Cientista de Dados Jr.

Desafio Técnico

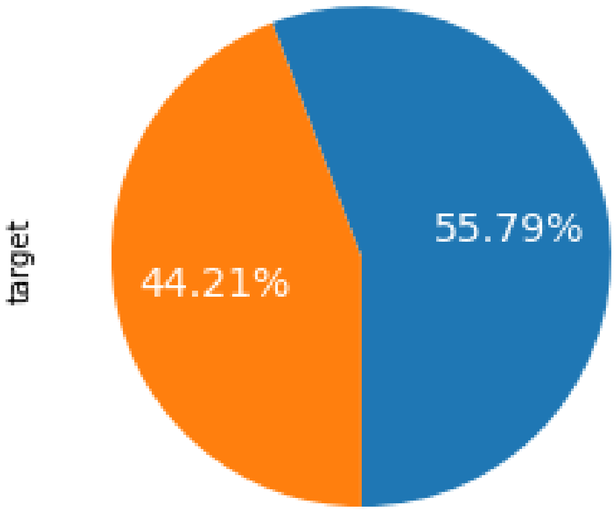
Leonardo Gabriel Ferreira Rodrigues

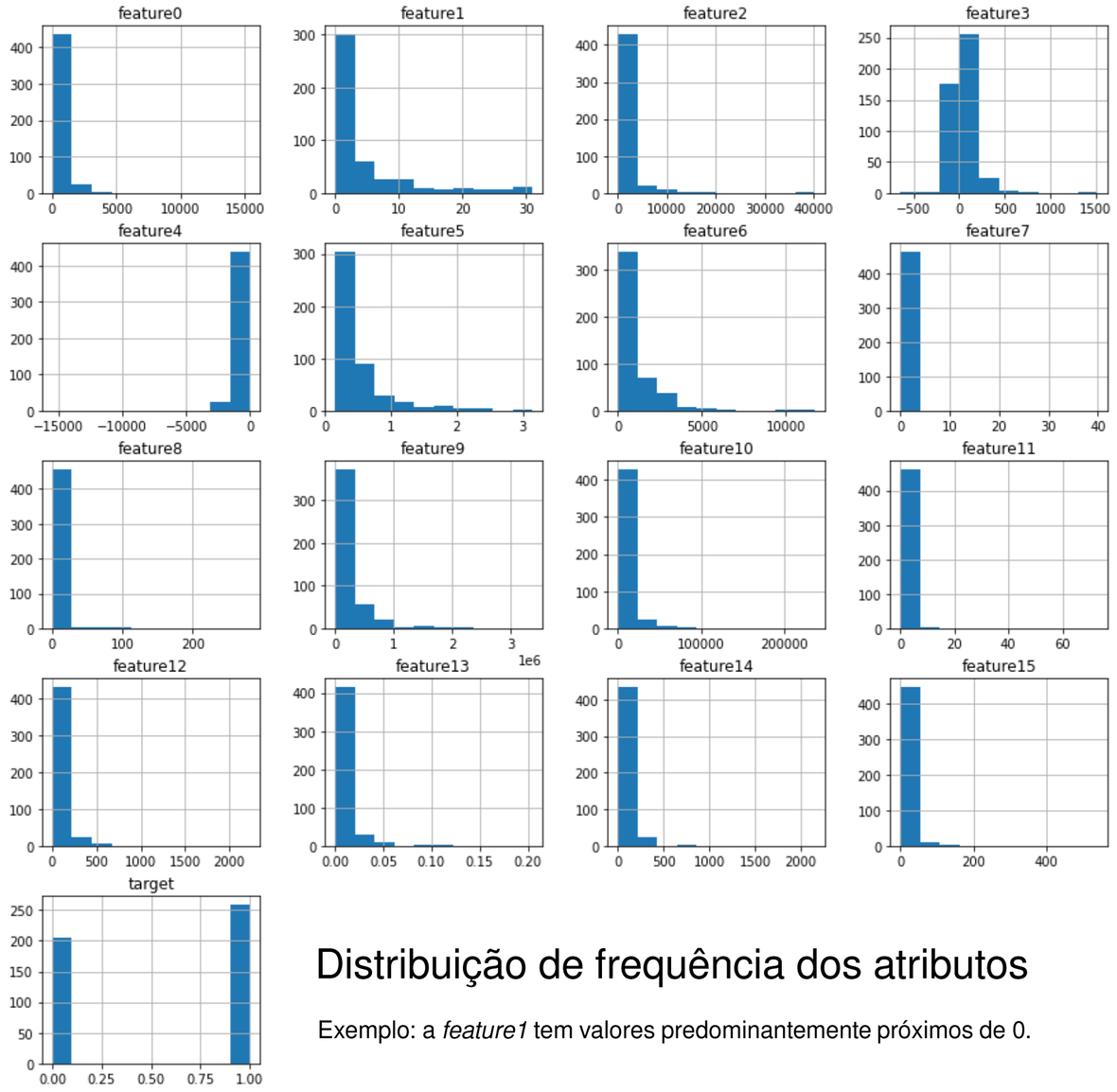
Conteúdo

☑ Análise Exploratória dos Dados

Descrição do Conjunto de Dados

Classe	Quantidade
0	206
1	260





Distribuição de frequência dos atributos

Exemplo: a *feature1* tem valores predominantemente próximos de 0.

☑ Análise Exploratória dos Dados

A matriz de correlação permite verificar os níveis de associação entre as *features*.

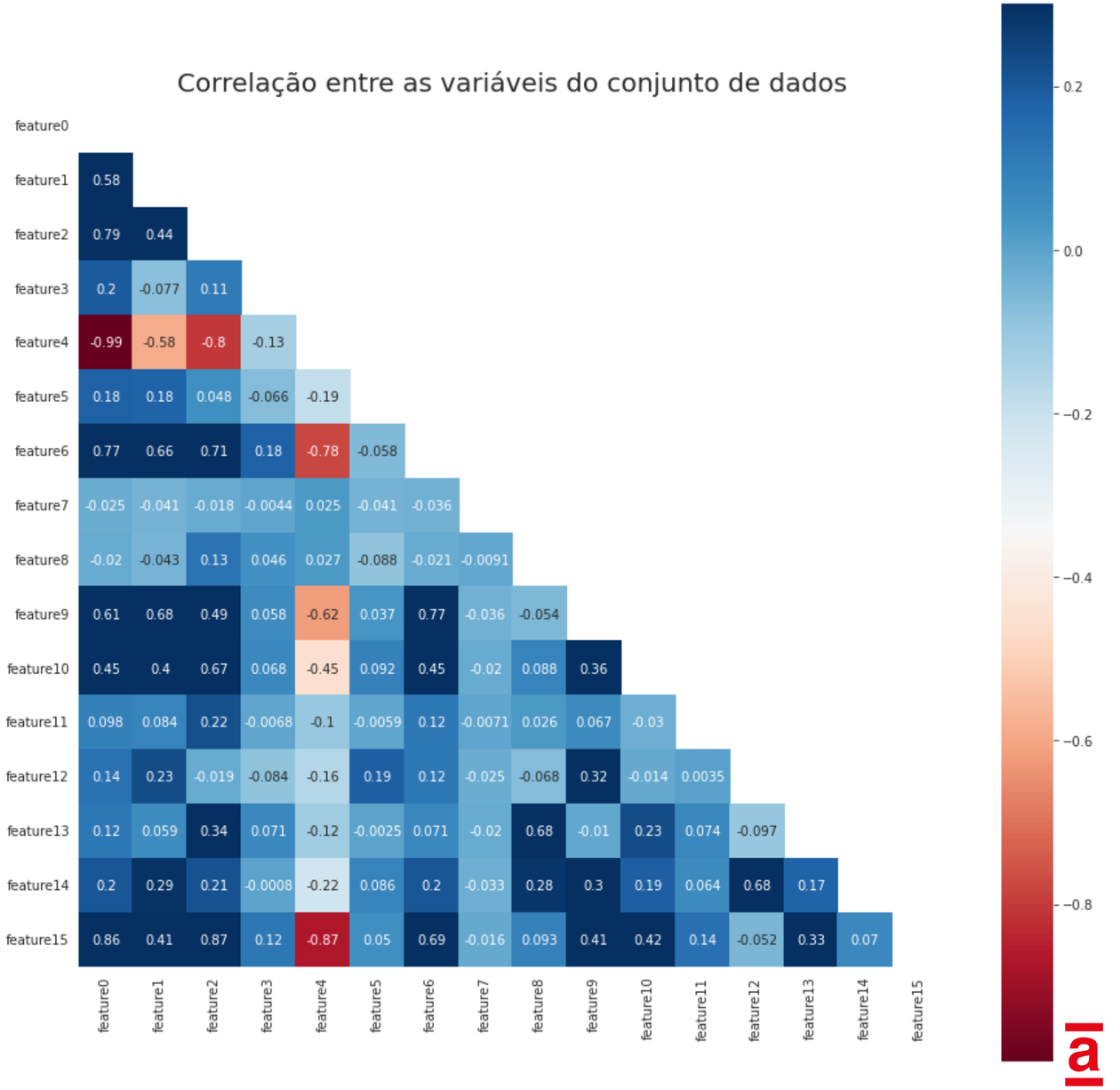
O coeficiente de correlação é uma medida de associação linear entre duas variáveis.

-1	+1
associação negativa perfeita	associação positiva perfeita

Exemplo:

Correlação Negativa
feature0 e feature4: quando feature0 aumenta, feature4 diminui

Correlação Positiva
feature0 e feature15: quando feature0 aumenta, feature15 também aumenta



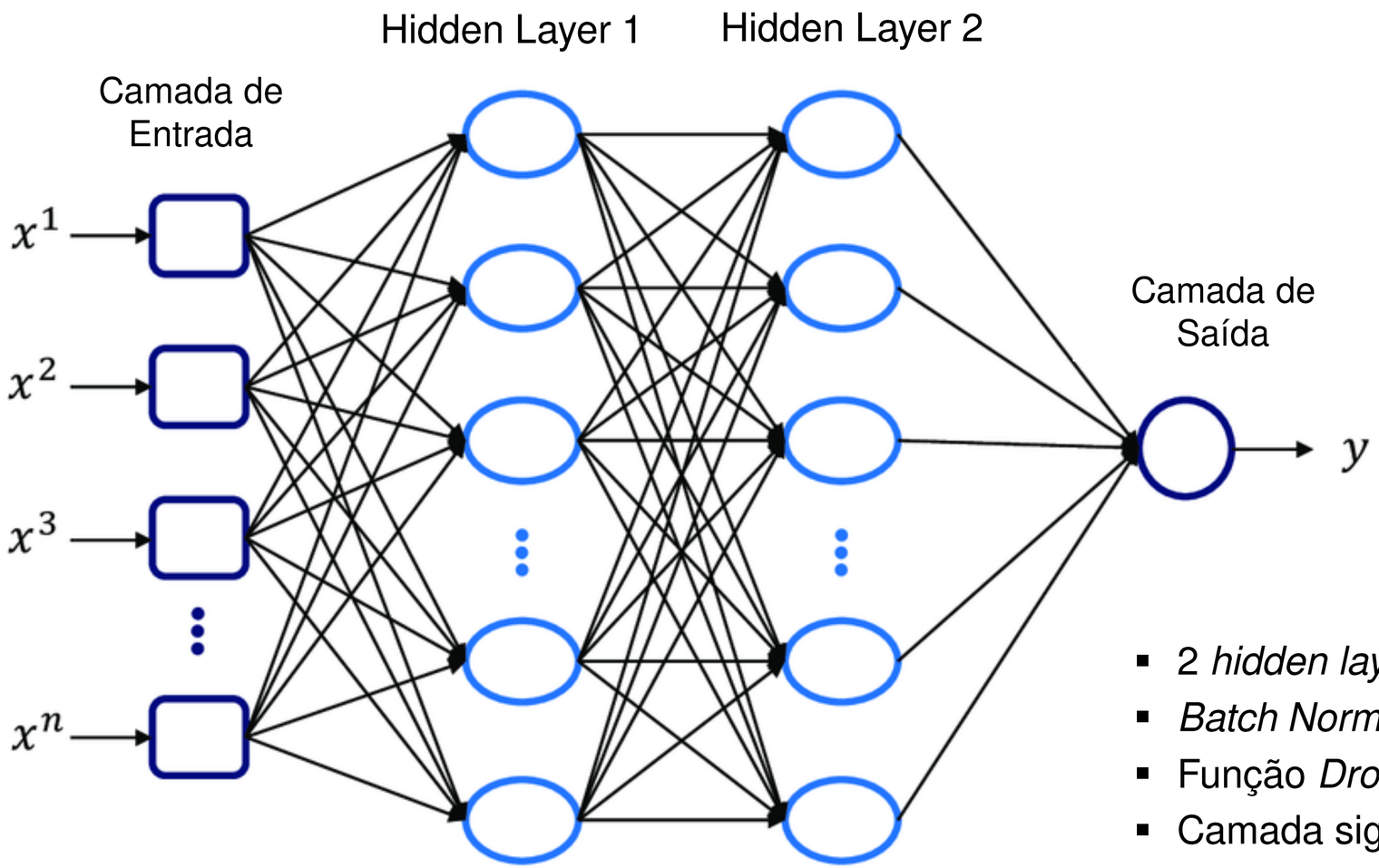
☑ Preparação dos Dados

- Normalização dos Dados
- Atribuições:
  - X: Dados
  - y: rótulos
- Exclusão da coluna “target” para classificação
- Distribuição dos Dados:

Treino 75%	Teste 25%

☑ Modelagem

Criação de um Multi-Layer Perceptron (MLP)
AmericanasModel



- 2 *hidden layers* (camadas ocultas)
- *Batch Normalization* entre cada camada
- Função *Dropout* antes de enviar a saída
- Camada sigmoide para realizar a classificação

☑ Avaliação e Performance do Modelo

Classe	Precisão (%)	Recall (%)	F1-Score (%)
0	65,52	73,68	69,42
1	71,70	63,33	67,26

Acurácia: 68,38%

		Valor Predito	
		0	1
Valor Real	0	**42**	15
	1	22	**38**

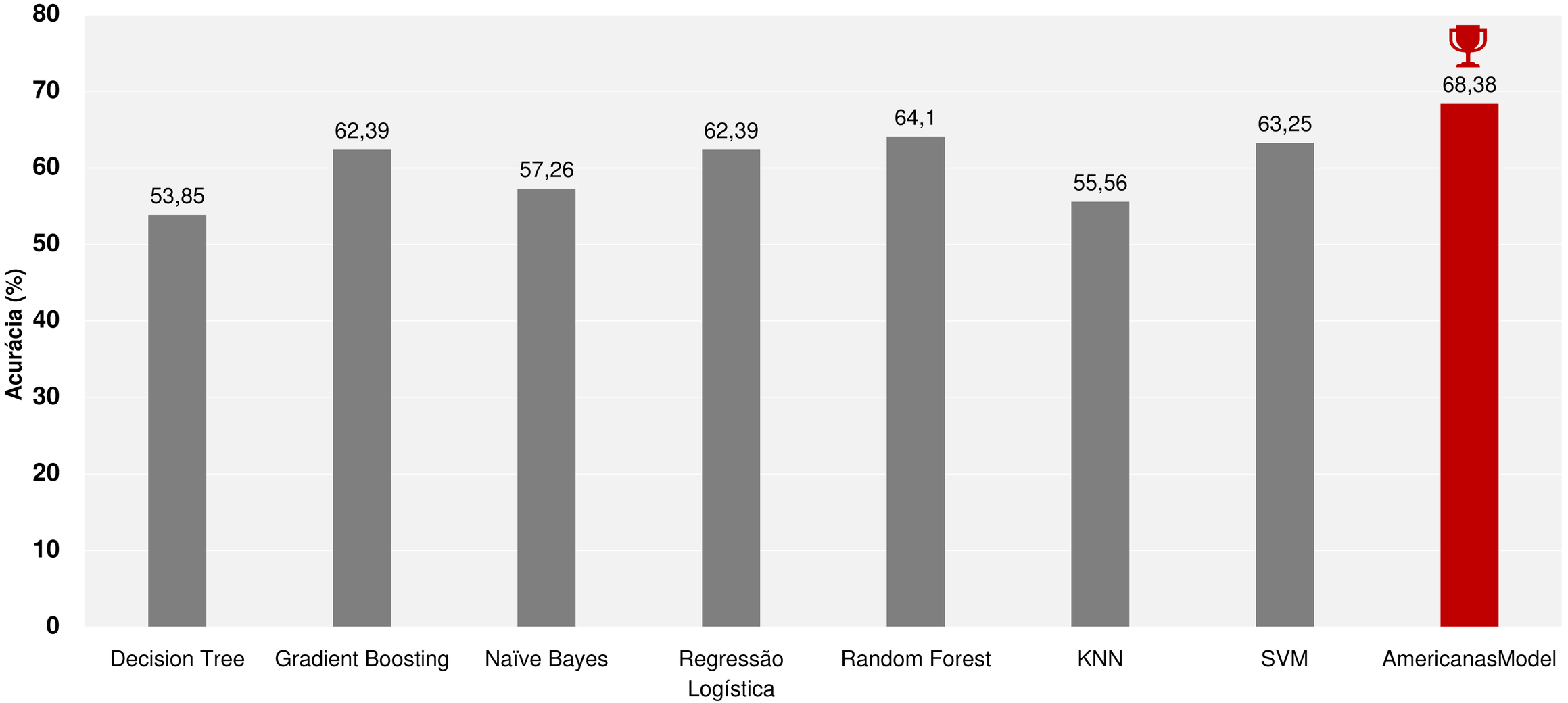
Matriz de confusão considerando o conjunto de teste

☑ Avaliação e Performance do Modelo

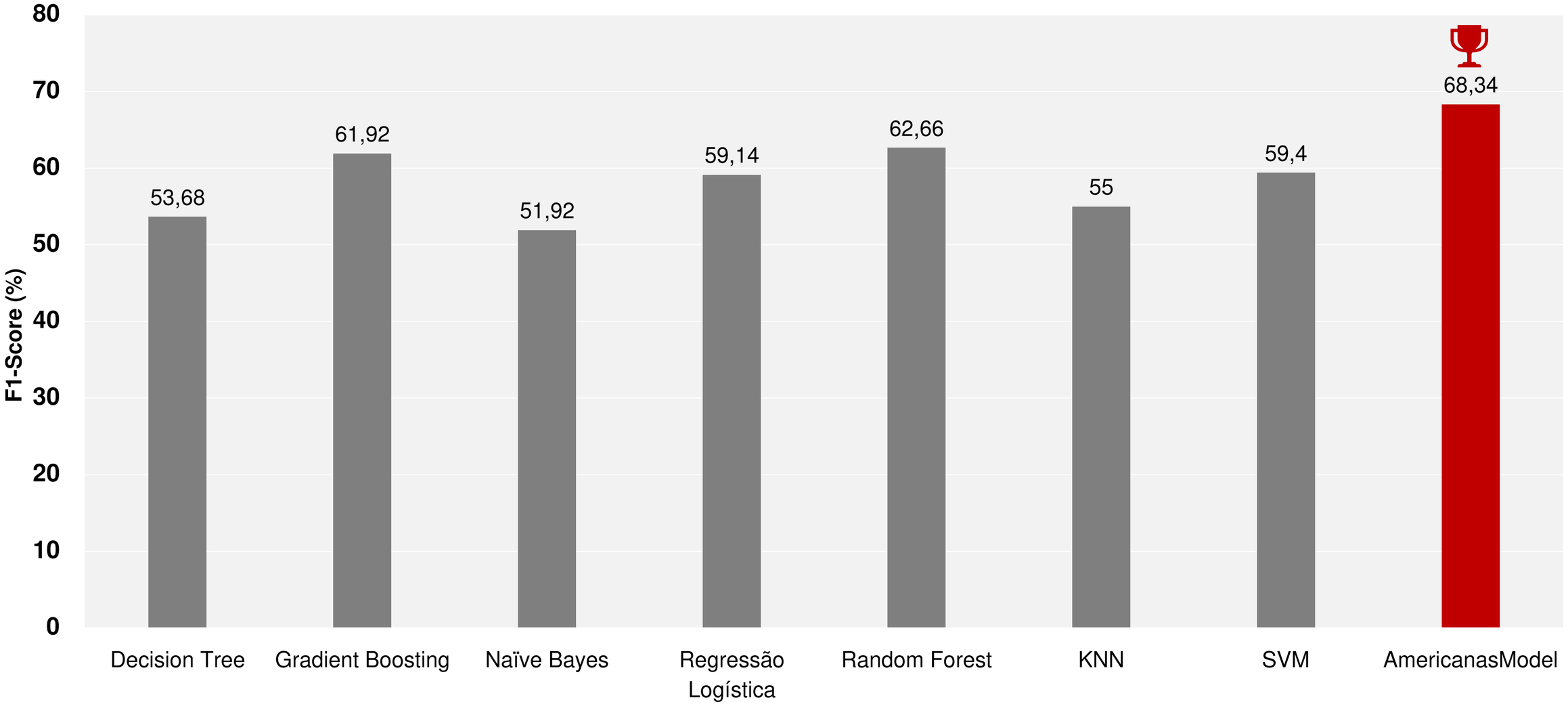
Análise Comparativa

Modelo	Acurácia (%)	Precisão (%)	Recall (%)	F1-Score (%)
Decision Tree	53,85	53,76	53,73	53,68
Gradient Boosting	62,39	62,33	62,15	61,92
Naïve Bayes	57,26	60,94	56,45	51,92
Regressão Logística	62,39	66,48	61,71	59,14
Random Forest	64,10	65,76	63,34	62,66
KNN	55,56	55,52	55,31	55,00
SVM (kernel RBF)	63,25	69,15	62,50	59,40
AmericanasModel	**68,38**	**68,66**	**68,51**	**68,34**

Avaliação e Performance do Modelo



Avaliação e Performance do Modelo



☑ Conclusão

- Utilizar apenas a **acurácia** como métrica **não é uma boa alternativa**, principalmente em casos de Falso Positivo, quando o modelo prevê a ocorrência do evento 1 quando o valor real é 0.
- Para superar essa limitação, as métricas **recall e F1-score são mais adequadas**, pois classificar o evento 1 como se fosse 0, pode gerar prejuízo para Americanas S.A.

AmericanasModel

- É mais adequado para a classificação do conjunto de dados proposto.
- Demonstrou robustez ao lidar com dados não linearmente separáveis.
- Superou os classificadores tradicionais.
- Pode ser implantado em ambientes de produção.

Cientista de Dados Jr.

Desafio Técnico

Leonardo Gabriel Ferreira Rodrigues

desafio-tecnico

leonardogfrodrigues

leonardogfrodrigues@gmail.com