Relatório Final Casos de Sífilis Congênita

Beatriz Silva Vilarim Santana[1], **João Ricardo Leitão Barros**[1], **João Victor Almeida Vieira de Melo**[1]

[1]C.E.S.A.R. School
Recife - PE - Brasil

bsvs@cesar.school, jrlb@cesar.school, jvavm@cesar.school

Abstract. *This article summarizes the analysis process of the Clinical and Sociodemographic Data dataset of Congenital Syphilis Cases, in Brazil between the years 2013 and 2021. The processes of pre-processing, hodout, training and calculation of metrics on the set of data will be shown along with conclusions drawn by the team.*

Resumo. *Este artigo resume o processo de analise do dataset Dados Clínicos e Sociodemográficos dos Casos de Sífilis Congênita (Clinical and sociodemographic data on congenital syphilis cases), no Brasil entre os anos de 2013 e 2021. Serão mostados os processos de pré-processamento, hodout, treino e cálculo de métricas sobre o dataset e conclusão tiradas pela equipe.*

1. Contexto

O estudo do artigo, busca identificar possíveis portadores de sífilis congênita atráves de diversos fatores, para assim, terem melhores possibilidades de tratamento preventivo de transmissão entre mãe e bebê. Os dados são de 2013 até 2021, e provêm da iniciativa Programa Mãe Coruja de Pernambuco, um programa do Sistema Público de Saúde (SUS) que busca dar suporte a mães grávidas antes e depois da gestão, colocando a criança em observação complementar até seus 5 anos.

2. Dataset

O Dataset contém cerca de 41,762 entradas e 25 colunas com dados catégoricos, mas apenas com um dado descritivo sendo 'AGE'(idade). Os dados levam em conta as variáveis mais comuns em relação a ter uma baixa imunidade (se é fumante ou consome álcool), assim como se já teve alguma doença grave ou, outras gravidez ou abortos, mas além disso, também leva em conta fatores do ambiente, porcurando saber se a mulher é a principal provedora da família, suas condições matrimoniais, alimentação, escolaridade e situação social.

A grande quantidade de informação parece ajudar a entender a situação das pacientes e tenta procurar fatores diretos e indiretos que possam influenciar uma maior vulnerabilidade para o desenvolvimento da doença, contudo, justamente por ter uma grande quantidade de colunas não relacionadas, o treinamento dos modelos se torna mais difícil.

3. Pré-Processamento

Para começar o processo de avaliação e chegarmos aos resultados da previsão, foi realizado uma breve análise dos dados e uma leitura do artigo de [Teixeira et al. 2023]. Foi

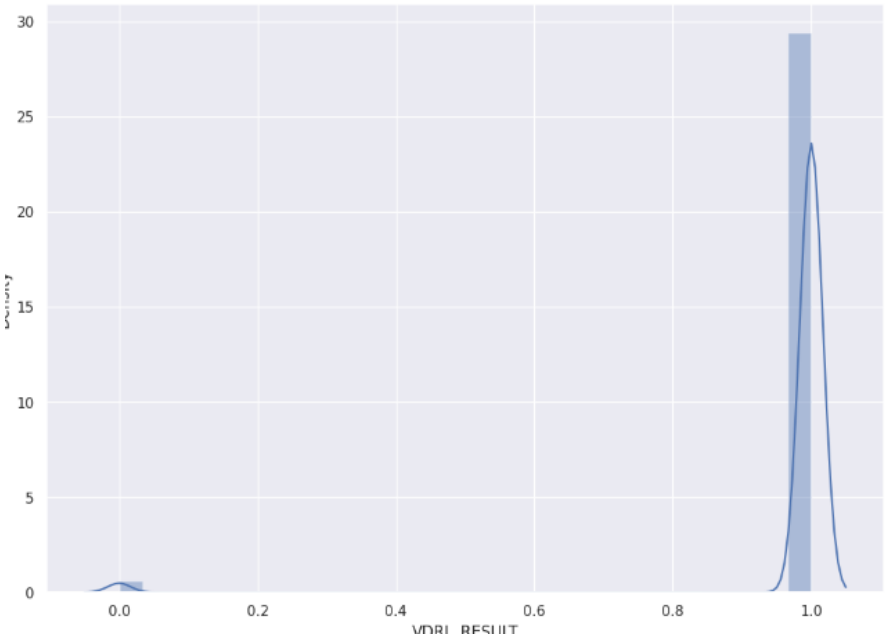

Figure 1. Exemplo da Dipersão do Dataset

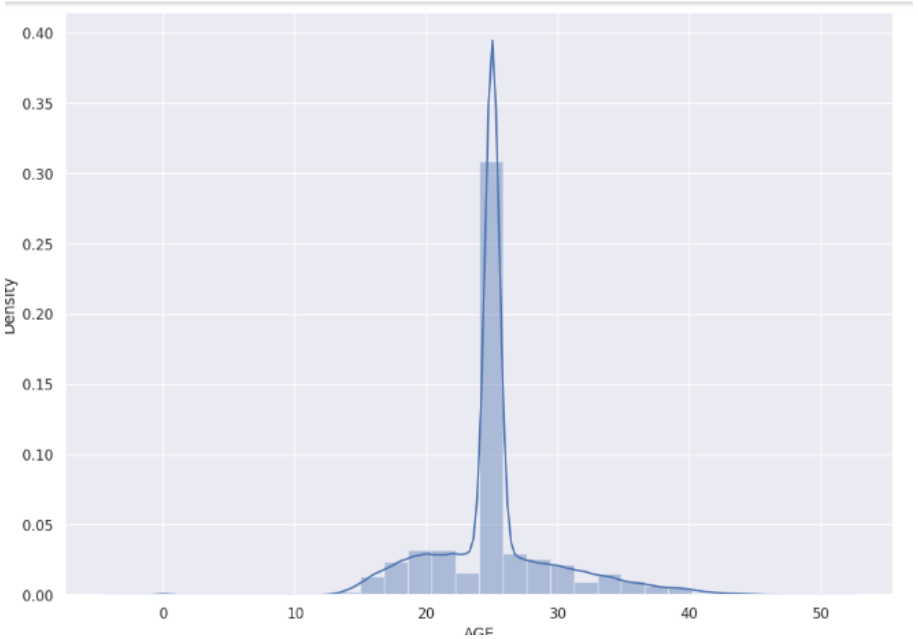

Figure 2. Exemplo da Dipersão do Dataset

observado que os dados disponibilizados no dataset já continha alguns ajustes realizados para melhor trabalhar com eles, por exemplo todos os valores contidos no dataset são numéricos, o que facilita na utilização de algoritmos especializados para realizar previsões baseando-se nos dados disponibilizados. Também foi decidido separar os dados do dataset em duas amostras, uma para uma análise classificatória, utilizando a coluna 'VDRL_RESULT' como variável alvo, e o outra para a análise de regressão com a coluna AGE como variável alvo. Para a classificação, foi observado que tamém que o dataset possui um desbalanceamento quanto a quantidade de dados contidos em 'VDRL_RESULT', pois aproximadamente 98% dos valores, 40936 para ser mais exato, encontrados nesta coluna contém o valor 1(casos negativos) e apenas 2% dos valores, 826 para ser mais exato, possuem um valor igual a 0(casos positivos). Foi concluído que este desbalanceamento viria a causar problemas para a previsão, assim foi decidido treinar os modelos de previsão utilizando um dataset que contém 100% dos valores de casos positivos, isto é os 826 casos encontrados originalmente, e pegar mais 826 casos negativos, de forma aleatória, para gerar um novo dataset que contém 1652 dados, e a partir dele, treinar os modelos de classificação.

4. Modelos

4.0.1. Classificação

Como dito na secção anterior, os dados analisados para a classificação serão do novo dataset gerado a partir da análise preliminar, onde foi constatado um desbalanceamento na variável alvo. Logo temos um dataset que contém 50% dos resultados positivos para sífilis congênita e 50% dos resultados negativos. Foi decidido utilizar a técnica de SMOOTEEN, como sugerido pelo professor, para tentar buscar valores de acurácia maiores que os obtidos por [Teixeira et al. 2023]. Mas primeiro foi realizada a separação dos dados do dataset em duas partes, 70% da base para treino e os 30% da base restantes para teste. Em seguida foi aplicado o SMOOTEEN nos valores da base de treino. Foi utilizado dois modelos de classificação para poder obter as previsões, sendo eles o Random Forest Classifier e o XGBoost Classifier. Onde buscamos validar esses algoritmos utilizando suas precisões, acurácias, recall e F1-score. E por fim comparar os resultados dos dois e observar caso haja um que se sobressaia quanto ao outro, além de validar se é possível prever casos positivos de sífilis congênita em bebês com as informações obtidas.

Primeiro iremos observar o random forest classifier. Foi observado que os valores de Precisão, Acurácia, Recall e F1-score estão em torno de 55% e 70%, com uma média em torno dos 60%. É importante constar que poderíamos ter obtido resultados melhores caso removendo colunas que possam ser consideradas redundantes ou até mesmo inúteis para a previsão, porém o grupo não consegui concluir quais colunas remover, dado ao tempo curto para a entrega deste trabalho, mas é uma opção que pode elevar os valores obtidos pelo modelo.

Agora, utilizando o XGBoost classifier para realizar a mesma operação realizada pelo random forest anteriormente, e comportou-se de forma parecida como o random forest, obtendo os valores das métricas já mencionadas em torno de 55% e 70%, onde é possível concluir que ambos os algoritmos tem performance muito parecidas para o problema apresentado, logo a escolha para a utilização de qualquer um dos dois para a previsão neste caso não é relevante visto que os resultados são minimamente diferentes.

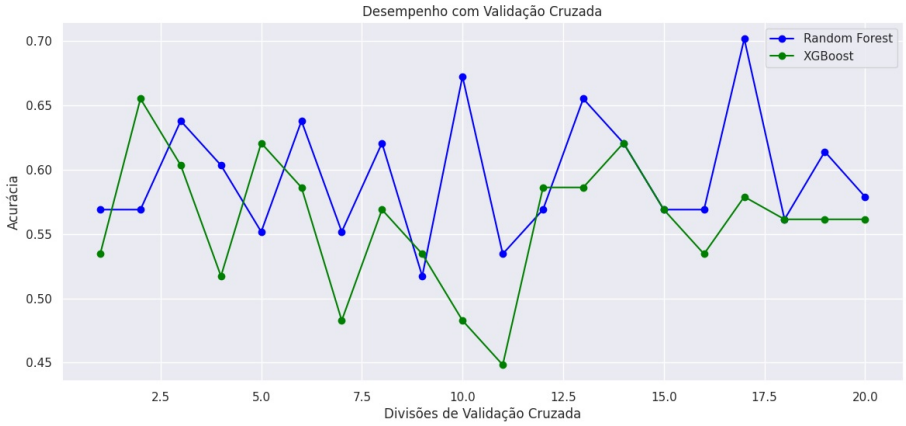

Figure 3. Comoparação da acurácia entre RFClassifier e XGBClassifier

Também é importante apontar, que para ambos os modelos de classificação escolhidos, foi aplicado a técnica de K-Folds para obter os valores da acurácia de cada modelo, sendo a quantidade de folds k = 20. Assim, é possível observar melhor o comportamento dos modelos, quanto a sua performance.

```
Accuracy score for each fold (Random Forest): [0.6, 0.64, 0.6, 0.64, 0.76, 0.56, 0.68, 0.64, 0.64, 0.56, 0.6, 0.68, 0.6, 0.72, 0.48, 0.6, 0.625, 0.75, 0.7083, 0.5833]
Average accuracy across 20 folds: 0.63
```

Figure 4. Valores da acurácia do RFClassifier depois de 20 folds

```
Accuracy score for each fold (XGB): [0.56, 0.64, 0.6, 0.48, 0.68, 0.56, 0.76, 0.52, 0.64, 0.6, 0.56, 0.64, 0.68, 0.72, 0.6, 0.56, 0.6667, 0.625, 0.625, 0.5833]
Average accuracy across 20 folds: 0.61
```

Figure 5. Valores da acurácia do XGBClassifier depois de 20 folds

4.0.2. Regressão

Para a regressão, o grupo tem como variável alvo a coluna de 'AGE', e para essa situação, decidimos remover algumas colunas do dataset, seguindo o sugerido pela PMCB segundo o artigo de [Teixeira et al. 2023]. As colunas removidas foram: 'CONS_ALCOHOL', 'RH_FACTOR', 'SMOKER', 'BLOOD_GROUP', 'TET_VACCINE', 'IS_HEAD_FAMILY', 'TYPE_HOUSE', 'HAS_FAM_INCOME', 'CONN_SEWER_NET', 'NUM_RES_HOUSEHOLD', 'HAS_FRU_TREE', 'HAS_VEG_GARDEN', 'HOUSING_STATUS', 'WATER_TREATMENT'. O objetivo da retirada dessas colunas é buscar atingir os melhores valores para os modelos possíveis, assim tendo previsões mais precisas.

Foi observado também que, a coluna que representa a variável alvo para a regressão, possui valores menores ou iguais a zero, logo para evitar qualquer problema, foram utilizadas apenas as linhas que contêm valores maiores que zero, foi obtido um dataset com 41730 linhas do original de 41762 linhas, uma redução de 32 linhas. Também foi alterado os valores das colunas que contém os dados categóricos, para evitar o uso de valores iguais a zero, com o intuito de evitar maiores problemas, os valores desses dados foram incrementados em um, logo colunas que continham valores como 0, 1 e 2, passaram a conter 1, 2 e 3, ainda mantendo a característica de categorização.

Assim como na classificação foram utilizados 70% da base para o treino e 30% para os testes. Porém para o problema de regressão foi aplicado apenas uma normalização utilizando o StandardScaler.

Com o random forest regressor, utilizando o MAE, foi apontado um erro de aproximadamente 3.15 anos, já o RMSE trouxe uma margem de erro de aproximadamente 4.45 anos. E como um todo, a aplicação dos dados neste modelo possui um erro médio de 13.36% segundo o cálculo do MAPE, o que consideramos relativamente alto.

Para o XGBoot, o valor do MAE foi de aproximadamente 2.98 anos, já o RMSE foi de aproximadamente 4.22 anos. E em geral o algoritmo apresentou um erro médio de aproximadamente 12.71% segundo o cálculo do MAPE.

É possível observar que os dois algoritmos possuem a mesma performance para esse problema apresentado, onde as diferenças entre eles são muito pequenas. Porém um ponto que chamou a atenção foi o quão elevado o erro médio foi, consideramos que um erro de mais de 10% é relativamente alto, porém é possível melhorar esse resultado, porém seria necessário a realização de mais análises preliminares para poder ajustar melhor a entrada de dados.

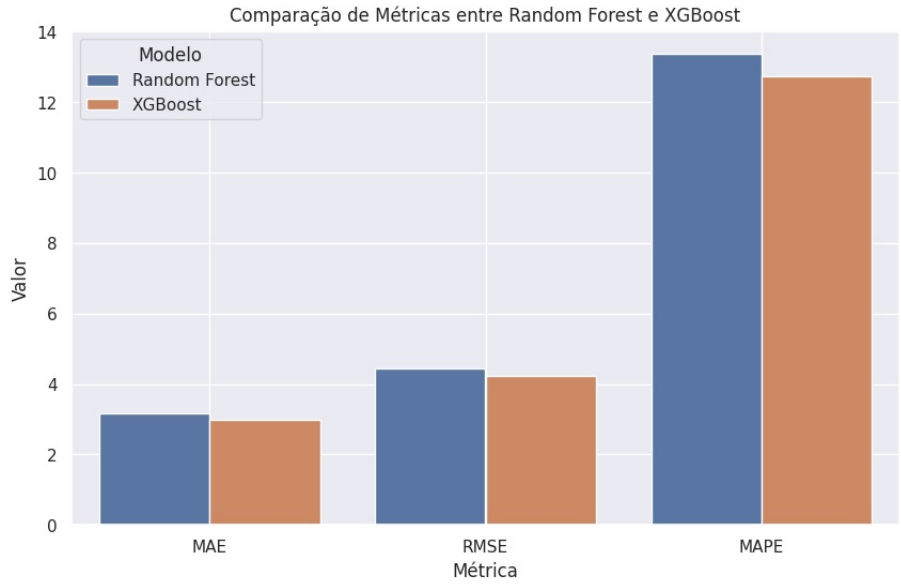

Figure 6. Grafico de barra msotrando a performace entre Random Forest e XGBoost

5. Métricas

Como mostrado na secção 4, sobre os modelos, foi decidido utilizar sete métricas para validar os modelos apresentados, quatro para validar os modelos de classificação: Precisão, Acurácia, Recall e F1-Score; e três métricas para validar os modelos de regressão: MAE, RMSE e MAPE.

6. Conclusões

Durante o desenvolvimento deste projeto, foram observados alguns pontos de melhoria, que infelizmente não puderam ser implementados. Realizar uma análise preliminar mais

robusta para poder buscar resultados melhores para os modelos desenvolvidos não pode ser realizada, logo muito do que será apontado nesta conclusão pode melhorar com uma pesquisa mais profunda e dedicada dos dados obtidos e da literatura de apoio.

Dito isso, primeiro observando a modelagem de classificação, ficou claro que com os dados obtidos e como eles foram aplicados, o grupo não conseguiu cumprir o objetivo de melhorar os valores de Precisão e Acurácia em comparação ao resultado obtidos por [Teixeira et al. 2023], novamente, talvez com uma entrada de dados mais elaborada seja possível atingir a meta mencionada. E observamos que os modelos obtidos não conseguem prever se um bebê irá testar positivo para sífilis congênita de forma eficaz, com uma precisão em torno de 60%, os resultados não são tão confiáveis.

Outro ponto a levantar é o desbalanceamento dos resultados de casos positivos e negativos, que precisam ser tratados para poder serem avaliados, talvez utilizando algum método não abordado neste projeto, a previsão dos casos possa ser mais assertiva.

Para a regressão, é possível notar o mesmo comportamento dos modelo como foi observado na classificação, ambos os modelos aplicados possuem performances muito parecidas, onde a decisão de escolha de qual utilizar não irá impactar muito nos resultados a serem obtidos. Mas é possível observar que o XGBoost Regressor possui resultados melhores que o Random Forest Regressor, porém essa diferença é relativamente pequena.

References

Teixeira, I. V., da Silva Leite, M. T., de Morais Melo, F. L., da Silva Rocha, É., Sadok, S., Pessoa da Costa Carrarine, A. S., Santana, M., Pinheiro Rodrigues, C., de Lima Oliveira, A. M., Vieira Gadelha, K., et al. (2023). Predicting congenital syphilis cases: A performance evaluation of different machine learning models. *PloS one*, 18(6):e0276150.