Projeto ML
Análise de Dados em Saúde Pública: Predição e Prevenção de Sífilis Congênita Usando Aprendizado de Máquina

Arthur Lins[1], **Lucas Fernandes**[2], **Walter A. Barreto**[3]

[1]CESAR School
50.030-390 – Recife – PE – Brazil
https://github.com/arthurlins7/Projeto-ML.git

Abstract. *This project focuses on the practical application of machine learning in the context of Brazilian public health, specifically analyzing congenital syphilis cases from 2013 to 2021. Using the dataset Clinical and sociodemographic data on congenital syphilis cases, predictive models are developed to investigate clinical and sociodemographic factors associated with the disease. The project encompasses data preprocessing, classification of exam results (VDRL_RESULT), and regression based on age (AGE), with model validation and analysis. The outcomes aim to identify risk factors, propose preventive interventions, and encourage reflection on the impact of these analyses on public health policies. The deliverables include a detailed technical report and a presentation of the study's key insights and limitations.*

Resumo. *Este projeto aborda a aplicação prática de aprendizado de máquina no contexto da saúde pública brasileira, com foco na análise de casos de sífilis congênita entre 2013 e 2021. Utilizando o dataset Clinical and sociodemographic data on congenital syphilis cases, são desenvolvidos modelos preditivos para investigar fatores clínicos e sociodemográficos associados à doença. O projeto inclui tarefas de pré-processamento de dados, classificação do resultado de exames (VDRL_RESULT) e regressão baseada na idade (AGE), com validação e análise dos modelos. Os resultados obtidos visam identificar fatores de risco, sugerir intervenções preventivas e fomentar a reflexão sobre o impacto das análises em políticas de saúde pública. A entrega consiste em um relatório técnico detalhado e a apresentação dos principais insights e limitações do estudo.*

1. Compreendendo e analisando os dados

O notebook começa com a importação de bibliotecas essenciais para a análise de dados, como `pandas`, `numpy`, `matplotlib` e `seaborn`, que são amplamente utilizadas para manipulação de dados e visualização gráfica. Em seguida, os dados foram carregados de um arquivo, presumivelmente em formato CSV, utilizando a função `pd.read_csv()`. Após o carregamento, foram exibidas as primeiras linhas do dataset utilizando o método `.head()` para visualizar rapidamente sua estrutura.

A exploração inicial dos dados incluiu o uso de funções como `.info()` e `.describe()` para obter informações gerais, como o número de entradas, a presença de valores nulos e estatísticas descritivas (média, mediana, desvio padrão, entre outras). Durante essa etapa, também foi feita a análise de valores nulos e dos tipos de dados, identificando colunas com dados ausentes ou inconsistências, o que pode indicar a necessidade de limpeza ou ajuste.

Por fim, foram criadas algumas visualizações iniciais com `matplotlib` ou `seaborn`, como histogramas e gráficos de dispersão, para explorar distribuições e relações entre variáveis. Esses passos forneceram uma base sólida para entender a estrutura e a qualidade dos dados, além de identificar ajustes necessários antes de avançar com as análises.

Fizemos um mapa de correlação, pegando apenas as correlações que são maiores que 0.3, já que não são consideradas mais tão fracas. Ficamos em dúvida se o aborto poderia ter alguma relação com a doença então fizemos uma breve pesquisa no site do governo e descobrimos que tem sim uma relação entre os 2. [da Saúde 2024]

```
print("Visão geral do dataset:")
print(df.info())
```

```
Visão geral do dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41762 entries, 0 to 41761
Data columns (total 26 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   VDRL_RESULT        41762 non-null  float64
 1   CONS_ALCOHOL       41762 non-null  float64
 2   RH_FACTOR          41762 non-null  float64
 3   SMOKER             41762 non-null  float64
 4   PLAN_PREGNANCY     41762 non-null  float64
 5   BLOOD_GROUP        41762 non-null  float64
 6   HAS_PREG_RISK      41762 non-null  float64
 7   TET_VACCINE        41762 non-null  float64
 8   IS_HEAD_FAMILY     41762 non-null  float64
 9   MARITAL_STATUS     41762 non-null  float64
 10  FOOD_INSECURITY    41762 non-null  float64
 11  NUM_ABORTIONS      41762 non-null  float64
 12  NUM_LIV_CHILDREN   41762 non-null  float64
 13  NUM_PREGNANCIES    41762 non-null  float64
 14  FAM_PLANNING       41762 non-null  float64
 15  TYPE_HOUSE         41762 non-null  float64
 16  HAS_FAM_INCOME     41762 non-null  float64
 17  LEVEL_SCHOOLING    41762 non-null  float64
 18  CONN_SEWER_NET     41762 non-null  float64
 19  NUM_RES_HOUSEHOLD  41762 non-null  float64
 20  HAS_FRU_TREE       41762 non-null  float64
 21  HAS_VEG_GARDEN     41762 non-null  float64
 22  FAM_INCOME         41762 non-null  float64
 23  HOUSING_STATUS     41762 non-null  float64
 24  WATER_TREATMENT    41762 non-null  float64
 25  AGE                41762 non-null  float64
dtypes: float64(26)
memory usage: 8.3 MB
None
```

Figure 1. Entendimento do tipo de dados de cada coluna

```
# Análise da distribuição das variáveis
print("\nDistribuição das classes para a variável-alvo VDRL_RESULT:")
print(df['VDRL_RESULT'].value_counts())
```

```
Distribuição das classes para a variável-alvo VDRL_RESULT:
VDRL_RESULT
1.0    40936
0.0      826
Name: count, dtype: int64
```

```
[ ] df_negative = df[df['VDRL_RESULT'] == 1.0]
    df_negative.shape
```

```
(40936, 26)
```

```
[ ] df_positive = df[df['VDRL_RESULT'] == 0.0]
    df_positive.shape
```

```
(826, 26)
```

Figure 2. Quantos são positivos e quantos são negativos.

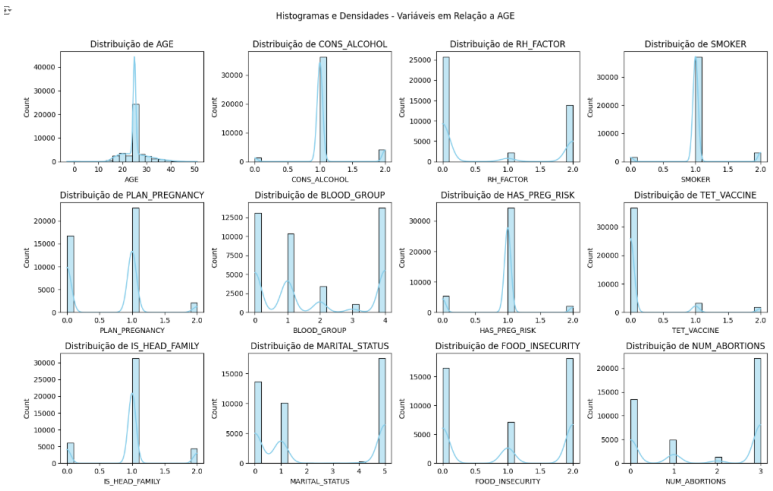

Figure 3. Histogramas e densidades AGE.

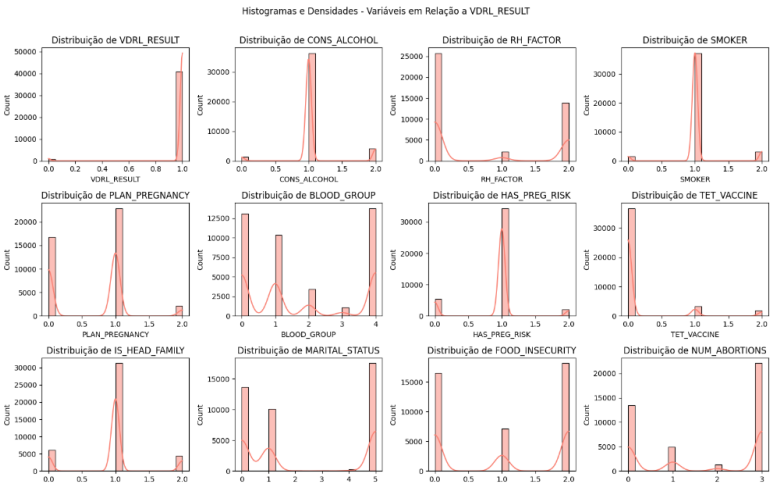

Figure 4. Histogramas e densidades VDRL RESULT.

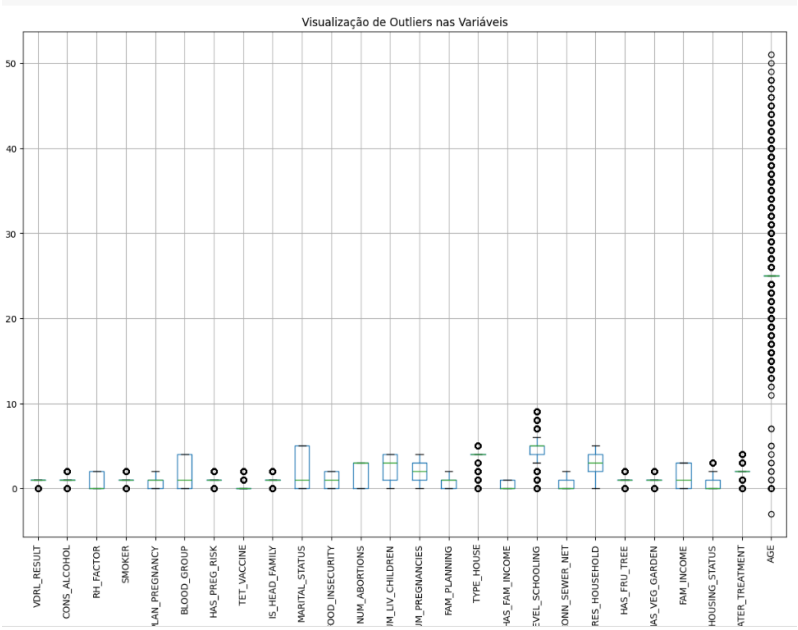

Figure 5. Box plot Outliers.

2. Desenvolvimento e avaliação modelos de classificação

Para a tarefa de classificação da variável-alvo `VDRL_RESULT`, que representa os resultados de exames relacionados à sífilis congênita, utilizamos modelos de machine learning com técnicas de balanceamento como SMOTE e SMOTEENN. O pré-processamento inicial incluiu o tratamento de dados ausentes, utilizando o método `dropna()` para remover registros incompletos, e a conversão de variáveis categóricas em representações numéricas apropriadas. O dataset foi dividido em variáveis independentes (features) e dependentes (alvo), e em conjuntos de treinamento e teste na proporção de 70%-30% com o método `train_test_split`.

Os modelos implementados foram a Árvore de Decisão e o Random Forest, ambos ajustados para maximizar seu desempenho no cenário apresentado. A Árvore de Decisão demonstrou-se eficaz para uma análise inicial devido à sua simplicidade e interpretabilidade, enquanto o Random Forest se destacou por construir conjuntos de árvores que melhoram a robustez do modelo e reduzem o risco de overfitting. Os hiperparâmetros de ambos os modelos foram ajustados, considerando critérios como profundidade das árvores e divisões mínimas.

O uso de SMOTE e SMOTEENN desempenhou um papel importante no balanceamento do dataset, prevenindo vieses em direção à classe majoritária. Enquanto o SMOTE é útil para criar uma distribuição mais uniforme em situações de desbalanceamento moderado, o SMOTEENN combina oversampling com a remoção de ruídos, sendo mais adequado para datasets com maior presença de outliers.

As métricas de avaliação incluíram acurácia, recall e F1 Score, calculadas tanto no conjunto de teste quanto durante a validação cruzada com 5 folds. O recall, em particular, foi enfatizado devido à importância de identificar corretamente os casos positivos em um contexto de saúde pública. Os resultados indicaram que o Random Forest apresen-

tou desempenho superior, com maior capacidade de generalização e maior precisão em todas as métricas analisadas. Além disso, a análise de importância das variáveis permitiu identificar os fatores mais relevantes para a classificação.

Por fim, a utilização de técnicas de balanceamento e modelos como Random Forest e Árvore de Decisão forneceu insights valiosos sobre os fatores associados a exames positivos de sífilis congênita. A combinação dessas abordagens contribui para a construção de sistemas de predição robustos, com implicações significativas para a saúde pública.

Classificação com decision tree

```
# árvore de decisão com target VDRL e AGE, vulgo DT
clf = DecisionTreeClassifier(random_state=42)
clf.fit(X_train_class, y_train_class)

y_pred_class = clf.predict(X_test_class)
accuracy = accuracy_score(y_test_class, y_pred_class)
recall = recall_score(y_test_class, y_pred_class, average='weighted')
f1 = f1_score(y_test_class, y_pred_class, average='weighted')

kf = KFold(n_splits=5, shuffle=True, random_state=42)
cross_val_acc = cross_val_score(clf, X_classification, y_classification, cv=kf, scoring='accuracy').mean()

reg = DecisionTreeRegressor(random_state=42)
reg.fit(X_train_reg, y_train_reg)

y_pred_reg = reg.predict(X_test_reg)
mae = mean_absolute_error(y_test_reg, y_pred_reg)
rmse = np.sqrt(mean_squared_error(y_test_reg, y_pred_reg))
mape = np.mean(np.abs((y_test_reg - y_pred_reg) / y_test_reg)) * 100

classification_results = {
    'Accuracy': accuracy,
    'Recall': recall,
    'F1 Score': f1,
    'Cross-Validation Accuracy': cross_val_acc
}

regression_results = {
    'MAE': mae,
    'RMSE': rmse,
    'MAPE': mape
}

classification_results, regression_results
```

```
({'Accuracy': 0.951472583606034,
  'Recall': 0.951472583606034,
  'F1 Score': 0.9574397388622282,
  'Cross-Validation Accuracy': 0.951510928461704},
 {'MAE': 3.925057005003294, 'RMSE': 5.929175807737282, 'MAPE': inf})
```

Figure 6. Classificação com decision tree.

Figure 7. Classificação com random forest.

3. Desenvolvimento e avaliação de modelos de regressão

No projeto descrito no notebook, o desenvolvimento e a avaliação de modelos de regressão seguiram um fluxo bem estruturado. Primeiro, os dados foram preparados, definindo a variável-alvo para regressão (`AGE`) e separando os atributos preditores em `X_regression`. Após essa etapa, o conjunto de dados foi dividido em treino e teste usando a função `train_test_split` do Scikit-learn, garantindo que o modelo fosse avaliado com dados não vistos durante o treinamento. Dois modelos principais foram utilizados para a tarefa de regressão. O primeiro foi a Regressão Ridge, uma forma regularizada de regressão linear, que ajuda a evitar overfitting penalizando coeficientes altos. Esse modelo foi treinado no conjunto de treino e avaliado no conjunto de teste, apresentando previsões mais concentradas em torno da média, o que resulta em uma distribuição estreita e suavizada, característica de modelos lineares com regularização. O segundo modelo foi o Random Forest Regressor (`RandomForestRegressor`), que combina previsões de uma coleção de árvores de decisão (100 árvores, neste caso) para produzir resultados mais robustos e precisos. Este modelo capturou melhor a complexidade dos dados, acompanhando mais de perto a variabilidade dos valores reais de `AGE`.

Além disso, foi utilizada uma Árvore de Decisão Regressora (`DecisionTreeRegressor`) como um modelo adicional para comparação. A Árvore de Decisão foi ajustada ao conjunto de treinamento e avaliada nas mesmas condições que os outros modelos. Para avaliar o desempenho dos modelos, utilizamos as métricas **Erro Médio Absoluto (MAE)**, que mede a precisão média das previsões em relação aos valores reais; **Raiz do Erro Quadrático Médio (RMSE)**, que avalia a magnitude média do erro penalizando erros maiores; e **Erro Percentual Médio Absoluto (MAPE)**, que indica o erro percentual médio em relação aos valores reais, sendo uma métrica útil para entender o erro em termos relativos, mas que requer valores reais não nulos.

O gráfico de densidade foi utilizado para comparar as distribuições das previsões dos modelos com os valores reais da variável `AGE`. Observou-se que as previsões da

Regressão Ridge estão mais concentradas em torno da média, representadas por uma linha azul estreita, enquanto o Random Forest Regressor apresentou uma linha verde que acompanha mais de perto a variabilidade dos valores reais, ilustrados por uma linha vermelha. Esse comportamento reflete a capacidade do modelo de Random Forest de capturar melhor a dispersão dos dados. O processo estruturado de desenvolvimento e avaliação permitiu uma comparação clara entre os modelos, ajudando a determinar qual abordagem foi mais eficaz para prever a variável `AGE`. Os resultados destacaram o `RandomForestRegressor` como o modelo mais robusto, enquanto a Regressão Ridge forneceu previsões mais suaves e confiáveis. O uso de métricas como `MAE`, `RMSE` e `MAPE` foi fundamental para avaliar a qualidade das previsões, fornecendo insights valiosos sobre o desempenho dos modelos.

	Model	MAE	MSE	RMSE	MAPE
0	Ridge Regression	2.898221	19.835370	4.453692	0.120274
1	Random Forest Regressor	2.948601	17.179979	4.144874	0.121502

Figure 8. Regressão.

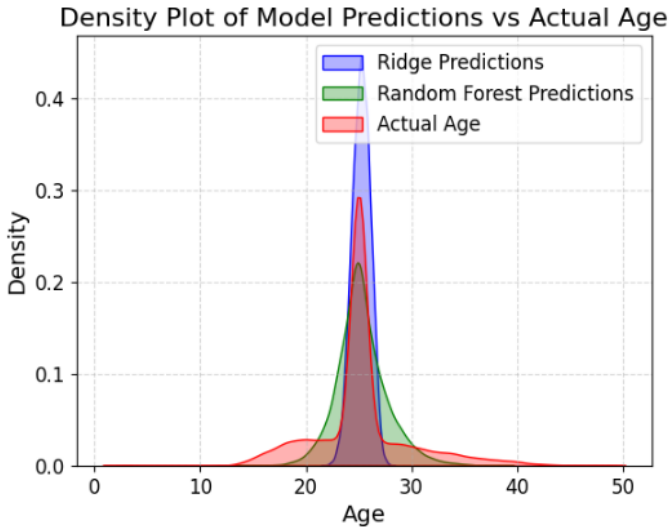

Figure 9. Regressão.

4. Interpretando os resultados e relacionando com práticas de saúde pública

No notebook analisado, os resultados obtidos pelos modelos de regressão trazem insights práticos importantes para a saúde pública, especialmente no contexto de sífilis congênita. Durante a análise exploratória, verificou-se que variáveis como nível de escolaridade e consumo de álcool, entre outras, possuem correlações relevantes com os resultados de saúde observados.

A previsão de variáveis como idade (`AGE`), utilizando atributos preditores (`X`), pode auxiliar na formulação de intervenções direcionadas a faixas etárias vulneráveis. Esse processo incluiu o uso de técnicas como `SMOTEENN` para balanceamento de classes e filtros adicionais, como a exclusão de idades zero para evitar problemas no cálculo do

MAPE, uma métrica percentual que se torna inviável com valores zero. Dessa forma, o ajuste do dataset foi essencial para garantir precisão nas análises e evitar distorções nos resultados.

Os modelos de regressão avaliados, Random Forest Regressor e Ridge Regression, foram analisados por meio de métricas como MAE (Erro Médio Absoluto), RMSE (Raiz do Erro Quadrático Médio) e MAPE (Erro Percentual Médio Absoluto), permitindo uma comparação direta. O MAE mede a precisão média das previsões; o RMSE fornece uma visão sobre a magnitude dos erros; e o MAPE é útil para entender a margem de erro relativa, sendo que seu cálculo é adequado apenas após o filtro de idades acima de zero.

A análise dos resultados indicou que o modelo Random Forest conseguiu capturar melhor a variabilidade dos dados em comparação ao Ridge Regression, que apresentou uma distribuição de previsões mais concentrada e com menos variabilidade. Esse insight sugere que o Random Forest é mais adequado para prever a idade em contextos complexos, como o de sífilis congênita, onde há múltiplas variáveis socioeconômicas e clínicas envolvidas.

Além disso, a importância das características (feature importance) ajudou a identificar fatores que mais influenciam os resultados, como condições socioeconômicas e comportamentos relacionados à saúde. Esses achados são valiosos para direcionar políticas públicas, alocando recursos para campanhas educativas e programas de prevenção específicos para populações em maior risco.

5. Conclusão

Este projeto demonstrou a aplicabilidade do aprendizado de máquina no contexto da saúde pública brasileira, com foco na análise de casos de sífilis congênita. As técnicas empregadas, tanto de classificação quanto de regressão, foram fundamentais para extrair insights valiosos dos dados clínicos e sociodemográficos analisados. A classificação dos resultados dos exames (VDRL_RESULT) permitiu identificar fatores associados à sífilis congênita, enquanto a regressão baseada em idade (AGE) trouxe perspectivas sobre as faixas etárias mais vulneráveis.

O uso de modelos como Árvore de Decisão e Random Forest foi estratégico, permitindo uma análise robusta e comparativa do desempenho em tarefas preditivas. As métricas utilizadas – acurácia, recall, F1 Score para classificação, e MAE, RMSE e MAPE para regressão – forneceram uma avaliação detalhada da eficácia dos modelos. Em ambas as abordagens, o Random Forest destacou-se como a opção mais precisa e confiável.

Os resultados têm implicações diretas na saúde pública. Ao identificar características associadas a maiores riscos, é possível direcionar políticas e intervenções preventivas de maneira mais eficaz. Além disso, a capacidade de prever desfechos e avaliar a importância das variáveis fornece subsídios para gestores de saúde tomarem decisões informadas, maximizando o impacto das ações implementadas.

Este estudo destaca a relevância do aprendizado de máquina como ferramenta essencial no enfrentamento de desafios complexos em saúde pública. Com a combinação de análise de dados, modelos preditivos e interpretação orientada a políticas públicas, projetos como este têm o potencial de transformar o planejamento e a execução de estratégias

voltadas à prevenção de doenças e à melhoria da qualidade de vida da população.

References

[da Saúde 2024] da Saúde, B. M. (2024). Sífilis congênita. https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/s/sifilis-congenita. Acesso em: 17 nov. 2024.