	VERDADEIRO	2.5	VERDADEIRO	3150	VERDADEIRO
3	FALSO	3.25	VERDADEIRO	2770	VERDADEIRO
2	FALSO	2.25	VERDADEIRO	1610	FALSO
5	VERDADEIRO	3.5	VERDADEIRO	3960	VERDADEIRO

Análise dos dados – Impureza de Gini

- Realizamos a seguinte substituição nos critérios:

  - VERDADEIRO = 1
  - FALSO = 0

Quartos > Média?	Banheiros > Média?	AC > Média?
1	1	1
1	1	1
0	1	1
0	1	0
1	1	1

Análise dos dados – Impureza de Gini

- Realização da Impureza de Gini para as colunas;
- Início do esboço da árvore.

True	True	True	True	True	True	True	True	True	True
Yes	***Yes***	***Yes***	***Yes***	***Yes***	***Yes***	***Yes***	***Yes***	***Yes***	***Yes***
106	61	136	71	60	75	114	88	39	94
No	***No***	***No***	***No***	***No***	***No***	***No***	***No***	***No***	***No***
74	113	177	119	48	99	97	72	72	109
Gini True	**Gini True**	**Gini True**	**Gini True**	**Gini True**	**Gini True**	**Gini True**	**Gini True**	**Gini True**	**Gini True**
0,4841975309	0,455344167	0,4914207555	0,4680886427	0,4938271605	0,4904875149	0,4967543406	0,495	0,4558071585	0,4972700138
False	**False**	**False**	**False**	**False**	**False**	**False**	**False**	**False**	**False**
Yes	***Yes***	***Yes***	***Yes***	***Yes***	***Yes***	***Yes***	***Yes***	***Yes***	***Yes***
97	142	67	132	143	128	89	115	164	109
No	***No***	***No***	***No***	***No***	***No***	***No***	***No***	***No***	***No***
129	90	26	84	155	104	106	131	131	94
Gini False	**Gini False**	**Gini False**	**Gini False**	**Gini False**	**Gini False**	**Gini False**	**Gini False**	**Gini False**	**Gini False**
0,4899757225	0,4748810939	0,4028211354	0,475308642	0,4991892257	0,4946492271	0,4961998685	0,4978848569	0,4937431772	0,4972700138
Gini Total	**Gini Total**	**Gini Total**	**Gini Total**	**Gini Total**	**Gini Total**	**Gini Total**	**Gini Total**	**Gini Total**	**Gini Total**
0,4874139626	0,4665081252	0,4711257687	0,4719298246	0,4977628635	0,4928656361	0,4964880301	0,4967479675	0,4833715071	0,4972700138

Elaboração da Árvore - Manual

- Implementar manualmente a árvore;
- Após recomendação do professor, retiramos a coluna **price** e **yr_built** e tivemos que recomeçar todo o processo;
- Alta complexidade;
- Após a tentativa, optamos pela utilização da biblioteca SKLEARN.

Elaboração da Árvore - SKLEARN

- Formatação da base de dados para o formato .CSV

```
Quartos,Banheiros,Area construida,Area do lote,Andares,Estado de conservacao,Nivel de construcao,Porao,Caro
1,1,1,0,1,0,1,0,1
1,1,1,0,0,1,1,1,0
0,1,1,0,1,0,1,1,0
0,1,0,0,1,1,0,0,0
1,1,1,1,1,0,1,1,0
```

Elaboração da Árvore - SKLEARN

- **Implementação do código:**

```
import pandas as pd
import numpy as np
from sklearn import tree

train = pd.read_csv('ArvoreDefinitiva.csv')
y_train = train['Caro']
x_train = train.drop(['Caro'], axis=1).values
decision_tree = tree.DecisionTreeClassifier(max_depth = 20)
decision_tree.fit(x_train, y_train)

with open("ArvoreDefinitiva.dot", 'w') as f:
     f = tree.export_graphviz(decision_tree,
                              out_file=f,
                              max_depth = 20,
                              impurity = True,
                              feature_names = list(train.drop(['Caro'], axis=1)),
                              class_names = ['False', 'True'],
                              rounded = True,
                              filled= True )
```

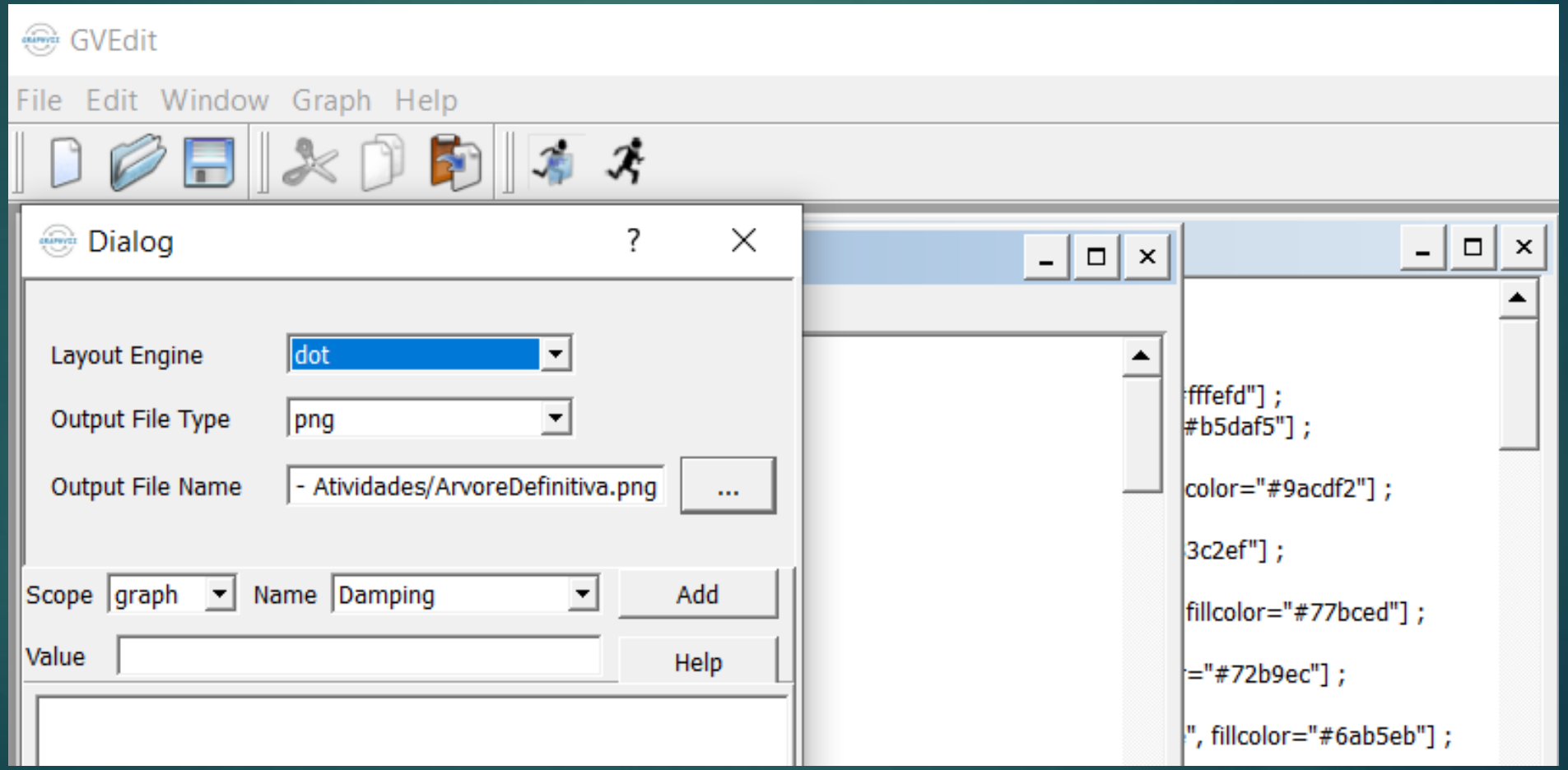
Elaboração da Árvore - SKLEARN

- **Arquivo .DOT:**

```
digraph Tree {
node [shape=box, style="filled, rounded", color="black", fontname=helvetica] ;
edge [fontname=helvetica] ;
0 [label="Quartos <= 0.5\ngini = 0.5\nsamples = 396\nvalue = [199, 197]\nclass = False", fillcolor="#fffefd"] ;
1 [label="Andares <= 0.5\ngini = 0.474\nsamples = 231\nvalue = [89, 142]\nclass = True", fillcolor="#b5daf5"] ;
0 -> 1 [labeldistance=2.5, labelangle=45, headlabel="True"] ;
2 [label="Area construida <= 0.5\ngini = 0.442\nsamples = 167\nvalue = [55, 112]\nclass = True", fillcolor="#9acdf2"] ;
1 -> 2 ;
3 [label="Porao <= 0.5\ngini = 0.397\nsamples = 143\nvalue = [39, 104]\nclass = True", fillcolor="#83c2ef"] ;
2 -> 3 ;
```

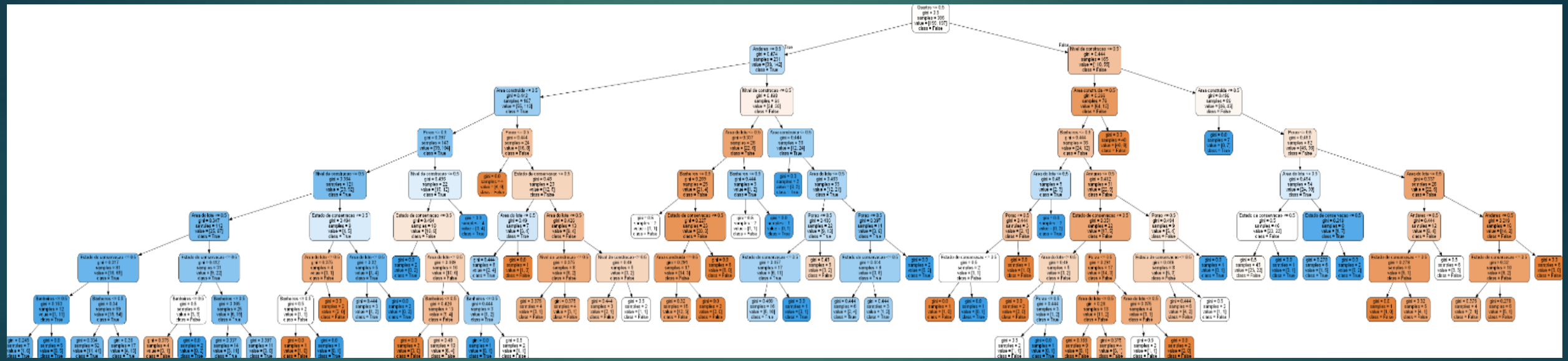
Elaboração da Árvore - SKLEARN

- **Arquivo .PNG - Utilizamos o GRAPHVIZ (gvedit.exe) para conversão .DOT -> .PNG**



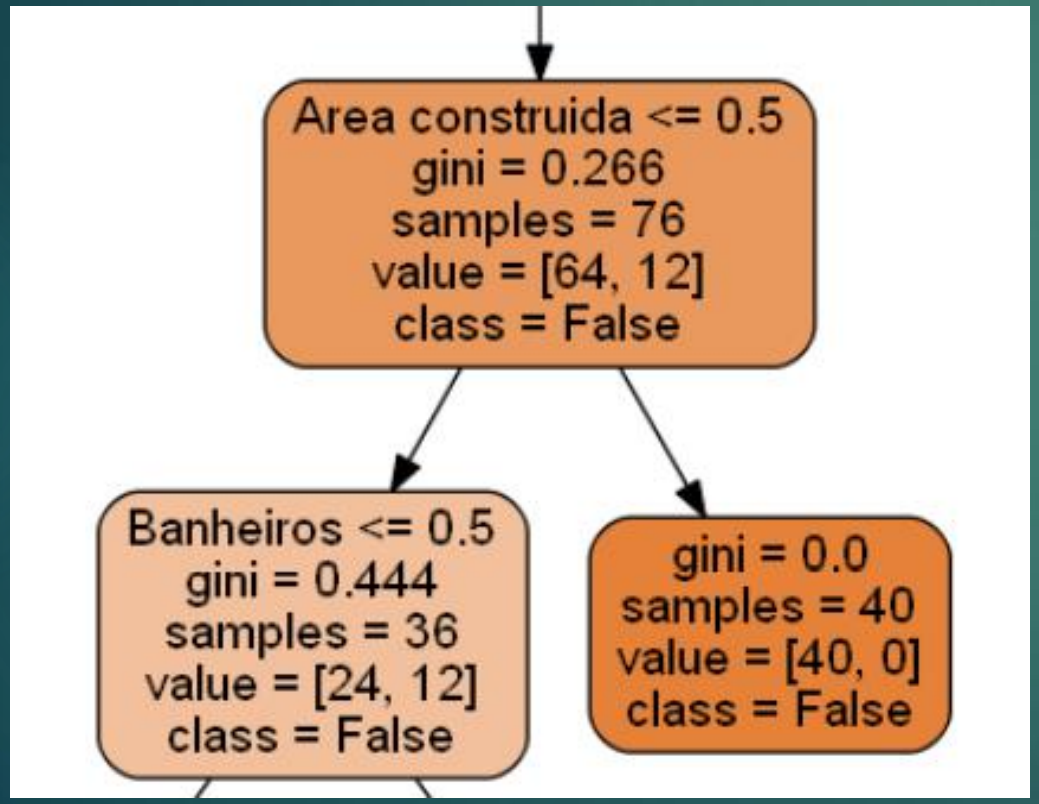
Elaboração da Árvore - SKLEARN

- **Gráfico da árvore:**



Elaboração da Árvore - SKLEARN

- **Gráfico da árvore:**



Implementação do código - CSV

- Utilização da árvore gerada para determinar os ID;
- Elaboração do .CSV que será utilizado no código:

```
1  ID,Pergunta,A,Nó A,B,Nó B
2  1,Mais que 3 quartos? ,SIM,2,NÃO,3
3  2,Mais que 1 andar? ,SIM,4,NÃO,5
4  3,Nível de construção maior que 7? ,SIM,6,NÃO,7
5  4,Área construída maior que 2017 metros quadrados? ,SIM,8,NÃO,9
6  5,Nível de construção maior que 7? ,SIM,10,NÃO,11
7  6,Área construída maior que 2017 metros quadrados? ,SIM,12,NÃO,13
8  7,Área construída maior que 2017 metros quadrados? ,SIM,14,NÃO,15
9  8,Tem porão? ,SIM,16,NÃO,17
10 9,Tem porão? ,SIM,18,NÃO,19
11 10,Área do lote maior que 9566 metros quadrados? ,SIM,20,NÃO,21
12 11,Área construída maior que 2017 metros quadrados? ,SIM,22,NÃO,23
13 12,Mais que 1 banheiro? ,SIM,24,NÃO,25
14 13,NÓ FOLHA,Caro! (Acima do valor de mercado).,,,
15 14,NÓ FOLHA,Barato! (Abaixo do valor de mercado).,,,
```

Implementação do código - Python

- Utilização do modelo proposto pelo professor;
- Modificação na formatação das saídas;
- Realização dos 10 casos testes;
- Taxa de acerto de 40% (4/10)