INTRODUÇÃO

O poder do big data e da ciência de dados(data science) está revolucionando o mundo atual. Das empresas ao nosso estilo de vida digital, as informações do data science nos levam à mudanças e benefícios nas mais diversas áreas. Neste trabalho, analisaremos um banco de dados de registros de admissão universitária nos utilizando de dois algoritmos conhecidos: K-NN(k- vizinhos próximos) e Árvore de Decisão. Partindo destas análises, obtivemos resultados de flagrante utilidade para nosso aprendizado dos conceitos trabalhados na disciplina de introdução à ciência de dados.

Fundamentos Teóricos e Metodológicos

A classificação é um modo de aprendizagem de máquina supervisionada: o algoritmo de classificação "aprende" com os dados rotulados(dados que já possuem a saída que se quer prever). Tais rótulos colaboram com os modelos para que tomem a decisão com base em regras lógicas bem definidas. Um algoritmo de agrupamento básico, como o método dos K-vizinhos próximos(KNN), ajuda a antever subgrupos contidos nos conjuntos de dados não rotulados. O K-NN é um classificador de aprendizagem de máquina supervisionado que se utiliza das observações que ele memoriza em um conjunto de dados de teste para antever as classificações aplicáveis à observações novas e não rotuladas. O K-NN faz previsões se utilizando da *semelhança* quanto mais as observações de treinamento se assemelham com as observações de entrada, muito provavelmente o classificador irá atribuí-las a uma classe igual. Para usar o K-NN, precisamos escolher um ponto de consulta na base de dados da amostra e calcular os k-vizinhos adjacentes até esse ponto. O ponto de consulta é classificado com um rótulo igual ao da maioria dos k pontos mais próximos em volta dele. Os K-vizinhos próximos são quantificados pela distância ou semelhança com base em outro atributo quantitativo. Vamos entender como isso funciona com um exemplo retirado de [1]:
Um conjunto de dados é representado por [1,1,4,3,5,2,6,2,4] e o ponto de consulta é igual a 5. Determinando que k é 3, pela distância, haveria três vizinhos mais próximos do ponto 5(pontos 4,4 e 6). Assim, de acordo com o algoritmo K-NN, o ponto de consulta será o 4. De forma idêntica, o K-NN continua definindo outros pontos de consulta se utilizando do mesmo princípio da maioria. A figura abaixo mostra como o K-NN se aplicaria a este caso.

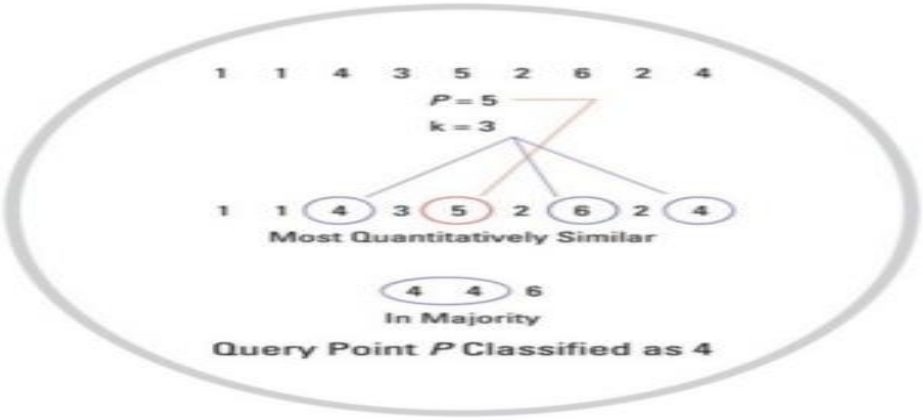


Figura 1: Funcionamento básico do K-NN

Uma estrutura de árvore é utilizada como ferramenta de suporte de decisão. Podemos utilizá-la para fazer modelos que antevejam prováveis consequências associadas a uma certa decisão. O algoritmo da árvore de decisão produz um conjunto de regras do tipo sim ou não, que podemos aplicar aos dados e ver como serão descritos pelos modelo. Esse tipo de modelo é preciso ser usado com muita cautela, pois temos um alto risco de propagação de erros, que acontece quando alguma regra do modelo está errada.Tomando outro exemplo de [1]:Usando o modelo da árvore de decisão e usando a famosa base de dados dos passageiros do Titanic, é possível prever se um passageiro o Titanic era mulher ou homem, com família grande ou pequena , se ele/ela sobreviveu ao terrível acidente. Vejamos:

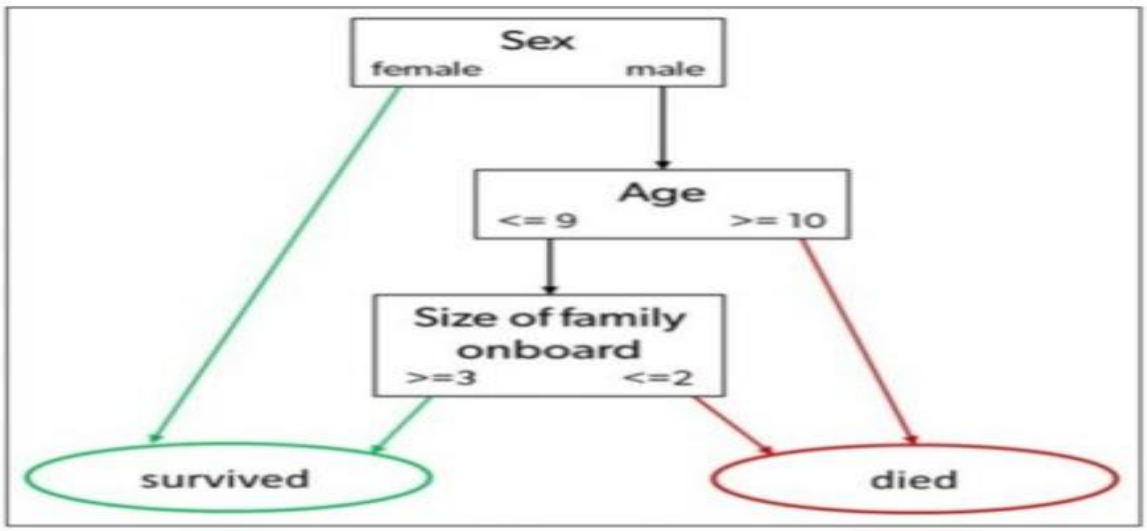


Figura 2: Utilização do algoritmo da árvore de decisão.

APLICAÇÃO

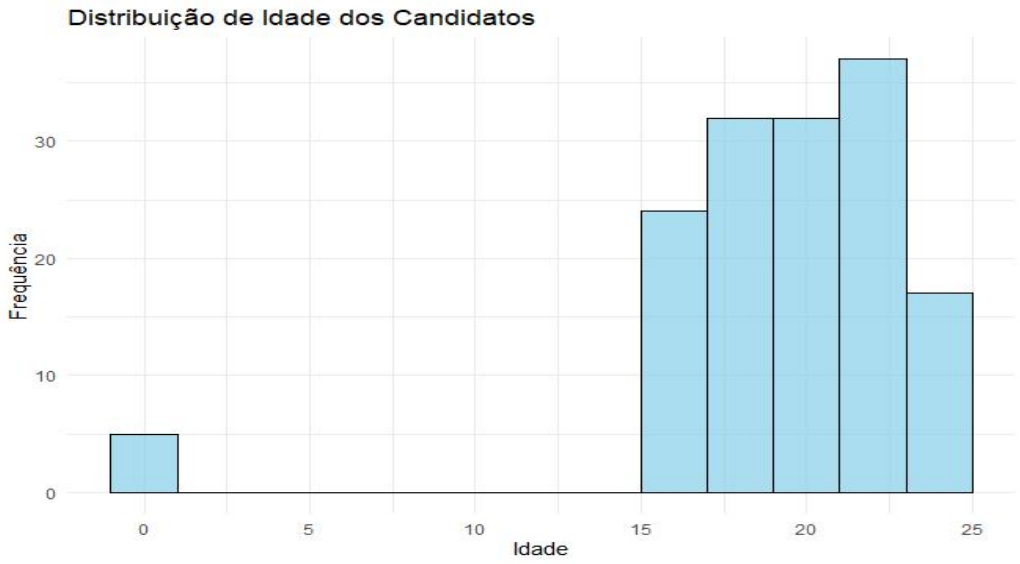
Segue abaixo a análise exploratória da base de dados bruta, com o objetivo de proporcionar uma visualização mais clara das informações, as quais serão fundamentais para a próxima etapa do processo, que envolve a aplicação dos modelos :

Tabela 1: resumo estatístico básico do conjunto de dados

	Age	Admission Test Score	High School Percentage
Min	-1.00	-5.00	-10.00
Q1	18.00	68.25	65.05
Median	20.00	79.00	77.55
Mean	19.68	77.66	75.68
Q3	22.00	89.00	88.31
Max	24.00	150.00	110.50
DP	4..540512	16.855343	17.368014

A idade média dos candidatos é **19,68 anos**, com uma mediana de **20 anos**, indicando que a maioria dos participantes está próxima dessa faixa etária. O desvio-padrão de **4,54 anos** mostra alguma variabilidade entre os candidatos. No entanto, foi identificado um valor mínimo de **-1**, o que representa uma inconsistência e deve ser corrigido. A pontuação média no teste de admissão é **77,66 pontos**, com uma mediana de **79 pontos** e um desvio-padrão de **16,86 pontos**, indicando uma dispersão moderada nos resultados. Contudo, valores mínimos de **-5** e máximos de **150** são

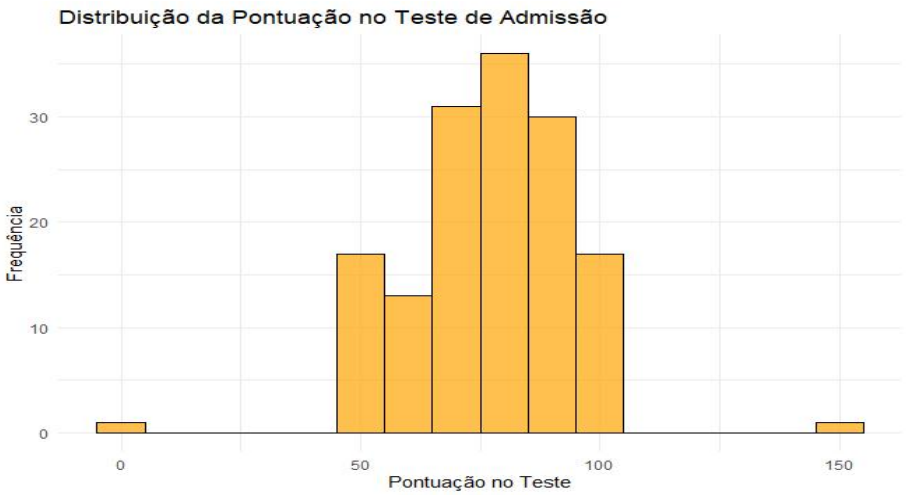
Gráfico 1: Distribuição da idade dos candidatos



Fonte: dados retirados do site kaggle

A distribuição de idade dos candidatos apresenta um padrão concentrado entre **17 e 24 anos**, com uma leve assimetria. O histograma mostra que a maioria dos estudantes que tentam a admissão se encontra na faixa etária de **17 a 21 anos.**

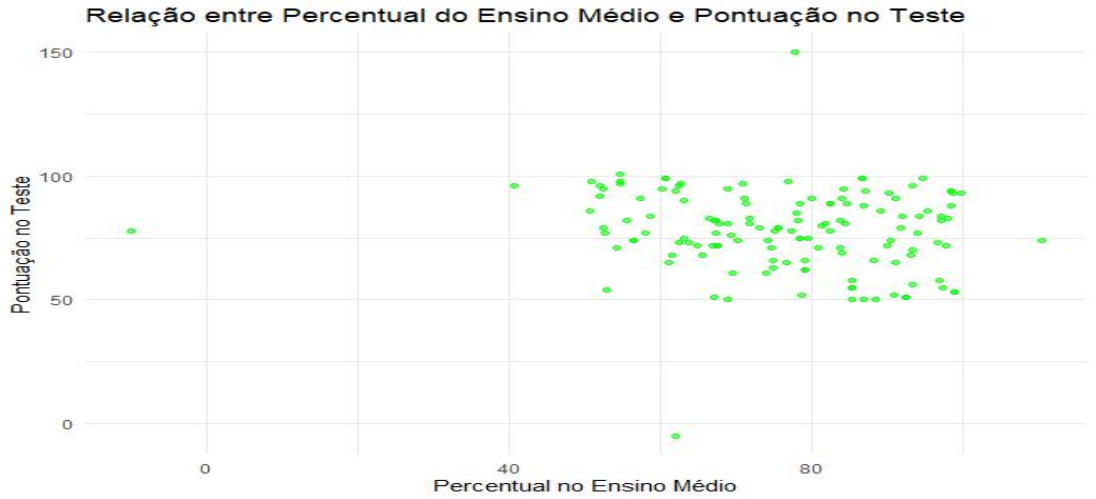
Gráfico 2: Distribuição da pontuação de admissão



Fonte: dados retirados do site kaggle

O histograma da pontuação do teste mostra que os candidatos apresentam pontuações distribuídas em um intervalo amplo, com um possível viés em direção a valores mais altos. A densidade maior está entre **50 e 100 pontos**, o que pode indicar que a maioria dos candidatos tem um desempenho moderado ou bom no exame.

Gráfico 3: Relação entre Percentual do Ensino Médio e Pontuação no Teste



Fonte: dados retirados do site kaggle

O gráfico de dispersão evidencia a relação entre o desempenho no ensino médio e a pontuação no teste de admissão. Espera-se uma **correlação positiva**, pois candidatos com notas mais altas no ensino médio tendem a ter bons desempenhos no teste. Entretanto,o gráfico sugere que não há uma relação forte entra eles, sendo uma correlação fraca ou inexistente.

A seguir será feita a aplicação dos modelos de machine learning já citados anteriormente para a classificação da variável **Admission Status,** será feita a comparação das métricas, gráfico de curva de aprendizado e a matriz confusão de ambos os modelos. Vale ressaltar que na análise exploratória havia dados incompletos,duplicados ou inexistentes e nesse passo isso tudo já foi tratado para não haver influência negativa ou positiva sobre os modelos.

Tabela 2: Métricas da Árvore de decisão

Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0.6296	0.6099	0.6296	0.6305	0.6265	0.2541	0.2570

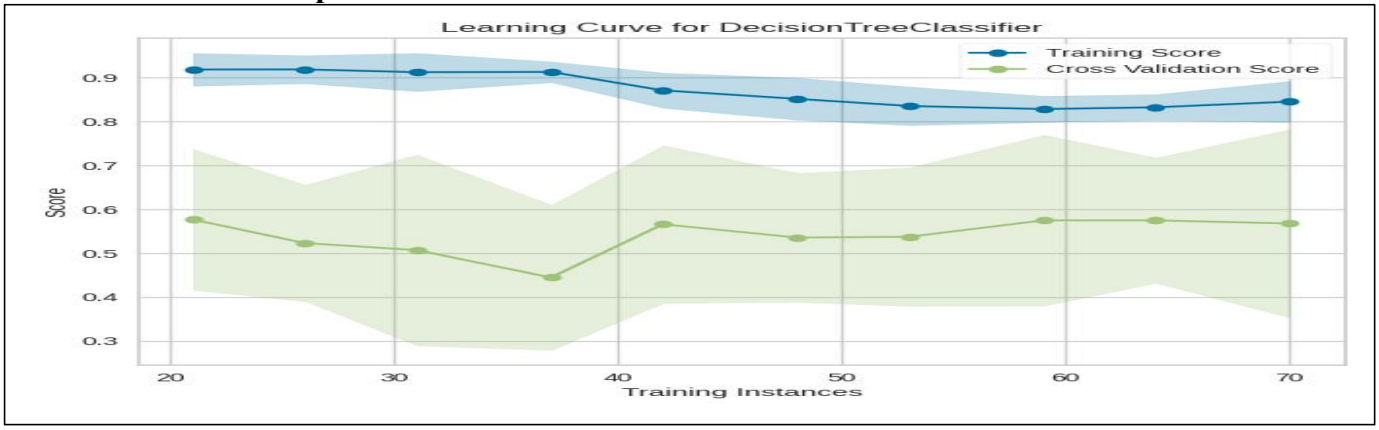
Fonte: dados retirados do site kaggle

Tabela 3: Métricas do KNN

Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0.5600	0.5673	0.5600	0.5590	0.5586	0.1158	0.1161

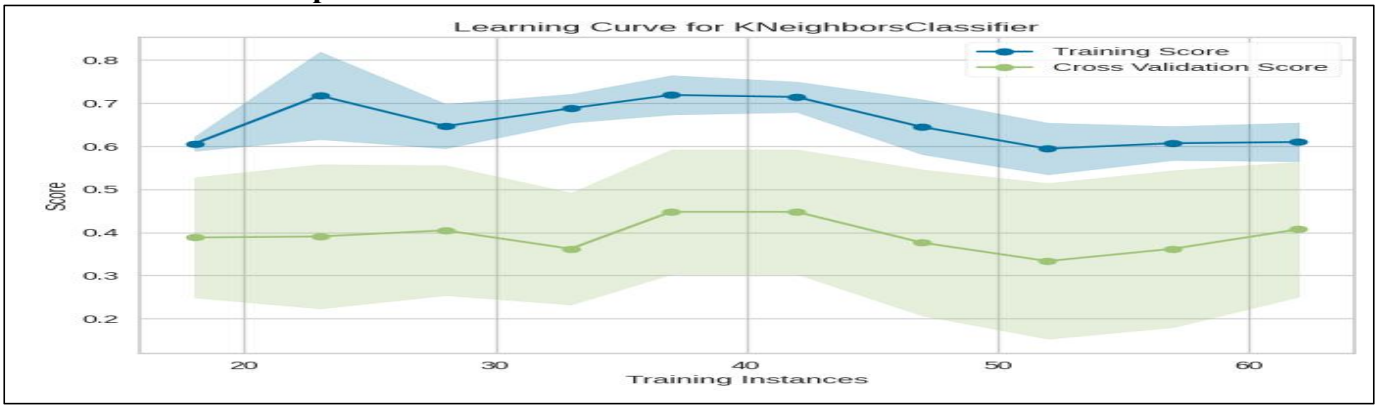
Fonte: dados retirados do site kaggle

Gráfico 4: Curva de aprendizado Árvore de decisão



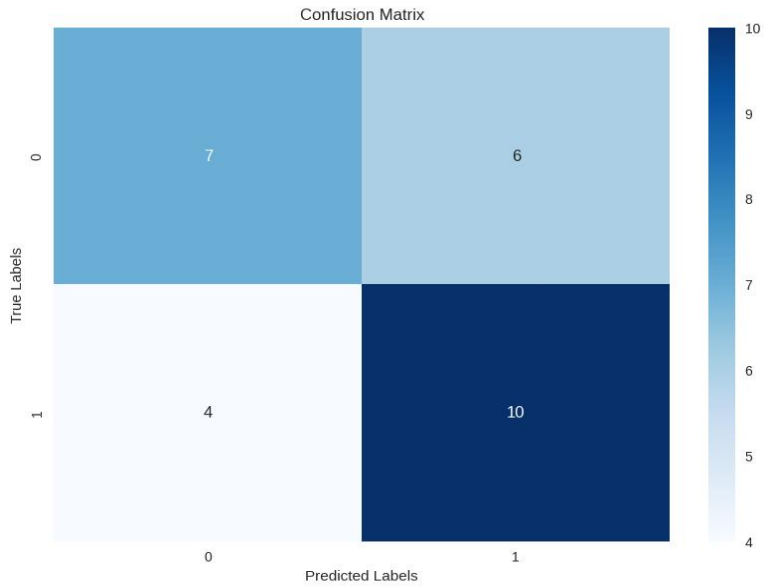
Fonte: base de dados retirados do site kaggle

Gráfico 5: Curva de aprendizado KNN



Fonte: base de dados retirados do site kaggle

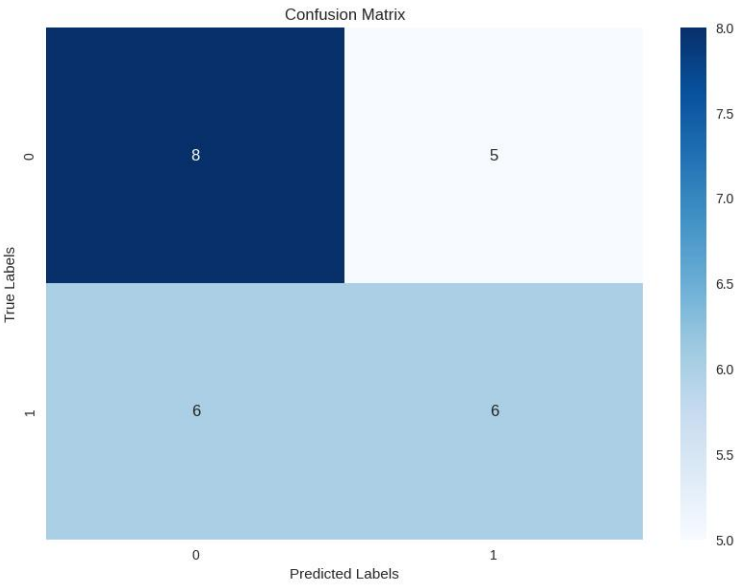
Gráfico 6: Matriz confusão Árvore de decisão



Accuracy:62.96%Kappa:0.25
F1(accepted):0.58,F1(rejected):0.67,
Recall(accepted):54%,Recall(rejected):71%,
Precision(accepted):64%,Precision(rejected):62%,
MediaMacro:0.62,MediaPonderada:0.63

Fonte: base de dados retirados do site kaggle

Gráfico 6: Matriz confusão KNN



Accuracy:56.00%,Kappa:0.1158,
F1(accepted):0.59,F1(rejected):0.52,
Recall(accepted):62%,Recall(rejected):50%,
Precision(accepted):57%,Precision(rejected):55%,
MediaMacro:0.56,MediaPonderada:056

Fonte: base de dados retirados do site kaggle

Com base nas métricas apresentadas, a Árvore de Decisão demonstrou um desempenho superior ao KNN em todos os critérios avaliados:

Acurácia: Árvore de Decisão (62,96%) vs. KNN (56,00%)
AUC: Árvore de Decisão (0,6099) vs. KNN (0,5673)
Recall: Árvore de Decisão (62,96%) vs. KNN (56,00%)
Precisão: Árvore de Decisão (63,05%) vs. KNN (55,90%)
F1-Score: Árvore de Decisão (62,65%) vs. KNN (55,86%)
Kappa: Árvore de Decisão (0,2541) vs. KNN (0,1158)
MCC: Árvore de Decisão (0,2570) vs. KNN (0,1161)

CONCLUSÃO

A análise realizada sobre a base de dados de admissão de alunos buscou avaliar a capacidade preditiva de dois modelos de classificação, Árvore de Decisão e KNN (K-Nearest Neighbors), para identificar quais candidatos possuem maior probabilidade de serem admitidos. Essa abordagem tem grande importância, pois permite otimizar o processo seletivo, tornando-o mais eficiente e justo, além de auxiliar instituições de ensino na tomada de decisões estratégicas.

Com base nas métricas apresentadas, foi possível observar que a Árvore de Decisão obteve um desempenho superior em todos os critérios avaliados, consolidando-se como o modelo mais adequado para o problema em questão.

Contribuições da equipe

JOÃO IGOR DO NASCIMENTO: Slide e Análise exploratória (50%)

YGOR REGIS DE SANTANA SILVA: Relatório e Aplicação dos modelos (50%)

REFERÊNCIAS

https://www.kaggle.com/datasets/zeeshier/student-admission-records/data

Pierson, L.; Data Science para leigos, 2ª ed. Rio de Janeiro: Alta Books, 2019. p. 92, 97, 101-103;

Morettin,P.A. e Singer J.M. ; Estatística e Ciência de dados, 1ª ed. , São Paulo, Blucher, 2022,p. 455- 476;

Slides sobre K-NN e Árvore de decisão(material da disciplina). Disponíveis em: https://jodavid.github.io/introducao_ds/ Último acesso: 20/03/2025 às 22:20.