IFPR Campus Pinhais – Técnico em Informática Documento de Requisitos	
Projeto: Análise da média geral do ENEM-2023 com base no ano de conclusão do ensino médio	**Versão**: 1.0
Cliente: Trabalho de Conclusão de Curso - IFPR Pinhais	**Data**: 03/04/2025

Alunos participantes

Miguel Correia Cruz Rodrigues & Jorge Henrique Kalluf

A – Visão Geral do Sistema

Código em Python com bibliotecas externas comuns nos estudos de ciência de dados, que serão utilizadas para esmiuçar e compreender os microdados do ENEM-2023

B - Requisitos Funcionais

REF01. O código deve deve gerar gráficos, como heatmap, histograma e scatter plot, utilizando as colunas e registros da base de dados, conforme solicitado pelos desenvolvedores do código;

REF02. A análise deve permitir filtrar os dados por diferentes critérios, como região, gênero, escolaridade, entre outros;

REF03. O código deve ser desenvolvido para treinar um modelo de machine learning;

REF04. O notebook deve incluir uma célula inicial que lista todas as bibliotecas necessárias, com instruções sobre o uso de cada uma.

C - Requisitos Não Funcionais

RNF01. O código deve utilizar um dataset aberto disponibilizado na plataforma do governo, gov.br;

RNF02. O código deve ser programado em Python utilizando as seguintes bibliotecas externas: Matplotlib, Seaborn, Numpy e Pandas;

RNF03. O código deve ser programado através da aplicação web Google Colab.

D – Regras de Negócio

RNE01 - O objetivo principal do sistema é realizar um estudo sobre a relação entre dados específicos disponíveis no dataset e a média geral das pessoas que realizaram a prova, sem objetivos comerciais ou de aplicação prática imediata;

RNE02 - O Google Colab deve ser organizado em seções claras, como introdução, preparação dos dados, análises, resultados e conclusões, facilitando a leitura e compreensão do código;

RNE03 - O código deve ser desenvolvido com base no treinamento de diferentes métodos de machine learning, como KNN, random forest e decision tree, permitindo ajustes e otimizações contínuas com base na qualidade dos dados e nas necessidades de aprimoramento das previsões;

RNE04 - Os resultados gerados pelo sistema devem ser validados através da comparação com estudos anteriores e dados conhecidos, assegurando a confiabilidade das conclusões;

RNE05. Os resultados obtidos através das análises não devem ser utilizadas para conclusões finais em discussões;

RNE06 - O código deve ter uma estratégia para lidar com dados faltantes.

E – Protótipo de Telas

F – Glossário

Termo	Descrição
Dataset	Coleção de dados usada para análise e modelagem organizada em um formato estruturado. Esse formato estruturado pode ser uma planilha do Excel, um arquivo CSV, um arquivo JSON ou outros formatos.
Google Colab	O Google Colab é uma plataforma gratuita baseada na nuvem que permite escrever e executar notebooks diretamente no navegador. Ele é especialmente popular entre quem trabalha com machine learning e ciência de dados.
Notebook	Um notebook é um ambiente interativo de programação que permite misturar código executável, texto explicativo (Markdown), gráficos e tabelas em um único documento.
Machine Learning (ML)	Machine Learning é um subconjunto da Inteligência Artificial que permite que um sistema aprenda a partir de métodos de aprendizagem ao ser alimentado com grandes quantidades de dados.
Python	Linguagem de programação orientada a objetos. Amplamente utilizada para Machine Learning e análise de dados.
Biblioteca externa	Recursos criados por terceiros que complementam as bibliotecas nativas da linguagem, úteis para projetos que precisam de funções adicionais.
Aplicação Web	Sistema executado em um navegador, sem a necessidade de estar instalado localmente.
Heatmap	Um heatmap (mapa de calor) é uma representação gráfica que utiliza cores para indicar a intensidade ou magnitude de valores em uma matriz, ajudando a destacar correlações, erros ou distribuições de dados.
Histograma	Um histograma é um gráfico de barras que representa a distribuição de frequências de um conjunto de dados contínuos.
Scatter Plot	Um scatter plot é um gráfico que mostra pontos individuais em um plano cartesiano, geralmente usado para visualizar a relação entre duas variáveis numéricas.
KNN	KNN(K — Nearest Neighbors) é um dos muitos algoritmos usados no campo de machine learning.
Decision Tree	A Decision Tree é um algoritmo de aprendizado de máquina que funciona como uma estrutura hierárquica com base em perguntas feitas sobre os dados. Cada nó da árvore representa uma decisão baseada em uma variável dos dados, cada ramo representa o resultado dessa decisão (por exemplo, “sim” ou “não”), e cada folha representa uma previsão final (uma classe ou valor numérico).
Random Forest	Em vez de confiar em uma única árvore de decisão (decision tree), que pode ser instável, o Random Forest constrói várias árvores de forma aleatória e combina suas previsões para obter um resultado final mais confiável.