	Especificação de Projeto	
C.E.S.A.R school	**DISCIPLINA:** Aprendizado de Máquina	**PERÍODO:** 2024.2 **UNIDADE:** 2a

1. Introdução

Este projeto visa proporcionar uma experiência prática de análise de dados aplicada à saúde pública, utilizando um problema real e atual no Brasil: os casos de sífilis congênita. Com base no dataset “Clinical and sociodemographic data on congenital syphilis cases, Brazil, 2013-2021,” você deverá aplicar técnicas de pré-processamento, classificação e regressão para investigar fatores clínicos e sociodemográficos associados aos desfechos de sífilis congênita. O projeto deverá utilizar os conceitos de aprendizagem de máquina vistos na disciplina para criar modelos preditivos e extrair *insights* sobre os fatores de risco e padrões de saúde pública, capacitando vocês a aplicarem conhecimentos teóricos em um contexto de impacto social.

A análise deverá se concentrar na variável **VDRL_RESULT** (resultado de exame) como alvo para a tarefa de classificação e **AGE** (idade) como alvo para a regressão. Você deverá realizar um processo de validação e justificar suas decisões em cada etapa, fornecendo um relatório final que documenta o fluxo de trabalho e os resultados obtidos. O projeto também incentivará a reflexão sobre o impacto dos fatores analisados para a prevenção e controle da sífilis congênita.

2. Objetivos

Desenvolver habilidades de análise de dados aplicadas à saúde pública. O projeto envolve pré-processamento de dados, classificação, regressão e análise dos fatores que contribuem para os desfechos de saúde, utilizando técnicas de validação e interpretação de modelos. Especificamente, os objetivos incluem:

1. **Compreender e preparar dados de saúde pública:**
    - Realizar a análise exploratória dos dados, identificar problemas de qualidade e aplicar técnicas de pré-processamento adequadas, como *one-hot encoding* e tratamento de valores ausentes.
2. **Desenvolver e avaliar modelos de classificação:**
    - Utilizar técnicas de classificação para prever **VDRL_RESULT**, justificando as escolhas de modelos, parâmetros e estratégias de validação e avaliando o desempenho.
3. **Desenvolver e avaliar modelos de regressão:**
    - Utilizar a variável **AGE** para identificar relações com variáveis clínicas e sociodemográficas
4. **Interpretar resultados e relacionar com práticas de saúde pública:**
    - Identificar fatores de risco ou associações relevantes para a prevenção de sífilis congênita.

5. **Desenvolver habilidades de comunicação técnica:**
    - Redigir um relatório que documenta todas as etapas do projeto, incluindo justificativas, metodologia e resultados, e realizar uma apresentação com os principais *insights*.

3. Descrição do dataset

- **Nome do Dataset**: Clinical and sociodemographic data on congenital syphilis cases, Brazil, 2013-2021 - **Link para Acesso**: Dataset no Mendeley
- **Artigo:** Predicting congenital syphilis cases: A performance evaluation of different machine learning models (https://doi.org/10.1371/journal.pone.0276150)
- **Características do Dataset**:
    - Contém dados clínicos e sociodemográficos relacionados a casos de sífilis congênita.
    - Variáveis incluem características do paciente, condições de saúde e fatores sociodemográficos.
    - O artigo descreve detalhes do dataset e apresenta uma avaliação que deve servir como *baseline* para o desenvolvimento do projeto.

4. Etapas do Projeto - Metodologia

1. **Análise Exploratória e Pré-processamento**:
    - **Objetivo**: Familiarizar e identificar problemas de qualidade relacionados aos dados, como valores ausentes, redundâncias e outliers.
    - **Atividades**:
        - Realizar uma análise exploratória inicial para entender a distribuição das variáveis (realizar análise de correlação entre os atributos e distribuição das classes).
        - **Considere aplicar pelo menos uma das estratégias utilizadas no artigo para lidar com o desbalanceamento dos dados**.
            1. **Bônus**: utilizar uma técnica de oversampling como o SMOTEENN ou qualquer outra não utilizada pelo artigo.
2. **Classificação (Variável-Alvo: VDRL_RESULT)**:
    - **Objetivo**: Construir modelos de classificação para prever o resultado do exame VDRL, que é um indicador de sífilis congênita.
    - **Atividades**:
        - Escolher e implementar pelo menos dois modelos de classificação (ex: Decision Trees e Random Forest).
        - Justificar a seleção dos modelos e os parâmetros escolhidos para cada um.
        - **Validação**: Aplicar validação cruzada (k-Fold) e documentar as métricas de desempenho (precisão, recall, F1-score).
        - **Interpretação**: Interpretar o impacto das variáveis no modelo, explicando as principais variáveis que influenciam a predição (feature importance).
3. **Regressão (Variável-Alvo: AGE)**:

- **Objetivo**: Modelar a relação entre os fatores sociodemográficos e clínicos e a variável **AGE** (idade).
- **Atividades**:
  - Escolher um modelo de regressão, como regressão linear ou regressão Ridge, e justificar a escolha (Ex.: RandomForestRegressor).
  - Avaliar o desempenho usando métricas como MAE, RMSE e MAPE.
  - Documentar e interpretar os fatores que mais impactam o modelo.

4. **Análise de Fatores e Discussão sobre Prevenção**:
   - **Objetivo**: Explorar o impacto de fatores sociodemográficos e clínicos em desfechos de saúde e discutir *insights*.
   - **Atividades**:
     - Identificar variáveis que contribuem para a ocorrência de sífilis congênita e sugerir intervenções preventivas.
     - Relacionar os resultados do modelo com potenciais políticas de saúde pública.
5. **Relatório e Apresentação dos Resultados**:
   - **Relatório Final**: Um relatório com (use este template):
     - Descrição detalhada do dataset e do processo de pré-processamento.
     - Explicação das escolhas metodológicas, resultados dos modelos, análise das variáveis mais influentes e sugestões para intervenção.
     - **Mínimo 6 páginas**
   - **Apresentação**: Uma apresentação (3-5 minutos) com uma visão geral dos principais resultados, *insights*, limitações e sugestões de melhoria no modelo.

5. Entrega

A entrega do projeto deverá conter código (*.**ipynb**) e o relatório. O relatório deverá ser submetido via Google Classroom e conter o link para o repositório do Github onde estará o código desenvolvido. O arquivo `README.md` deverá conter as seguintes informações:

1. Nome e sobrenome dos membros do projeto e seus respectivos **usuários no GitHub (@fulano, @beltrano, @sicrano)**.
2. Nome da disciplina: **Aprendizado de Máquina - 2024.2**.
3. Nome da instituição de ensino: **CESAR School**.

6. Critérios de Avaliação

1. **Exploração e Pré-processamento do Dataset** (10%):
   - Avaliação da compreensão do dataset e justificativas das técnicas de pré-processamento.
2. **Modelos de Classificação** (30%):
   - Qualidade das justificativas para a escolha dos modelos de classificação para **VDRL_RESULT** e análise das métricas de desempenho.
   - Aplicação e interpretação dos resultados da validação cruzada.
3. **Modelos de Regressão** (30%):
   - Justificativas e interpretação dos modelos de regressão para **AGE**.
   - Análise de desempenho e impacto das variáveis contínuas (idade).
4. **Relatório e Apresentação** (30%):

- Qualidade do relatório, com explicação das metodologias e análise dos resultados.
- Clareza e organização na apresentação dos principais *insights*.