ANTÔNIO MENEGHETTI FACULDADE

BACHARELADO EM SISTEMAS DE INFORMAÇÃO

INTELIGÊNCIA ARTIFICIAL II: TRABALHO 1

ANDERSON STEUERNAGEL MENEGASSI

RECANTO MAESTRO

2025

Problema e objetivo:

Contexto e Motivação:

A depressão é um problema de saúde mental crescente entre estudantes de universidades, afetando desempenho acadêmico, relações sociais e bem-estar da saúde do estudante. Pretende-se identificar pessoas com alto risco de depressão e conseguirmos realizar intervenções preventivas e direcionadas para cada universitário.

A depressão é a doença do século, e segundo o site Publico, 1 em cada 4 universitários tomam medicação psiquiátrica. Link: https://www.publico.pt/2024/12/13/p3/noticia/quatro-universitarios-toma-medicacao-psiquiatrica-aponta-estudo-2115587

Objetivo

O objetivo desse trabalho é desenvolver e avaliar modelos de **aprendizado de máquina,** capazes de prever a depressão em estudantes universitários, considerando variáveis de: estudos, sociais e comportamentais. Os algoritmos serão treinados, e poderão ser utilizados para testes. Esse trabalho não tem o objetivo de fazer acusações, e apenas será usado para o trabalho da disciplina.

Hipótese

Espera-se ter um resultado satisfatório dos algoritmos selecionados, com uma acurácia acima de 80%, a fim de podermos prever, de acordo com as variáveis disponíveis, e mais as variáveis criadas, a possibilidade de depressão de cada estudante. Não será utilizado para alguma atividade profissional, porque necessitaria de mais dados para treinamento.

2. Dados

Origem e Licença

O dataset público foi baixado do Kaggle em CSV.
Link do dataset: https://www.kaggle.com/datasets/adilshamim8/student-depression-dataset

Principais Variáveis

- Depressão : variável alvo (0 = não, 1 = sim)
- Pensou em se matar: pensamentos de se suicidar
- Pressão: pressão acadêmica normalizada
- Estresse financeiro: estresse financeiro normalizado
- Histórico familiar: histórico familiar de doenças mentais
- Horas de estudo: carga horária semanal normalizada (0 = não, 1 = sim)
- Nota na faculdade: satisfação e desempenho acadêmico normalizados
- Tempo dormindo, hábitos alimentares, gênero, cidade, Graduação, idade, satisfação nos estudos

Limpeza e Engenharia de Atributos

- Remover as colunas que possuem muitos dados iguais
    - Profissão (quase todos eram estudantes).
    - Pressão no trabalho (poucos trabalhavam).
    - Satisfação no trabalho (poucos trabalhavam.

- Conversão de texto para numérico:

  - Gênero: 1- masculino e 0-feminino.
  - Já teve pensamentos de se suicidar: 1-sim e 0-não.
  - Se a família já tem histórico de doenças mentais: 1-sim e 0-não

- Normalização dos dados para ficar entre 0 à 1:
  - Idade, pressão acadêmica, Nota média na faculdade, satisfação nos estudos, horas de estudo e estresse financeiro.

- Substituição de cidades com poucos registros para “outras_cidades”.

- Adição de uma nova coluna: Área de estudo, de acordo com cada graduação de cada aluno, classificado em : Saúde, Engenharia/Tecnologia, Administração e Negócios e Outros.

- Utilização do **get_dummies** para a transformação das variáveis categóricas para variáveis numéricas.

- Preenchimento de dados na variável de Estresse financeiro, com os valores médios.

Vazamento de dados:

Foi retirado a variável de “depressão”, para evitar que os modelos tenham a resposta no treinamento, o que aumentaria bastante a acurácia dos modelos, tendo um overfitting.

Divisão Treino/Teste

- O dataset foi dividido em 80% dos dados para teste, e 20% para treinos, sendo esses dados escolhidos aleatoriamente durante o processo de separação dos dados.

3. Metodologia

Pipeline do Projeto

1. Carregamento e seleção de colunas

2. Tratamento de valores ausentes

3. Normalização
4. Conversão de campos categóricos em numéricos
5. Divisão treino/teste
6. Treino de múltiplos modelos com **RandomizedSearchCV,** para combinações aleatórias de hiperparâmetros.
7. Avaliação de métricas de classificação
8. Seleção do melhor modelo por acurácia

Algoritmos Testados

- Random Forest
- Bagging
- AdaBoost
- Gradient Boosting
- XGBoost
- LightGBM

Hiperparâmetros

- Hiperparâmetros específicos definidos para cada modelo (n_estimators, max_depth, learning_rate, num_leaves etc.)
- Busca aleatória do RandomizedSearchCV com 10 combinações por modelo

Avaliação

- **Acurácia** (Valor de acurácia dos modelos)

4. Experimentos e Resultados

Treinamento e Busca de Hiperparâmetros

Foi utilizado alguns valores distintos aleatórios, para testar o mesmo modelo várias vezes, com hiperparâmetros diferentes, para cada modelo.

Resultados de Teste

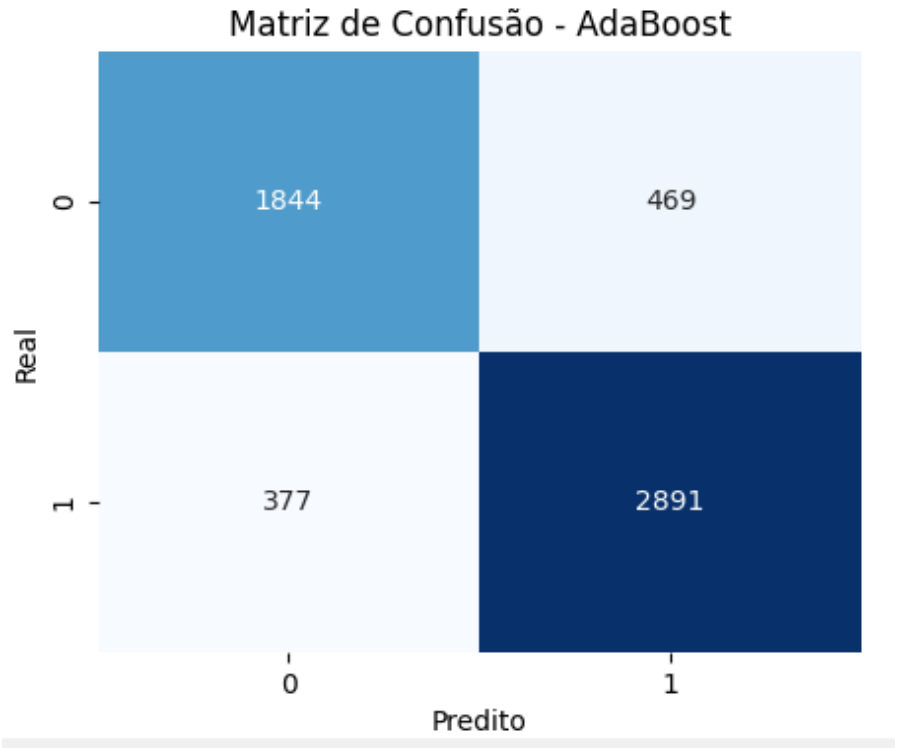
A seguir, temos uma tabela que indica a acurácia que cada modelo teve durante o processo de treinamento.

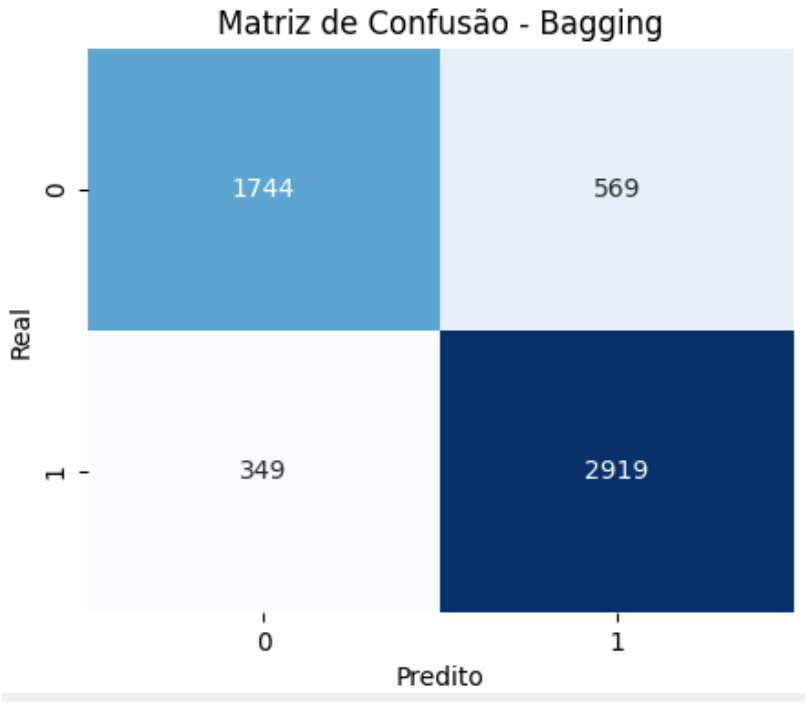
Modelo	Acurácia
Random Forest	0.8404
Bagging	0.8355
AdaBoost	0.8484
Gradient Boosting	0.8464
XGBoost	0.8466
LightGBM	0.8472

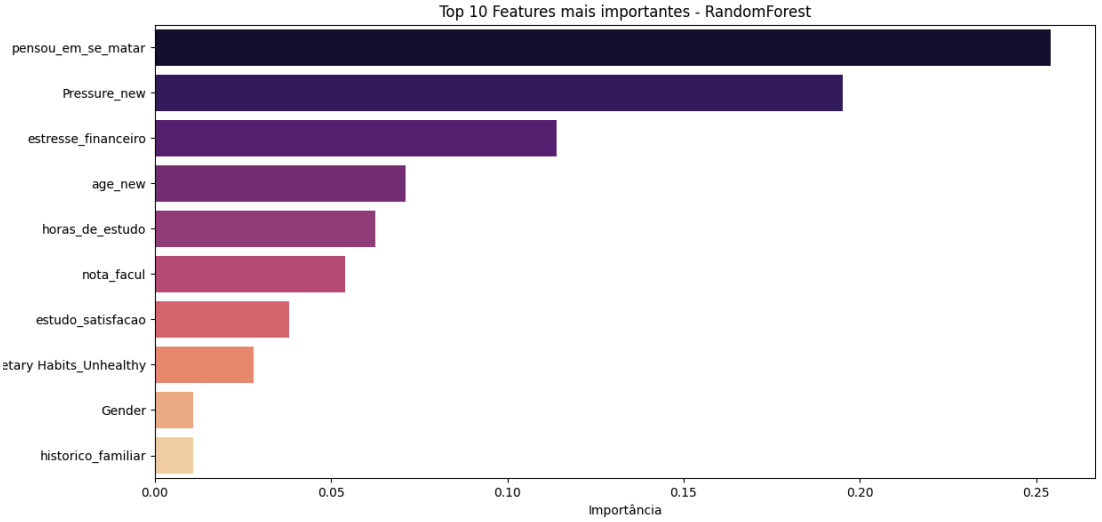
Análise Crítica

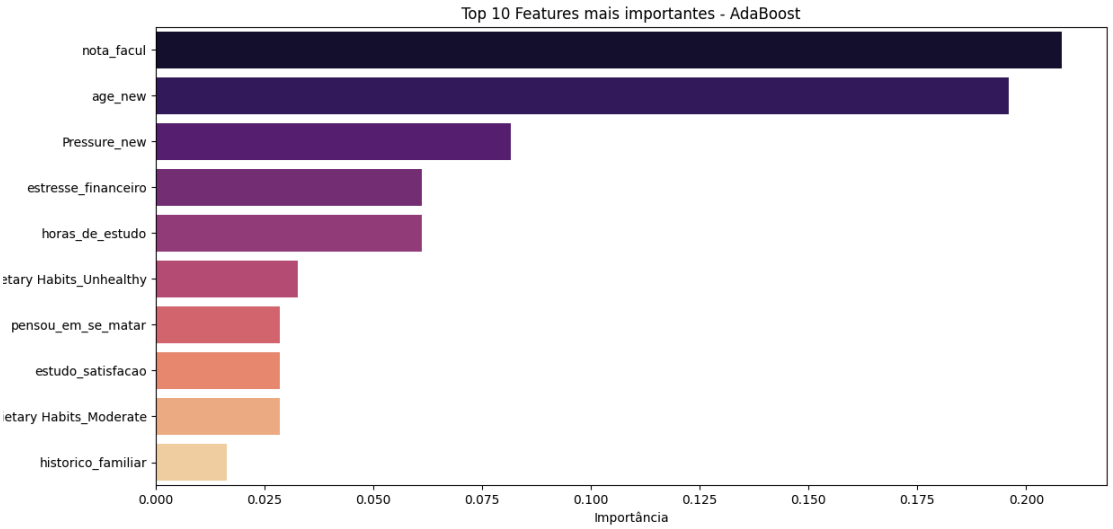
- Modelos AdaBoost, Gradient Boosting, XGBoost e LightGBM apresentam melhores resultados.
- Random Forest e Bagging tiveram os desempenhos mais baixos.
- Identificamos que os modelos tiveram uma taxa de acerto com valores bastante próximos, mas mesmo assim, tivemos diferenças notáveis nos algoritmos.

Imagens:




Matriz de Confusão - Bagging
1744
569
349
2919
0
1
Real
0
1
Predito


Top 10 Features mais importantes - RandomForest
pensou_em_se_matar
Pressure_new
estresse_financeiro
age_new
horas_de_estudo
nota_facul
estudo_satisfacao
etary Habits_Unhealthy
Gender
historico_familiar
0.00
0.05
0.10
0.15
0.20
0.25
Importância


Top 10 Features mais importantes - AdaBoost
nota_facul
age_new
Pressure_new
estresse_financeiro
horas_de_estudo
etary Habits_Unhealthy
pensou_em_se_matar
estudo_satisfacao
ietary Habits_Moderate
historico_familiar
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Importância

Com a matriz de confusão, conseguimos perceber que o pior modelo Baggin, obteve uma predição menor de valores em que o estudante tinha depressão, e o modelo apontou que não tinha, em comparação com o AdaBoost.

Mas em comparação aos valores em que o modelo fez a predição dizendo que a pessoa tinha depressão, e na verdade não tinha, o modelo Baggin possui 100 registros a mais em comparação ao modelo Adaboost.

Percebemos uma diferença entre as variáveis que o AdaBoost previu como as mais importantes para sua análise, em comparação com o modelo DecisionTree. A variável de alimentação aparece duas vezes no modelo AdaBoost, uma como hábito não saudável, e outra, como hábito moderado. Isso se dá pelo fato de que foi utilizado o get_dummies, que separou todas as variáveis categóricas, e numéricas.

5. Interpretação

- Principais insights:
  - Maior pressão acadêmica e estresse financeiro aumentam risco na decisionTree.

  - Histórico familiar influencia fortemente a probabilidade de depressão na decisionTree

  - Histórico familiar influencia fortemente a probabilidade de depressão na decisionTree

  - Pressão acadêmica, idade e nota acadêmica apresentam efeito são as principais variáveis no modelo adaboost

6. Refinamentos

- Ajustes de hiperparâmetros melhoram a acurácia, como a alteração dos valores que são utilizados para o treinamento de cada modelo..
- Utilização de variáveis criadas aumentaram a acurácia, como por exemplo: Área do curso de estudo
- Mudanças na ordem das variáveis, também fizeram com que os valores da acurácia mudassem nos modelos.
- Retirada de algumas variáveis: idade, gênero e cidade diminuíram um pouco o valor da acurácia, por isso elas foram deixadas nos treinamentos.
- A junção das cidades que tinham poucos registros, para o valor de “Outras Cidades”, fizeram a acurácia aumentar também.

7. Conclusões e Próximos Passos

- AdaBoost e Light GBM foram os modelos mais eficazes. Poderia ser feito o teste com mais algoritmos.

- Limitações:Algumas variáveis com poucos dados ou muitos dados iguais, como por exemplo: pressão no trabalho e satisfação no trabalho, continham bastante dados que não agregariam no resultado, pois a maioria dos universitários eram apenas estudantes, não trabalhavam.

- Futuro: Deveria ser incluído variáveis psicológicas, buscar estudantes que trabalham, para poder incluir mais alguns campos relevantes no treinamento. Seria necessário fazer uma pesquisa em diversos lugares, para podermos ter um modelo que possa ser utilizado em qualquer região do mundo.

8. Ética e Limitações

- **Viés de dados**: população restrita a estudantes universitários de determinadas cidades apenas, e não um estudo de todo o mundo.

- **Generalização limitada** a outras populações. Necessitaria de universitários de todo o mundo, com a adição de mais campos.

- **Risco de uso indevido**: diagnósticos automatizados sem acompanhamento profissional. Necessitaria de muitos dados para podermos ter uma melhor acurácia, com mais variáveis para utilizar no treinamento. Possíveis falsos apontamentos de depressão, pois o melhor modelo apresentou apenas 84% de acurácia. Precisaria ser feito um estudo muito mais aprofundado.