MODELO PARA A ENTREGA DAS ATIVIDADES

COMPONENTE CURRICULAR:	Projeto Aplicado II
NOME COMPLETO DO ALUNO:	Willians Carvalho da Silva -10416087 João Pedro Santos Oliveira – 10423752 Cesar Valentim Silva – 10416087 Luciano Guimaraes Costas – 10289655

UNIVERSIDADE PRESBIRIANA MACKENZIE

CURSO DE TECNOLOGIA EM CIÊNCIA DE DADOS

PROJETO APLICADO II

Relatório Técnico – Correção de Redações Automatizada com Aprendizado de Máquina

1. Introdução

A avaliação de redações é uma tarefa complexa e subjetiva, com desafios relacionados à consistência e ao tempo necessário para a análise manual. O presente projeto tem como objetivo automatizar esse processo por meio de técnicas de aprendizado de máquina, utilizando os algoritmos Random Forest e XGBoost.

Essa solução foi projetada para prever as notas de redações com base em características textuais e oferecer um sistema de correção escalável e eficiente.

O projeto foi dividido em etapas sistemáticas, desde a definição do escopo até a entrega dos resultados e a elaboração de storytelling para apresentação. A seguir, são descritas as atividades desenvolvidas em cada etapa, os métodos aplicados e os resultados obtidos.

2. Etapa 1: Definição da Empresa e Escopo

Período: Semana 1 a Semana 2

Nesta etapa inicial, foram definidos os membros do grupo e as premissas fundamentais do projeto, incluindo a escolha da empresa, a base de dados e os objetivos.

Empresa e área de atuação:

Hugging Face foi selecionada como a organização de referência, devido à sua expertise em Processamento de Linguagem Natural (NLP).

A área escolhida foi a educação, com foco no desenvolvimento de ferramentas baseadas em IA para análise textual.

Base de dados utilizada:

A base de dados selecionada foi o aes_enem_dataset, contendo textos de redações, notas atribuídas e metadados relevantes.

Cronograma de atividades:

Um cronograma foi elaborado, com prazos estimados para cada etapa, organizando o fluxo de trabalho até a entrega final.

Resultado: Documento de definição do escopo, contendo a descrição da empresa, a área de atuação, os dados utilizados e o planejamento das atividades.

3. Etapa 2: Definição do Método Analítico

Período: Semana 3 a Semana 4

A segunda etapa concentrou-se no planejamento e na implementação do método analítico para análise dos dados e modelagem.

Linguagem de programação e ferramentas:

Utilizou-se Python, com bibliotecas como pandas, scikit-learn, xgboost e matplotlib.

Análise exploratória dos dados (EDA):

Identificação de padrões iniciais nos dados, incluindo a distribuição das notas.

Extração de características como contagem de palavras e ano de submissão.

Pré-processamento dos dados:

Conversão de variáveis textuais para numéricas com o uso do TF-IDF (n-grams).

Divisão da base em conjuntos de treinamento (70%) e teste (30%).

Modelagem:

Implementação dos modelos Random Forest e XGBoost:

O Random Forest foi ajustado com hiperparâmetros via GridSearchCV.

O XGBoost foi configurado manualmente, priorizando simplicidade e desempenho.

Planejamento da avaliação:

As métricas selecionadas para análise de desempenho foram:

Erro Quadrático Médio (MSE).

Coeficiente de Determinação (R^2).

Distribuição de resíduos.

Resultado: Relatório metodológico detalhado, com descrição do pipeline de pré-processamento, modelagem e planejamento de avaliação.

4. Etapa 3: Implementação e Análise dos Resultados

Período: Semana 5 a Semana 6

A terceira etapa consistiu na execução do modelo e análise detalhada dos resultados.

Treinamento dos modelos:

Os dados foram utilizados para treinar os algoritmos, gerando previsões de notas baseadas nas características textuais.

Resultados obtidos:

Random Forest:

MSE: 2087.28.

R^2: 0.26.

Acurácia: 66.73%.

XGBoost:

MSE: 2143.59.

R^2: 0.24.

Acurácia: 64.82%.

Visualizações geradas:

Gráfico de Dispersão: Demonstra a relação entre notas reais e preditas, evidenciando maior precisão em faixas médias.

Matriz de Confusão (binned): Avalia o desempenho dos modelos em faixas específicas de notas, destacando erros em extremos.

Distribuição dos Resíduos: Mostra a consistência dos modelos, com erros centrados em torno de zero.

Análise dos Resultados:

Os modelos demonstraram potencial para capturar padrões básicos nas notas de redações.

Dificuldades foram identificadas na predição de notas extremas, devido ao desbalanceamento dos dados e à subjetividade da avaliação humana.

5. Etapa 4: Storytelling e Apresentação

Período: Semana 7 a Semana 8

A última etapa foi dedicada à consolidação dos resultados e à elaboração de uma apresentação em formato de storytelling.

Estrutura da apresentação:

Introdução ao problema, destacando a relevância da automatização da correção de redações.

Descrição da metodologia utilizada, com ênfase nos modelos aplicados.

Resultados e análises, acompanhados de gráficos e visualizações.

Proposta de impacto educacional:

Para escolas públicas: Democratizar o acesso à correção de redações, promovendo equidade.

Para escolas privadas: Posicionar-se como diferencial competitivo, oferecendo feedback personalizado aos alunos.

Resultado: Apresentação em formato de PPT, com storytelling alinhado ao propósito do projeto.

6. Conclusão

O projeto mostrou que os algoritmos de aprendizado de máquina podem contribuir significativamente para a automação da correção de redações, com impactos positivos tanto em termos de eficiência quanto de democratização do acesso.

Próximos Passos:

Refinamento do modelo:

Implementar técnicas avançadas de NLP, como embeddings (Word2Vec e BERT).

Ajustar os parâmetros para reduzir os erros em faixas críticas de notas.

Tratamento do desbalanceamento:

Utilizar métodos como SMOTE para lidar com a desproporção entre as classes de notas.

Teste em ambientes reais:

Avaliar a eficácia da solução em escolas públicas e privadas, ajustando o sistema às demandas do contexto educacional.