Sobre o Conjunto de Dados

Um conjunto de dados sobre depressão estudantil normalmente contém informações voltadas para analisar, entender e prever os níveis de depressão entre estudantes. Pode incluir características como informações demográficas (idade, gênero), desempenho acadêmico (notas, presença), hábitos de estilo de vida (padrões de sono, exercícios, atividades sociais), histórico de saúde mental e respostas a escalas padronizadas de depressão.

Esses conjuntos de dados são valiosos para pesquisas em psicologia, ciência de dados e educação, a fim de identificar fatores que contribuem para a depressão entre estudantes e criar estratégias de intervenção precoce. Considerações éticas, como privacidade, consentimento informado e anonimização dos dados, são cruciais ao trabalhar com essas informações sensíveis.

Estrutura do Arquivo

- **Formato:** CSV
- **Linhas:** Cada linha representa um estudante individual.
- **Colunas:** Cada coluna representa uma característica ou atributo específico.

Colunas

- **ID**: Identificador único, como um número de matrícula ou código atribuído a cada estudante.
- **Gênero**: Masculino, Feminino ou outras opções, dependendo das preferências do estudo.
- **Idade**: Número de anos do estudante.
- **Cidade**: Nome da cidade ou região geográfica do estudante.
- **Profissão**: Cargo ou área de atuação profissional do estudante, como "Estudante", "Engenheiro", etc.
- **Pressão Acadêmica**: Escala de 1 a 10, por exemplo, indicando o nível de estresse relacionado aos estudos.
- **Pressão no Trabalho**: Escala de 1 a 10, indicando a pressão no ambiente de trabalho, se aplicável.
- **CGPA**: Média ponderada das notas (Cumulative Grade Point Average), variando entre 0 a 10 ou 0 a 4, dependendo do sistema.
- **Satisfação com os Estudos**: Escala de 1 a 10, indicando o nível de satisfação com o curso ou a área de estudo.
- **Satisfação no Trabalho**: Escala de 1 a 10, indicando o nível de satisfação com o trabalho ou estágio, se aplicável.
- **Duração do Sono**: Média de horas de sono por noite (ex: 7 horas).

- **Hábitos Alimentares**: Tipo de dieta ou hábitos, como "Saudável", "Desbalanceada", "Vegetariano", etc.
- **Grau**: Conclusão de um curso superior (Bacharelado, Mestrado, Doutorado, etc.), ou se o estudante está em curso.
- **Você já teve pensamentos suicidas**: Resposta binária (Sim/Não) ou uma escala de intensidade.
- **Horas de Trabalho/Estudo**: Total de horas dedicadas ao trabalho e/ou estudos por dia ou semana.
- **Estresse Financeiro**: Nível de preocupação com finanças, em uma escala de 1 a 10, por exemplo.
- **Histórico Familiar de Doença Mental**: Se o estudante tem histórico familiar de problemas de saúde mental, é indicado por sim/não.

Link para o dataset

Análise do Dataset

Primeiramente, foi realizada uma verificação para identificar valores nulos. Foram encontrados 3 valores nulos na coluna **Estresse Financeiro**, uma das principais do modelo. Como alternativa, optou-se por excluir as linhas com valores ausentes.

Sleep Duration e Dietary Habits: pra ter uma ideia de quantas variáveis dummy seriam criadas, se deixaria o processo muito pesado.

Work Pressure: ao analisar apenas com o olho me deparei com uma inconsistência, com o count descobri que 99% das coluna eram zeros, optei por não utilizar a coluna.

Depression: havia uma diferença de 5.000 a menos dos alunos que não tinham depressão. Em vez de utilizar o SMOTE para criar variáveis sintéticas, optei por selecionar 13063 amostras (esse número foi alterado várias vezes até eu chegar nos valores que eu queria) dos que tinham depressão para balancear os dados.

Colunas Selecionadas

X (Variáveis Independentes)

- Gênero
- Idade
- Pressão Acadêmica
- CGPA
- Satisfação com os Estudos
- Duração do Sono
- Hábitos Alimentares
- Você já teve pensamentos suicidas
- Horas de trabalho/estudo
- Estresse financeiro

- Histórico familiar de doença mental

y (Variável Dependente - Rótulo)

- Depressão

Separação e Transformação dos Dados

- **Divisão dos Dados:** 80% treino e 20% teste, sem aleatorização na divisão de dados.
- **Divisão Validação:** 72% treino e 8% validação, sem aleatorização na divisão de dados.
- **LabelEncoder:** para converter as colunas que continham duas categorias de strings para 0 e 1.
- **MinMaxScaler:** Normalização das colunas no qual eu não utilizei LabelEncoder.
- **OneHotEncoder:** Para converter colunas com mais de duas categorias de string, para criar variáveis dummy .

Rede Neural para Classificação

- **EarlyStopping:** Interrompe o treinamento quando a perda da validação não melhora por 15 épocas.
- **Estrutura da Rede:** 2 camadas ocultas com 212 neurônios cada.
- **Dropout:** 50% para evitar o overfitting.
- **Função de Ativação:**
    - **ReLU:** Usado nas camadas ocultas para evitar o desaparecimento do gradiente.
    - **Sigmoid:** Utilizado na saída para classificação binária.
- **Otimizador:** Adam, por ser seguro e eficiente na grande maioria dos casos.
- **Função de Perda:** Binary Crossentropy, apropriada para classificação binária.
- **Métrica de Avaliação:** Accuracy (Acurácia).
- **Hiperparâmetros:**
    - **Épocas:** 100 (com EarlyStopping para interromper antes, se necessário).
    - **Batch Size:** 32 (atualiza os pesos a cada 616 amostras processadas).

Resultados

Para avaliar o desempenho do modelo, foram gerados três gráficos:

1. **Acurácia** do treinamento e da validação.
2. **Perda** do treinamento e da validação.
3. **Matriz de confusão**, para visualizar a classificação correta e incorreta dos dados.

Métricas Finais

	precision	recall	f1-score	support
0	0.84	0.83	0.83	2319
1	0.85	0.86	0.85	2607
accuracy			0.84	4926
macro avg	0.84	0.84	0.84	4926
weighted avg	0.84	0.84	0.84	4926