Justificativa Técnica - Projeto de Previsão de Nível de Obesidade

1. Introdução

Este projeto tem como objetivo prever o **nível de obesidade** com base em variáveis de estilo de vida, hábitos alimentares e características físicas, utilizando técnicas de *Data Analytics* e *Machine Learning*.
O trabalho abrange todo o pipeline de dados — desde a extração, tratamento e modelagem até a análise interpretativa dos resultados.

---

2. Arquitetura do Projeto

O projeto foi desenvolvido em **Python**, utilizando o **VS Code** como ambiente principal.
A arquitetura foi organizada para garantir **reprodutibilidade, modularidade e segurança** dos dados, conforme a estrutura abaixo:

- `data/` — armazenamento dos dados em diferentes estágios do pipeline
- `src/` — código modular (ETL, engenharia de atributos, modelagem, aplicação)
- `models/` — modelos treinados salvos em formato `.pkl`
- `notebooks/` — análise exploratória e testes
- `reports/` — relatórios e figuras de apoio
- `docs/` — documentação técnica (justificativa e storytelling)

Essa estrutura segue boas práticas de **MLOps** e **Data Engineering**, separando claramente código, dados e relatórios.

---

3. Ferramentas e Tecnologias Utilizadas

Lista e justificativa breve das escolhas:

Categoria	Ferramenta	Justificativa
Linguagem	**Python 3.11+**	Linguagem padrão em análise de dados e ML
IDE	**VS Code**	Desenvolvimento local e modular, sem custos
Bibliotecas	`pandas, numpy, scikit-learn, matplotlib, seaborn, streamlit`	Cobrem todo o ciclo de ML e visualização
Versionamento	**Git + GitHub**	Controle de versão, colaboração e histórico
Documentação	**Markdown e PDF**	Facilita leitura e publicação
Segurança	`.env, .gitignore`	Proteção de dados e variáveis sensíveis

---

4. Segurança e Conformidade

Explique as boas práticas aplicadas à manipulação dos dados:

- Separação entre dados **brutos (raw)** e **processados**.
- Dados brutos mantidos fora do versionamento (`.gitignore`).
- Variáveis sensíveis (ex: caminhos, senhas, chaves) controladas via `.env`.
- Implementação de **anonimização** para possíveis informações pessoais (hash de identificadores).
- Garantia de **reprodutibilidade**: todos os passos podem ser refeitos com os mesmos resultados.

Essas práticas estão alinhadas à **LGPD (Lei Geral de Proteção de Dados)** e aos princípios de segurança de projetos de dados.

---

5. Pipeline de Dados e Modelagem

Explique o fluxo geral do projeto — da leitura à previsão:

O pipeline foi projetado em etapas sequenciais, utilizando scripts modulares no diretório `src/`:

1. **Ingestão:** leitura do dataset original (`obesity.csv`) e criação de cópia segura (`data/interim/`).

2. **Pré-processamento:** padronização de tipos, tratamento de valores ausentes e criação do índice de massa corporal (IMC).

3. **Engenharia de atributos:** codificação de variáveis categóricas e normalização de variáveis numéricas.

4. **Modelagem:** treinamento e validação de modelos supervisionados (RandomForest, XGBoost, etc.).

5. **Avaliação:** comparação de métricas e seleção do modelo com melhor acurácia (>75%).

6. **Deploy:** desenvolvimento de interface interativa em **Streamlit** para demonstração.

---

6. Escolha e Justificativa do Modelo

Foram testados diferentes algoritmos de classificação supervisionada.
O modelo escolhido foi o **Random Forest Classifier**, por equilibrar **desempenho**, **robustez** e **interpretabilidade**.
Além disso, ele permite fácil integração com o pipeline do `scikit-learn`.

Caso seja necessário expandir, está prevista a comparação com **XGBoost** e **LightGBM**.

---

7. Métricas de Avaliação

- **Principal:** Acurácia (mínimo exigido: 75%)

- **Complementares:** F1-Score, Precision, Recall e Matriz de Confusão.

- **Explicabilidade:** SHAP values para entender o impacto das variáveis no resultado.

---

8. Storytelling e Comunicação

O storytelling será desenvolvido com base nos **insights analíticos** obtidos durante a exploração dos dados e no desempenho do modelo.
O objetivo é traduzir os resultados técnicos em **informações compreensíveis e acionáveis**, para diferentes públicos (técnico e não técnico).

As ferramentas previstas para essa etapa são:

- **Streamlit** (dashboard interativo)
- **PowerPoint/PDF** (narrativa de resultados)
- **Gráficos e análises** (matriz de confusão, SHAP, distribuições)

---

9. Próximos Passos

- Implementar scripts restantes de feature engineering e modelagem.
- Gerar relatório analítico e dashboard completo.
- Refinar modelo com validação cruzada e tuning de hiperparâmetros.
- Criar storytelling final e vídeo de apresentação.

---

10. Referências Técnicas

- Documentação oficial do scikit-learn
- Google Cloud Architecture Center (para inspiração estrutural)
- Guias de boas práticas em projetos de Data Science e MLOps

Resultado final esperado

Um pipeline completo, seguro e reprodutível, acompanhado de:

- Justificativa técnica (este documento)

- Storytelling analítico
- Dashboard interativo
- Modelo com acurácia > 75%