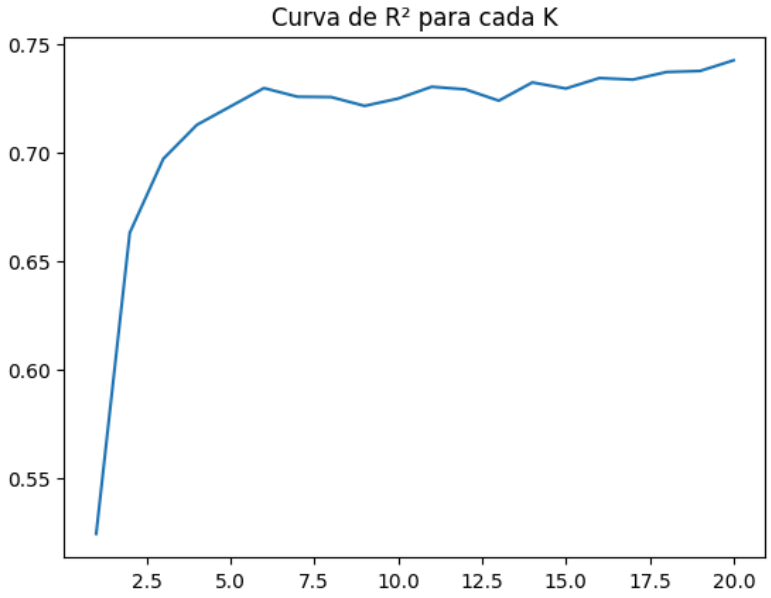
Relatório sobre o Desafio Prático

O dataset representa dados de usuários que interagiram com um site de imóveis e a interação final foi registrada como "comprou" ou "não comprou". Com pouco mais de 200 linhas disponíveis, o conjunto de dados apresentava alguns problemas que precisavam ser resolvidos para a finalização do desafio.

Problemas

Nos dados, havia uma coluna que representava o tempo em minutos que os usuários passaram no site. No entanto, observou-se um problema, pois alguns usuários aparentavam ter passado apenas 1 minuto no site, o que parecia incoerente. Por segurança, optei por remover todas as linhas correspondentes a esses casos (5 no total), uma vez que poderia haver um problema na coleta dos dados desses usuários e outros dados poderiam estar comprometidos.

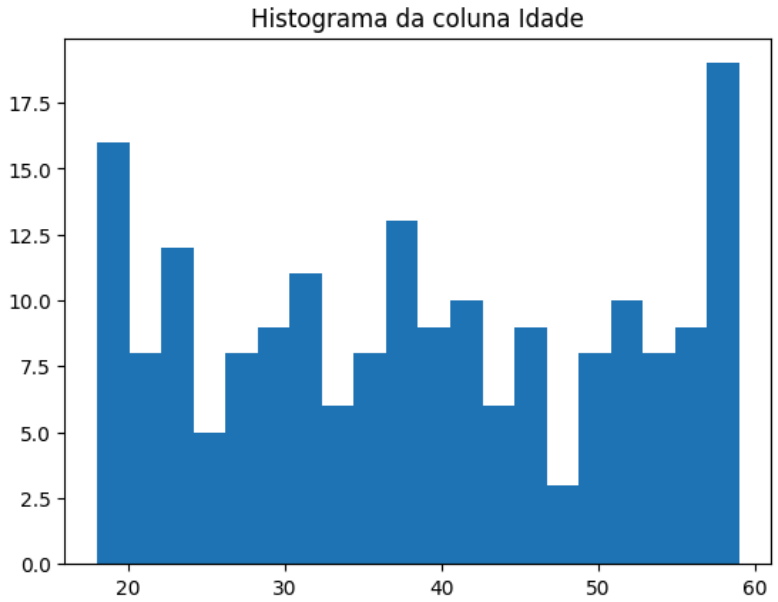
Além disso, havia 30 usuários com valores ausentes, distribuídos quase que uniformemente entre as variáveis explicativas. Considerando a quantidade restrita de dados disponíveis para alimentar os modelos, concluí que não seria viável simplesmente descartar essas linhas. Assim, realizei uma imputação dos valores ausentes utilizando o método K-Nearest Neighbors Imputer, com a busca pelo valor de k que maximizasse o R^2, mas sem ser tão grande a ponto de causar sobreajuste (overfitting) e viés nos dados. A imputação foi então realizada com base nessa metodologia.



Análise Exploratória

Antes da imputação de valores faltantes, foi realizada uma Análise Exploratória com o objetivo de compreender a distribuição dos dados, as correlações e possíveis padrões. Um

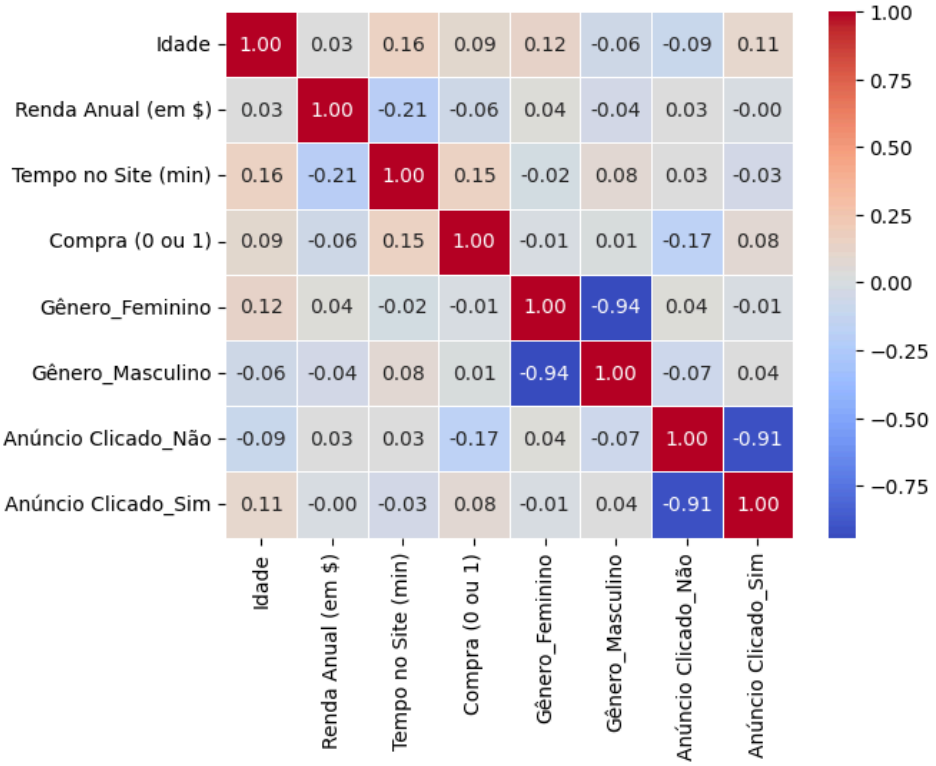
dos destaques foi a boa variância observada em algumas variáveis explicativas, como Idade e Tempo no Site, o que pode contribuir para o desempenho de modelos preditivos.



Nota-se uma boa distribuição das idades, com um intervalo que vai de 18 a 59 anos. Apesar desse teto, a variância é suficiente para ser explorada por modelos paramétricos.

Por outro lado, há um desequilíbrio nas classes-alvo: o número de exemplos de não compradores é pelo menos o dobro do número de compradores, o que representou um desafio na modelagem.

Ao analisar as correlações, fica evidente que não há padrões facilmente detectáveis por modelos mais simples, o que reforça a necessidade de técnicas mais avançadas para capturar as relações existentes nos dados.



Observou-se uma pequena correlação negativa entre a renda e o tempo gasto no site. Esse padrão sugere que usuários com maior renda tendem a passar menos tempo analisando preços e diferentes opções, possivelmente realizando compras de forma mais rápida. Esses usuários poderiam ser incentivados a realizar compras ainda mais ágeis por meio de anúncios direcionados e estratégias de marketing específicas.

Entretanto, devido à baixa quantidade de dados disponíveis e ao desbalanceamento da amostra, é possível que algumas correlações relevantes tenham sido omitidas, não sendo detectadas nesta análise.

Modelagem

Diante da vasta quantidade de modelos de aprendizado de máquina disponíveis, optou-se por restringir a análise a três candidatos: Random Forest, Regressão Logística e Decision Tree. Para comparar o desempenho desses modelos, foi utilizada uma metodologia estatística paramétrica composta pelas seguintes etapas:

- 1º: Cross-validation com 10 folds aplicado a todos os modelos, avaliando as métricas F1-Score, Recall, Precisão e Acurácia.
- 2º: Aplicação de um Teste ANOVA para cada métrica, buscando identificar diferenças estatísticas entre os grupos.
- 3º: Realização de um Teste pós-hoc de Tukey para determinar quais modelos apresentaram desempenho superior.

Entretanto, a metodologia enfrentou limitações devido ao não atendimento da suposição de normalidade nas métricas de desempenho, causado principalmente pelo baixo volume de dados e pelo sub-ajuste dos modelos.

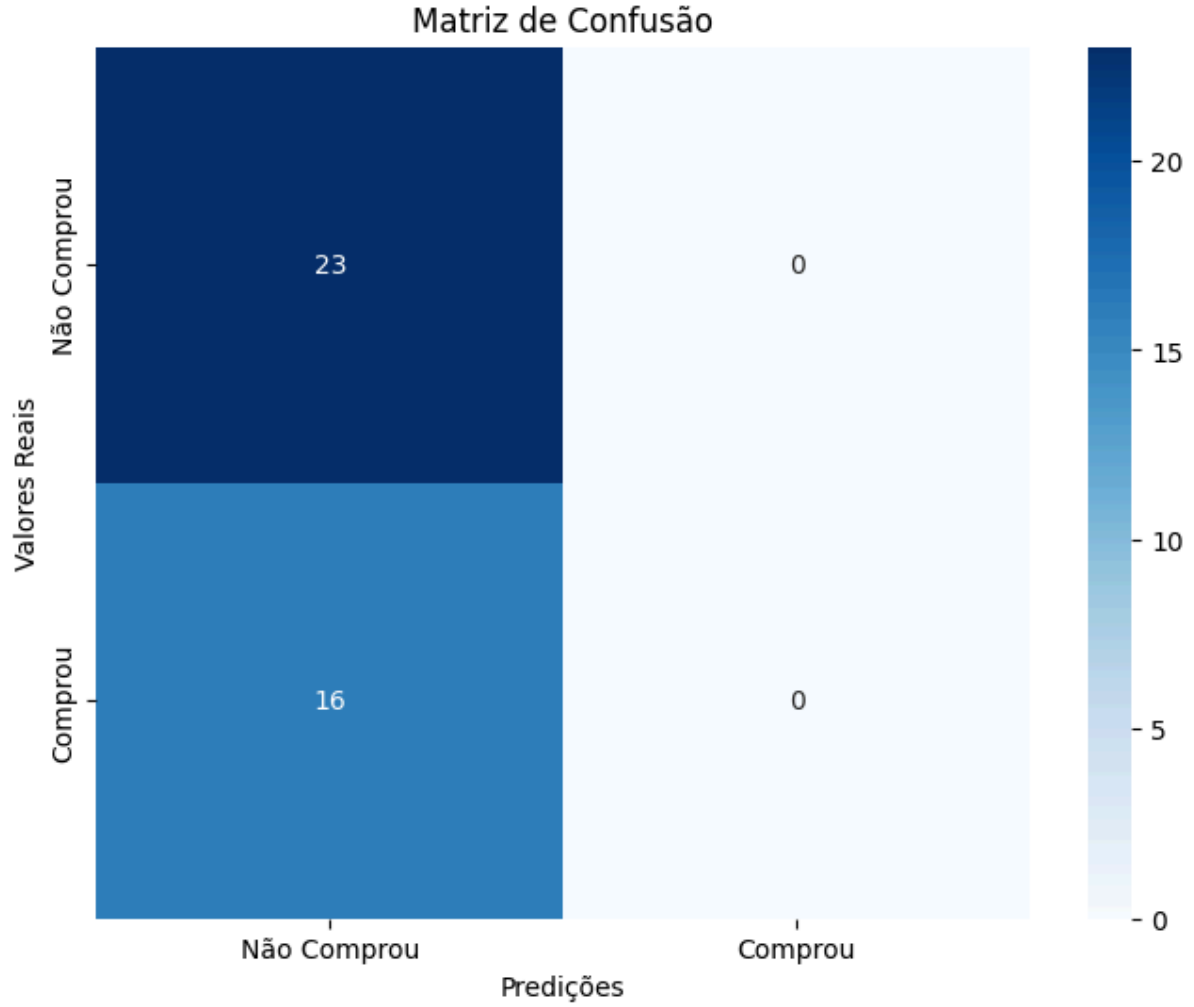
Os resultados dos testes indicaram que tanto o Random Forest quanto a Regressão Logística apresentaram desempenho superior à Decision Tree. Contudo, não foi possível distinguir estatisticamente uma diferença significativa entre Random Forest e Regressão Logística.

```
model_1 = RandomForestClassifier()
model_2 = LogisticRegression()
model_3 = DecisionTreeClassifier()

Teste de Tukey Acurácia:
  Multiple Comparison of Means - Tukey HSD, FWER=0.05
=====================================================
 group1  group2 meandiff p-adj   lower   upper  reject
-----------------------------------------------------
model_1 model_2   0.0133 0.9473  -0.088  0.1147  False
model_1 model_3  -0.1156 0.0214 -0.2169 -0.0142   True
model_2 model_3  -0.1289 0.0089 -0.2303 -0.0275   True
-----------------------------------------------------
```

Devido à maior possibilidade de explorar o papel de cada feature no resultado proporcionada pela Regressão Logística, optou-se por utilizá-la como modelo final.

No teste final, realizado com dados inéditos previamente separados, observou-se que o modelo foi impactado pelo desbalanceamento de classes. A matriz de confusão apresentada abaixo evidencia a incapacidade do modelo de solucionar o problema adequadamente.



A matriz indica que os dados disponíveis não são suficientes para treinar um modelo capaz de realizar previsões com algum nível de precisão. Para minimizar o erro do tipo 1, foi assumido que todos os usuários não são compradores. Como consequência, nenhuma previsão de comprador foi feita, tornando impossível o cálculo da métrica de Precisão, já que ela depende de pelo menos uma previsão para cada classe (neste caso, não houve previsões de compradores).

Além disso, métricas como Recall e F1-Score também resultaram em zero, pois não houve nenhuma previsão de Verdadeiro Positivo. Tentativas de ajustar o hiperparâmetro de balanço de classes não foram bem-sucedidas, resultando apenas em alterações na quantidade de Falsos Negativos e Verdadeiros Negativos.

Conclusão

Não foi possível construir um modelo suficientemente robusto para prever possíveis compradores ou interpretar as características que levam um possível comprador a finalizar a compra. Isso provavelmente se deve à baixa quantidade de dados disponíveis e à possibilidade de a amostra utilizada não representar adequadamente a população real.