CODERHOUSE

Modelo de Machine Learning para predição de carga cognitiva de uma pessoa

Data Science (Turma 63825)
Lucy Gomes de Souza

Objeto de pesquisa: Exame clínico neuropsicológico

Contexto:

Durante atividades de educação, os estudantes frequentemente enfrentam níveis variados de estresse e engajamento emocional, o que pode impactar diretamente seu desempenho e bem-estar. No entanto, as instituições de ensino carecem de ferramentas eficazes para monitorar e avaliar, em tempo real, o estado psicológico dos alunos, dificultando intervenções proativas e personalizadas.

Problema:

A falta de um sistema integrado que utilize dados fisiológicos, comportamentais e ambientais coletados por biossensores para avaliar o estado psicológico dos estudantes durante atividades educacionais resulta em:

- Dificuldade em identificar sinais precoces de estresse e desengajamento emocional.
- Falta de personalização no suporte oferecido aos estudantes.
- Perda de oportunidades para melhorar a eficácia das atividades educacionais e o bem-estar mental dos alunos.

Solução proposta:

Desenvolver uma plataforma baseada em machine learning que analise, em tempo real, dados coletados por biossensores para:

- Identificar o estado psicológico, níveis de estresse e engajamento emocional dos estudantes.
- Fornecer insights e alertas para professores e gestores educacionais, permitindo intervenções imediatas.
- Personalizar o suporte e as atividades educacionais com base no estado emocional de cada estudante.

Benefícios:

Melhoria do bem-estar mental e do desempenho acadêmico dos estudantes.

Aumento do engajamento emocional durante atividades educacionais.

Redução do estresse e prevenção de problemas de saúde mental.

Otimização do ambiente educacional e das metodologias de ensino.

Público-Alvo:

Instituições de ensino, empresas de tecnologia educacional e pesquisadores interessados em aplicar biossensores e machine learning para melhorar a experiência educacional e o suporte à saúde mental.

Objetivos:

Criar uma solução inovadora que utilize tecnologia de ponta para transformar dados em ações práticas, promovendo um ambiente educacional mais saudável, engajador e eficaz.

Pergunta:

A construção de um sistema integrado que utilize dados fisiológicos, comportamentais e ambientais coletados por biossensores possui suficiente correlação com a carga cognitiva para que possamos construir um algoritmo de machine learning capaz, de forma confiável, de inferir a carga cognitiva de uma pessoa?

Dataset (Data Acquisition)

Site de origem: https://www.kaggle.com/datasets/ziya07/psychological-state-identification-dataset

Sobre este arquivo

O conjunto de dados representa informações fisiológicas, comportamentais, ambientais e relacionadas a tarefas, coletadas por biossensores de indivíduos, principalmente estudantes, durante atividades educacionais. Foi criado para facilitar tarefas de aprendizado de máquina, como identificação em tempo real do estado psicológico, análise de estresse e avaliação do engajamento emocional.

Principais Destaques

Formato do Arquivo: CSV. Total de Registros: 1.000 linhas. Total de Recursos: 20 colunas, incluindo dados fisiológicos (ex.: HRV, GSR), condições ambientais, métricas de engajamento comportamental e detalhes demográficos.

TARGET

```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 20 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   ID                        1000 non-null   int64
 1   Time                      1000 non-null   object
 2   HRV (ms)                  1000 non-null   float64
 3   GSR (µS)                  1000 non-null   float64
 4   EEG Power Bands           1000 non-null   object
 5   Blood Pressure (mmHg)     1000 non-null   object
 6   Oxygen Saturation (%)     1000 non-null   float64
 7   Heart Rate (BPM)          1000 non-null   int64
 8   Ambient Noise (dB)        1000 non-null   float64
 9   Cognitive Load            1000 non-null   object
 10  Mood State                1000 non-null   object
 11  Psychological State       1000 non-null   object
 12  Respiration Rate (BPM)    1000 non-null   int64
 13  Skin Temp (°C)            1000 non-null   float64
 14  Focus Duration (s)        1000 non-null   int64
 15  Task Type                 1000 non-null   object
 16  Age                       1000 non-null   int64
 17  Gender                    1000 non-null   object
 18  Educational Level         1000 non-null   object
 19  Study Major               1000 non-null   object
dtypes: float64(5), int64(5), object(10)
```

Preparação dos Dados (Data Wrangling)

Bandas de Potência EEG: Captura a atividade cerebral nas bandas Delta, Alfa e Beta.

EEG_Band_1	EEG_Band_2	EEG_Band_3
0.7583653347946298	1.423247998317594	0.6157696670741735

Pressão Arterial (Sistólica/Diastólica) (mmHg): Mede a resposta cardiovascular.

Blood Pressure (systolic)	Blood Pressure (diastolic)
114	79

Ajustando as colunas com tipo 'object'

pd.get_dummies(df, columns=cols_to_encode)

'Mood State', 'Psychological State', 'Task Type', 'Gender', 'Study Major'

drop_first=False (valor padrão) foi mantido, pois:

- É necessário todas as categorias e os algoritmos não tem problemas com multicolinearidade;
- Importante para uma melhor interpretabilidade dos resultados.

	Task Type_Assignment	Task Type_Exam	Task Type_Group Discussion	Task Type_Lecture	Gender_Female	Gender_Male	Gender_Other	Study Major_Arts	Study Major_Engineering	Study Major_Science
	False	True	False	False	True	False	False	False	True	False

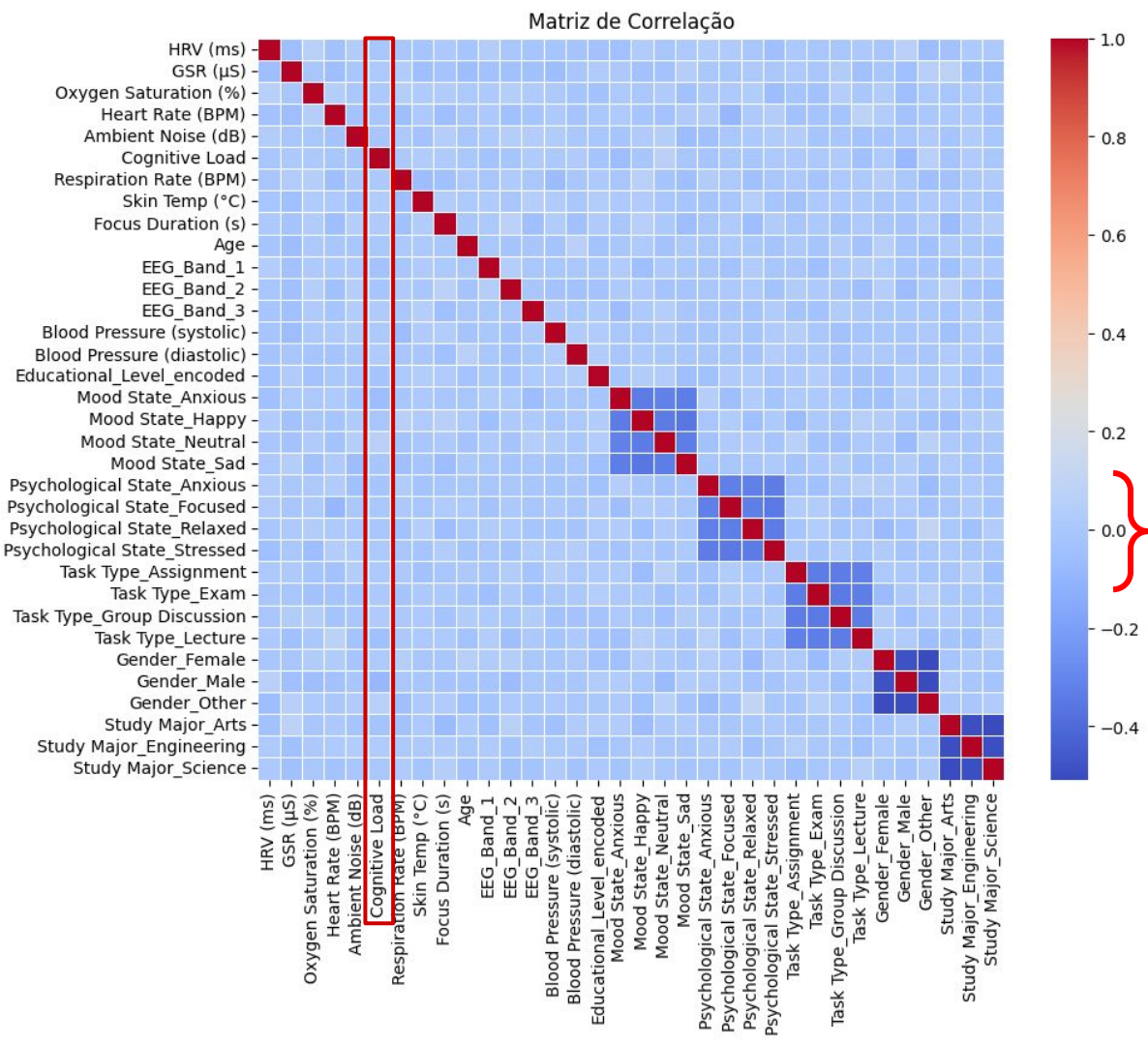
CONVERSÃO DE BOOLEANO PARA NUMÉRICO

	Task Type_Assignment	Task Type_Exam	Task Type_Group Discussion	Task Type_Lecture	Gender_Female	Gender_Male	Gender_Other	Study Major_Arts	Study Major_Engineering	Study Major_Science
	0	1	0	0	1	0	0	0	1	0

Análises exploratórias do dataset

Relação entre as features e o target:

O heatmap mostra que todas as features do dataset têm baixa correlação linear com o target, sendo a maior positiva *Mood State_Neutral* (0,0638) e a maior negativa *Gender_Male* (-0,0816). Vale ressaltar que a correlação só capta relações lineares; se houver padrões não lineares, o modelo ainda pode extrair informações relevantes (por isso nenhuma feature será removida). Contudo, a ausência de correlação significativa pode sugerir que as variáveis preditoras não influenciam substancialmente a variável resposta.



Divisão dos dados

O dataset será separado em conjuntos de treino, validação e teste (respectivamente: 70% treino, 20% validação, 10% teste).

Escolha dos Algoritmos

1.Árvores de Decisão (Decision Trees): São simples, interpretáveis e lidam bem com dados numéricos e categóricos. Podem capturar relações não lineares entre as features e o target.

2.Florestas Aleatórias (Random Forests): Melhoram as árvores de decisão ao combinar várias árvores, reduzindo o risco de overfitting. São robustas à features irrelevantes.

3.Gradient Boosting (escolhido o XGBoost): São modelos poderosos que combinam várias árvores de decisão de forma sequencial, corrigindo os erros das árvores anteriores. Tendem a ter alta precisão.

4.Support Vector Machines (SVM): São eficazes em problemas de classificação com um número moderado de features. Podem lidar com relações não lineares usando kernels.

5.K-Nearest Neighbors (KNN): É um algoritmo simples baseado em distância que pode ser eficaz quando há padrões claros nos dados.

6.Regressão Logística (Logistic Regression): É um modelo linear simples e interpretável, útil para problemas de classificação multiclasse.

Escolha das métricas

$$\text{Acurácia} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Porcentagem de previsões corretas (total de acertos).

$$\text{Precisão} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Proporção de verdadeiros positivos entre todos os previstos como positivos (evitar falsos positivos).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Proporção de verdadeiros positivos identificados corretamente (evitar falsos negativos).

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}$$

Média harmônica entre Precisão e Recall (balanceia as duas métricas).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Magnitude média dos erros de previsão (quanto menor, melhor o modelo).

Análise (Treinamento e Avaliação do Modelo)

	Accuracy
KNN	0.4000
Random Forest	0.3500
Decision Tree	0.3350
Logistic Regression	0.3200
SVM	0.3100
XGBoost	0.3000

Melhor Modelo: KNN com 0.4000 de acurácia!

Resultados Finais com Métricas para Todos os Modelos:

	CV Accuracy	Validation Accuracy	Precision	Recall	F1-score	RMSE
KNN	0.3614	0.4000	0.3761	0.4000	0.3716	1.1023
Random Forest	0.3443	0.3500	0.3469	0.3500	0.3305	1.0977
Decision Tree	0.3486	0.3350	0.3348	0.3350	0.3329	1.1511
Logistic Regression	0.3243	0.3200	0.3011	0.3200	0.3036	1.1576
SVM	0.3557	0.3100	0.2003	0.3100	0.2340	1.0607
XGBoost	0.3300	0.3000	0.2950	0.3000	0.2954	1.2166

Melhor Modelo: KNN com 0.4000 de acurácia na validação!

Ajuste de hiperparâmetros

Técnicas de Otimização

Random Search: Testa amostras aleatórias dos hiperparâmetros. Mais rápido, mas pode não encontrar a melhor combinação.

Média e desvio padrão da validação Cruzada (Cross-Validation): Divide os dados em K partes (K-Folds) e treina o modelo várias vezes. Reduz o risco de overfitting.

Grid Search: Testa todas as combinações possíveis de hiperparâmetros. Mais preciso, mas demorado para grandes combinações.

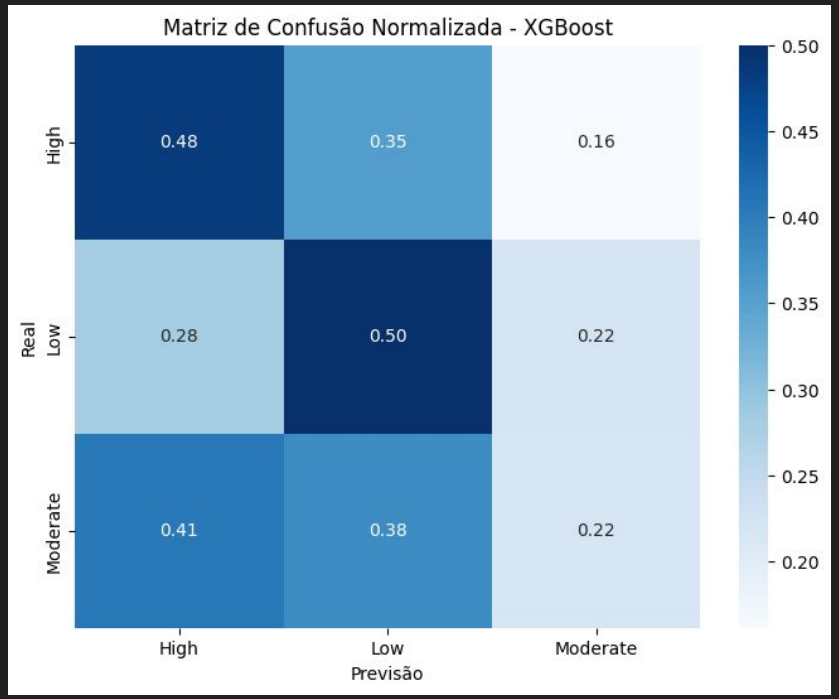
Resultados Finais:

	Accuracy
KNN	0.3700
Decision Tree	0.3400
Random Forest	0.3400
Logistic Regression	0.3200
XGBoost	0.3150

Melhor Modelo: KNN com 0.3700 de acurácia!

No geral, RandomizedSearchCV ajudou a encontrar boas combinações iniciais, enquanto GridSearchCV refinou os ajustes, beneficiando especialmente KNN e Random Forest. Já XGBoost teve desempenho inferior, possivelmente devido à necessidade de um pré-processamento mais cuidadoso. No entanto, os resultados do modelo KNN sem ajustes segue sendo o melhor encontrado.

Teste:



Modelo: Decision Tree
Acurácia: 0.2600
Precisão: 0.2788
Recall: 0.2600
F1-Score: 0.2626

Modelo: Random Forest
Acurácia: 0.3100
Precisão: 0.3252
Recall: 0.3100
F1-Score: 0.2851

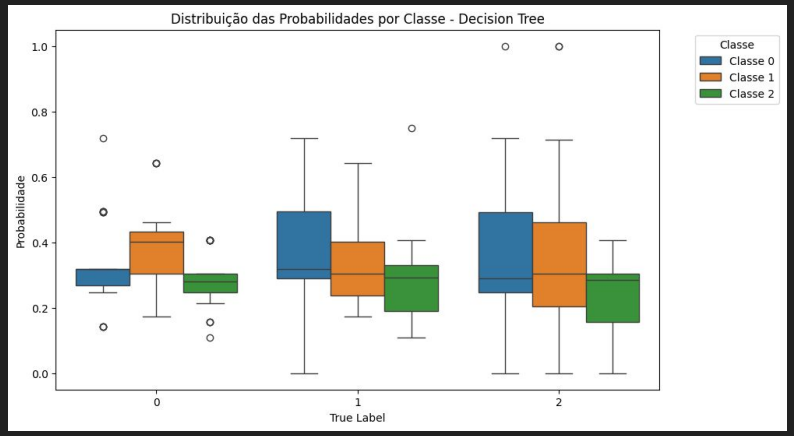
Modelo: XGBoost
Acurácia: 0.3900
Precisão: 0.3921
Recall: 0.3900
F1-Score: 0.3770

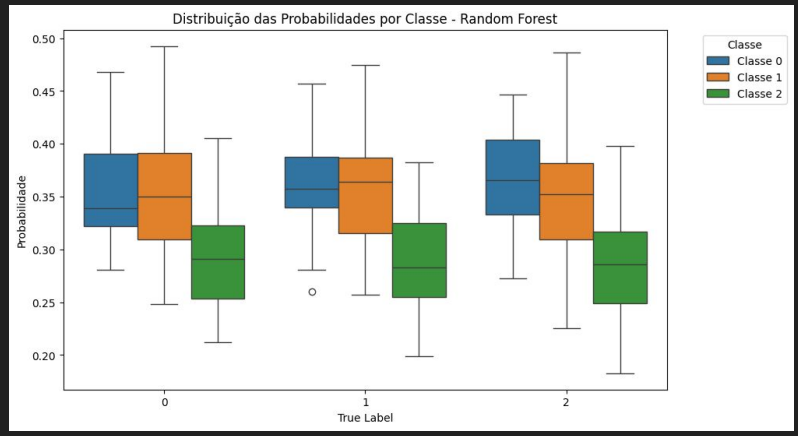
Modelo: KNN
Acurácia: 0.3700
Precisão: 0.3289
Recall: 0.3700
F1-Score: 0.3285

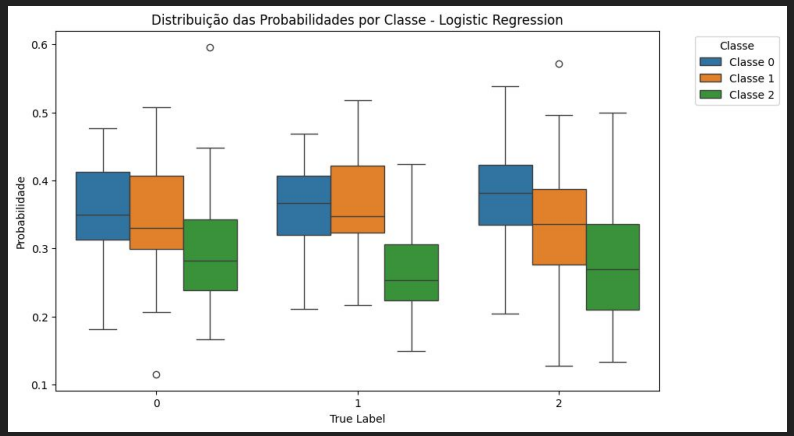
Modelo: Logistic Regression
Acurácia: 0.3400
Precisão: 0.3740
Recall: 0.3400
F1-Score: 0.3243

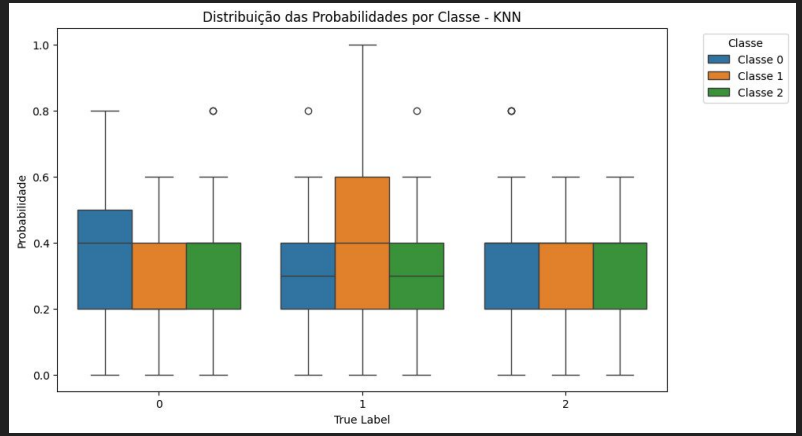
todos os modelos apresentaram dificuldades na classificação, especialmente para a classe 2, o que sugere desafios na separação das classes.

Relatório de Classificação Detalhado



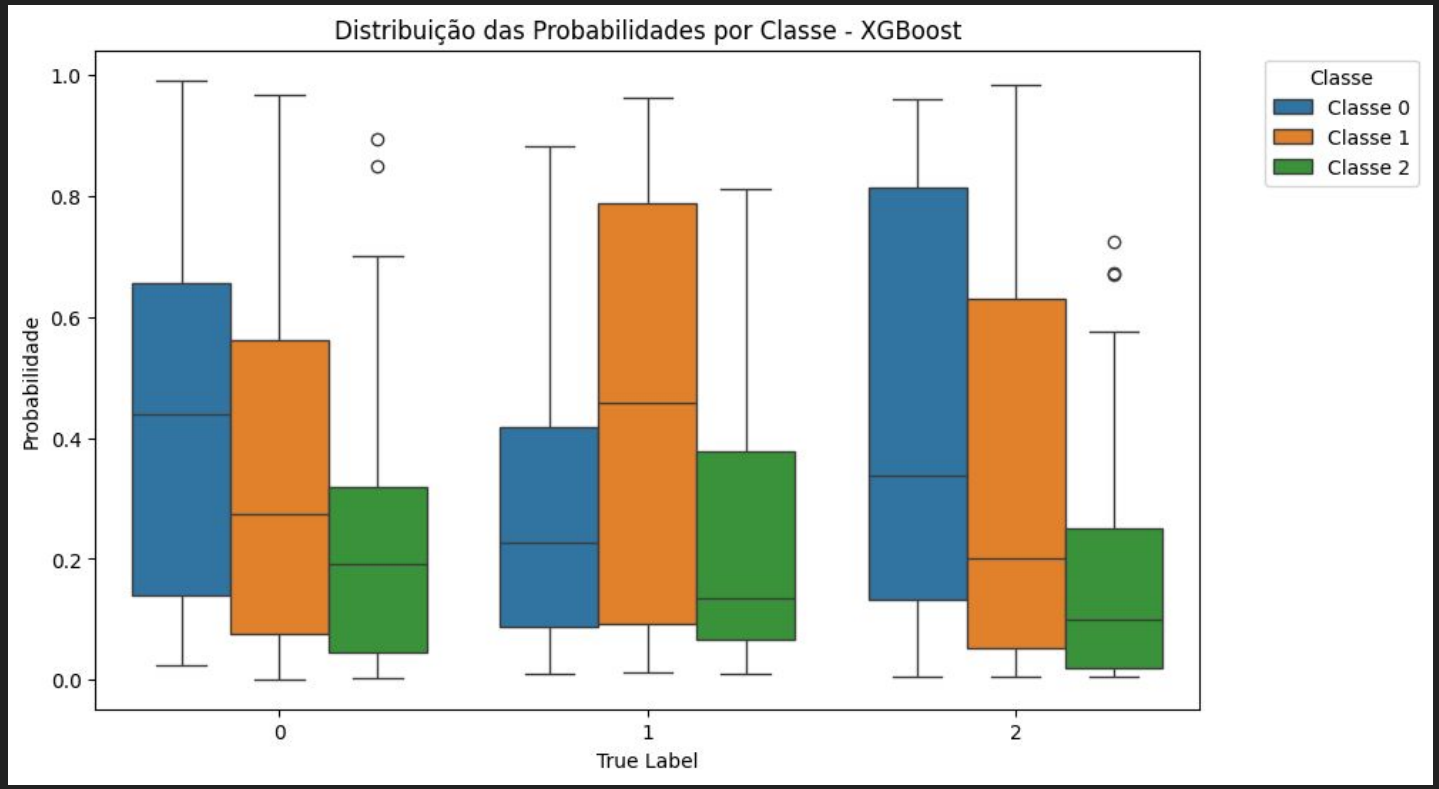






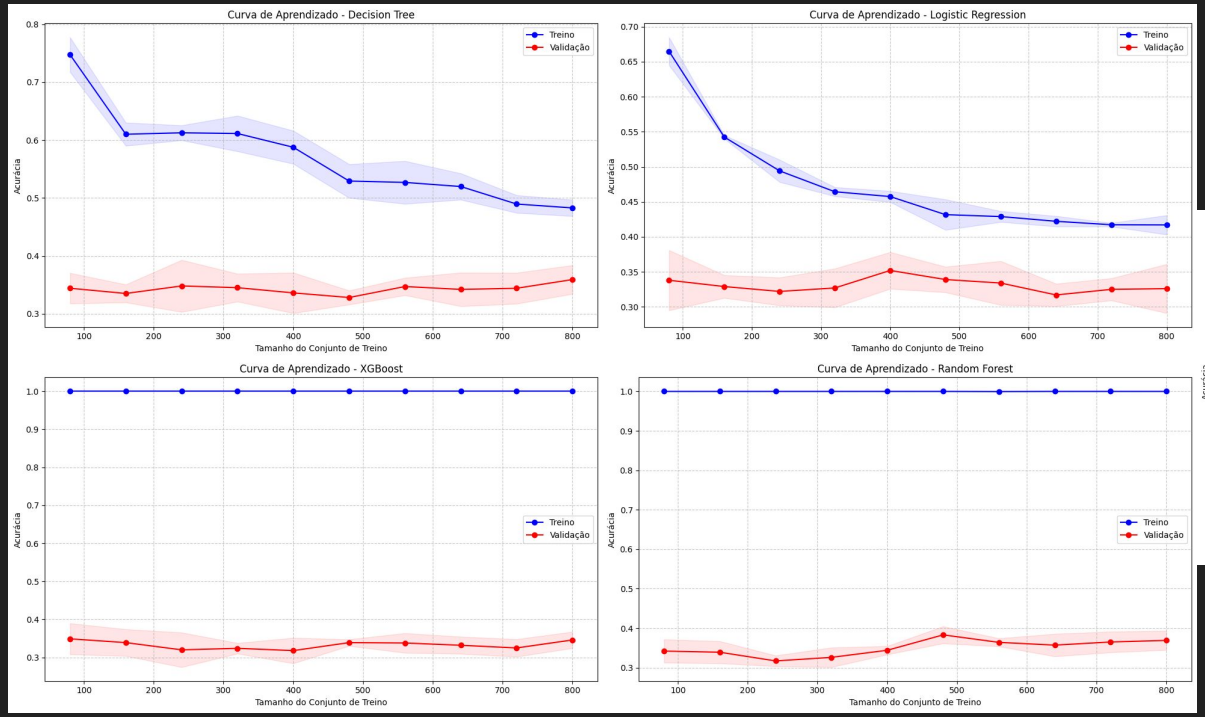
observação: 'High' -> 0, 'Low' -> 1, 'Moderate' -> 2

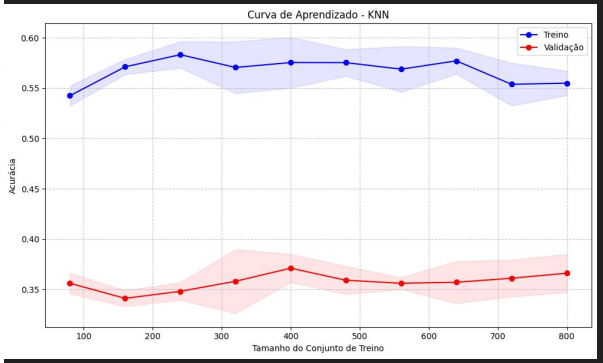
Relatório de Classificação Detalhado



observação: 'High' -> 0, 'Low' -> 1, 'Moderate' -> 2

Learning Curve (Curva de Aprendizado)





Todos os modelos apresentam curvas de validação lineares e de baixo desempenho, indicando falta de aprendizado ou generalização, possivelmente devido a underfitting, dados insuficientes, features pouco informativas ou hiperparâmetros inadequados. Decision Tree e Logistic Regression mostram underfitting, com acurácia decrescente em treino e validação, sugerindo modelos excessivamente simples. Random Forest e XGBoost exibem overfitting extremo, com 100% de acurácia no treino, indicando memorização dos dados. Já o KNN tem desempenho próximo ao acaso (~0,55 de acurácia), caracterizando underfitting por incapacidade de capturar padrões relevantes.

Pergunta:

A construção de um sistema integrado que utilize dados fisiológicos, comportamentais e ambientais coletados por biossensores possui suficiente correlação com a carga cognitiva para que possamos construir um algoritmo de machine learning capaz, de forma confiável, a carga cognitiva de uma pessoa?

Resposta:

Com o dataset utilizado não foi possível a construção de um modelo eficiente, prático e aplicável ao mundo real. Portanto, respaldada pelo que foi aqui apresentado concluímos que não foi possível construir um modelo confiável capaz de predizer a carga cognitiva de uma pessoa.

Perspectivas futuras

- O dataset precisa ser aprimorado com novas features que tenham maior correlação com o target, especialmente as relacionadas a estados emocionais e características fisiológicas.
- Redes neurais recorrentes (RNN) são promissoras para análise desses dados, mas exigem aumento significativo de amostras e feature engineering.
- A abordagem EOO-DLSTM combina LSTM com otimização para emoções, sendo eficaz para dados sequenciais e análise psicológica.
- Essa tecnologia pode revolucionar áreas como educação e saúde mental, melhorando precisão e aplicações.
- É essencial garantir práticas éticas, privacidade e interpretabilidade dos resultados para seu uso efetivo.

EOO-DLSTM (Emotion-Oriented Optimization - Deep Long Short-Term Memory)