UNIVERSIDADE FEDERAL DO CEARÁ

TECHNICAL REPORT

Aluno: Luis Sávio Gomes Rosa

1. Introdução

1.1. Sobre o dataset 1:

O dataset fornecido contém informações sobre a decisão de compra de um computador com base em várias características demográficas e financeiras dos indivíduos, seja ela idade, se é estudante, o preço do produto, entre outras.

1.2. Dataset e suas variáveis:

O dataset ubisoft possui 7 colunas, sendo elas:

- **RID:** Identificador único para cada registro.
- **Age:** Faixa etária do indivíduo, categorizada em três grupos: youth (jovem), middle_aged (meia-idade) e senior (sênior).
- **Income:** Nível de renda do indivíduo, classificado como high (alto), medium (médio) e low (baixo).
- **Student:** Indica se o indivíduo é estudante (yes ou no).
- **Credit_rating:** Classificação de crédito do indivíduo, que pode ser fair (justa) ou excellent (excelente).
- **Buys_computer:** Variável de resposta que indica se o indivíduo comprou um computador (yes ou no).

A problemática estabelecida foi utilizar os cálculos de Gini e Entropy para determinar as duas possibilidades de nó raíz da árvore de decisão.

1.3. Resultados e discussões sobre o dataset1:

	Gini	Entropy
Age	0.34	0.69
Student	0.36	0.78
Credit rating	0.42	0.89
income	0.44	0.91

Dado os valores de gini, podemos observar que as colunas com menores resultados forma as colunas Age e Student, logo, a entropia delas sendo as menores mostra que elas são as mais corretas para serem as duas possibilidades de nó raiz da árvore de decisão. No caso da entropy, como ocorrido nos cálculos de gini, as colunas Age e Student também tiveram os menores resultados. Diante disso, tanto a entropia quanto o gini mostram as colunas age e student como as melhores possibilidades de nós raiz.

2. Introdução ao dataset 2:

Neste conjunto de dados de "Plant Growth Data Classification", a tarefa envolve classificar o marco de crescimento das plantas com base nos fatores ambientais e de gerenciamento fornecidos com base em variáveis como tipo de solo, horas de luz solar, frequência de rega, tipo de fertilizante, temperatura e umidade. Esta ação pode ajudar a compreender como diferentes condições influenciam o crescimento das plantas e pode ser valiosa para otimizar as práticas agrícolas ou a gestão de estufas.

2.1. Pré Processamento:

A priori, a única ação de pré-processamento dos dados foi a transformação do dados categóricos em dados numéricos, visto que o dataset não possui não valor faltante sua distribuição da target é bem balanceada.

2.2. Resultados usando Árvore de decisão:

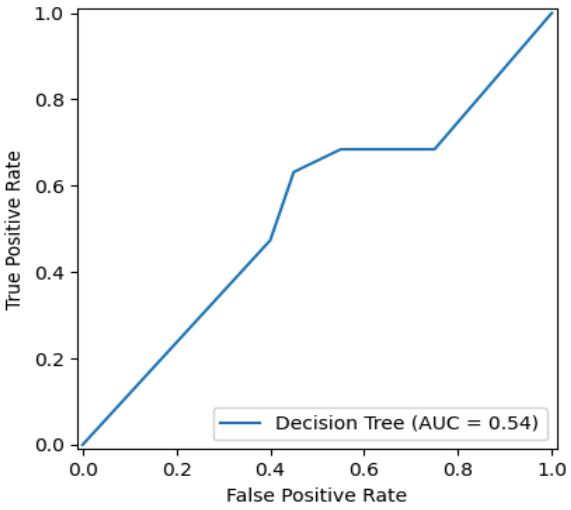
1. Tabela com a avaliação do modelo usando “classification_report”:

Acurácia: 0.59

	precision	**recall**	**f1-score**	**support**
0	0.67	0.55	0.58	20
1	0.57	0.63	0.60	19
accuracy			**0.59**	39
macro avg	0.59	0.59	0.59	39
weighted avg	0.59	0.59	0.59	39

Raiz do Erro quadrado médio: 0.641

2. Curva Roc:



O modelo de árvore de decisão apresentou uma acurácia bem insatisfatória, com um desempenho de 59%, onde o desempenho foi levemente melhor na identificação da classe 1 em comparação à classe 0. A RMSE de 0.64 complementa essa avaliação mostrando uma média das discrepâncias entre as previsões e os valores reais. A curva roc com outro avaliador, fomenta a que o modelo não possui uma boa capacidade de distinção de classes, quase chegando à 0.5 (que significa quase que o modelo está chutando).

3. Implementação manual do Random Classifier

Na terceira etapa, foi criado um algoritmo de Random Classifier, que é uma extensão do Random Forest. Diferente do Random Forest tradicional, que utiliza apenas variações de árvores de decisão, nesta atividade foi desenvolvida uma versão personalizada do algoritmo, capaz de trabalhar tanto com classificadores KNN (K-Nearest Neighbors) quanto com árvores de decisão.

3.1. Tabela com acurácias:

Decision Tree	0.64
knn	0.61

Como observado na tabela acima, o modelo de árvore de decisão acabou tendo um melhor desempenho, mesmo que no geral não seja um bom desempenho, que o modelo de vizinho mais próximo.

4. Implementação de múltiplos modelos:

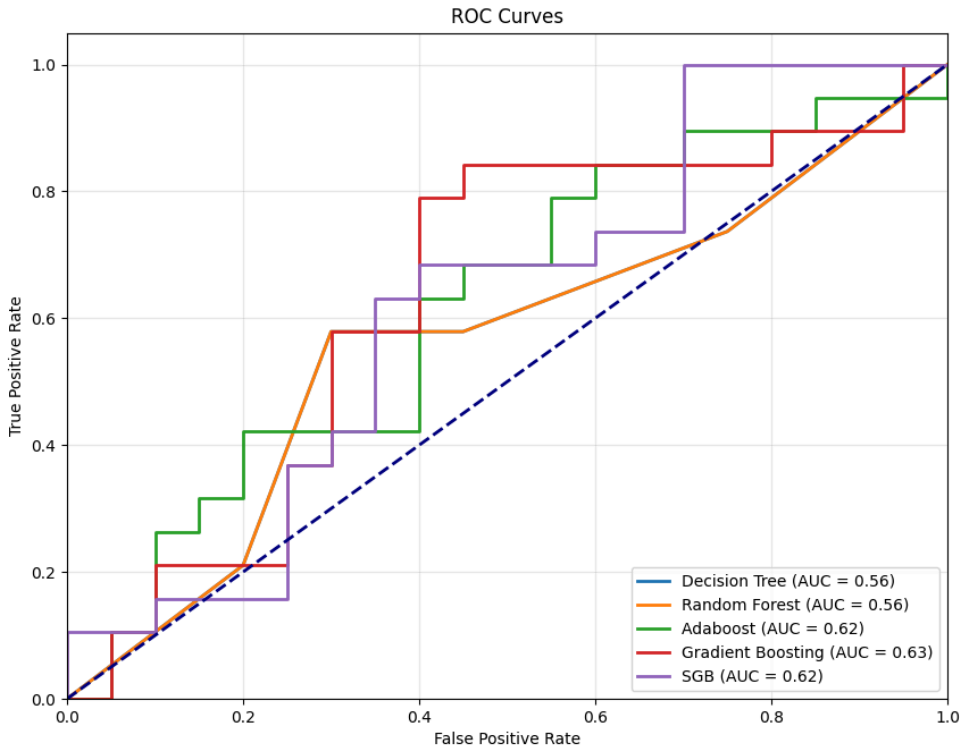
Nessa seção foram implementados 5 algoritmos de classificação nos dados. Foram eles: árvore de decisão, random forest, adaboost, gradientBoost e SGB.

4.1. Tabela com os resultados:

Algoritmo	acurácia
Decision Tree	0.64
Random Forest	0.64
Adaboost	0.59
Gradient Boosting	0.69
SGB	0.64

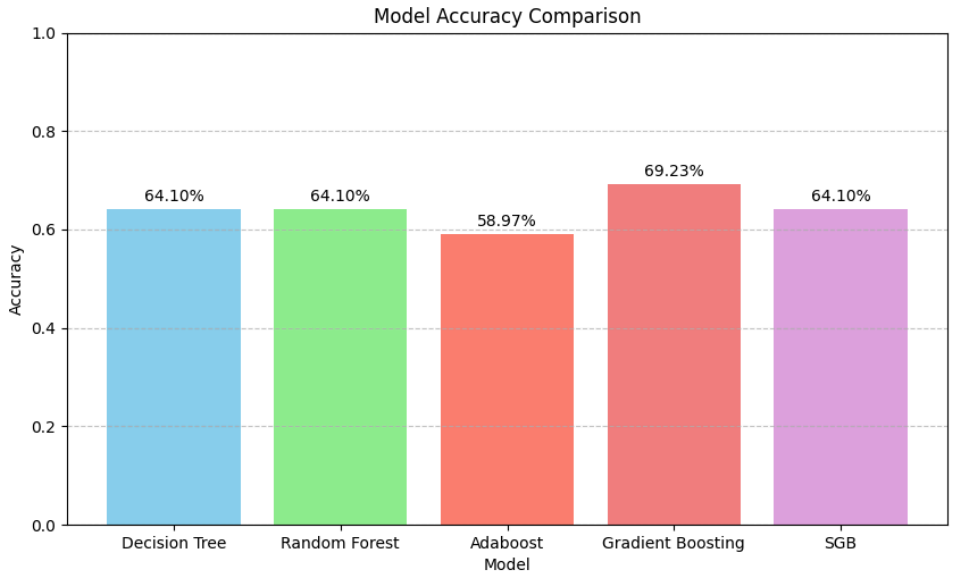
4.2. Curva roc com a acurácia dos modelos:

O gráfico a seguir mostra o desempenho dos modelos utilizando a curva de roc com o intuito de avaliar a capacidade dos modelos de distinção das classes que foram classificadas.



4.3. Gráfico de barras com a acurácia dos modelos:

A tabela a seguir nos mostra uma melhor visualização e comparação das acuracias dos modelos.



4.4. Conclusão:

Embora o Gradient Boosting apresentou a melhor acurácia, com 0.69, sendo o modelo mais eficaz para os dados analisados, ainda não foi atingida uma avaliação satisfatória. Diante disso, faz-se necessário uma melhor qualidade de dados. Algumas sugestões foram deixadas na parte de Próximos Passos.

5. Próximos Passos:

- **Uso de conjuntos de dados maiores:** A análise de conjuntos de dados maiores e mais diversos poderia fornecer insights mais generalizáveis.

- **Incorporação de Variáveis Adicionais:** Incluir variáveis como espécies de plantas, níveis de nutrientes no solo e métodos de controle de pragas poderia oferecer uma compreensão mais abrangente do crescimento das plantas.

6. Referências

GORTOROZYANNNN. **Plant Growth Data Classification**. Disponível em: <https://www.kaggle.com/datasets/gorororororo23/plant-growth-data-classification/data>. Acesso em: 5 ago. 2024.