UNIVERSIDADE
DE PERNAMBUCO

Universidade de Pernambuco
Escola Politécnica de Pernambuco
Academic Graduate Program in Computer Engineering

Igor Vitor Teixeira

Predicting congenital syphilis cases: a performance evaluation of different machine learning models

Masters dissertation

Recife, October 2022

Universidade de Pernambuco
Escola Politécnica de Pernambuco
Academic Graduate Program in Computer Engineering

Igor Vitor Teixeira

Predicting congenital syphilis cases: a performance evaluation of different machine learning models

Masters dissertation

Dissertation presented to the Academic Graduate Program in COMPUTER ENGINEERING at the *Universidade de Pernambuco* as a partial requirement for obtaining the Master's degree in Computer Engineering.

Prof. PhD. Patricia Takako Endo
Supervisor

Recife, October 2022

Dissertação do Mestrado apresentada por **Igor Vitor Teixeira**, à Pós-Graduação em Engenharia de Computação da Escola Politécnica de Pernambuco da Universidade de Pernambuco, sob o título **"Predicting congenital syphilis cases: a performance evaluation of different machine learning models",** orientado pela Professora Patricia Takako Endo **-** Doutora, onde foi aprovado pela Banca Examinadora formada pelos professores:

Patricia Takako Endo – Doutora

(Orientadora - Primeira Examinadora)

Carmen Simone Grilo Diniz - Doutora

(Examinadora Externa)

Wellington Pinheiro dos Santos – Doutor
(Examinador Interno)

Visto e permitido a impressão.

Recife, 24 de Outubro de 2022.

gov.br Documento assinado digitalmente
CLEYTON MARIO DE OLIVEIRA RODRIGUES
Data: 16/02/2023 09:50:33-0300
Verifique em https://verificador.iti.br

Prof. Drº Cleyton Mário de Oliveira Rodrigues
Coordenador da Pós-Graduação em Engenharia de Computação
da Escola Politécnica de Pernambuco da Universidade de Pernambuco

Dados Internacionais de Catalogação-na-Publicação (CIP)
Universidade de Pernambuco – Recife

Teixeira, Igor Vitor
T266p Predicting congenital syphilis cases: a performance evaluation of different machine learning models. / Igor Vitor Teixeira. – Recife: UPE, Escola Politécnica, 2022.

58 f.: il.

Orientadora: Profa. PhD. Patrícia Takako Endo

Dissertação (Mestrado – Inteligência Computacional) Universidade de Pernambuco, Escola Politécnica, Programa de Pós-graduação em Engenharia da Computação, 2022.

1. Sífilis. 2. Sífilis Congênita. 3. Machine Learning. 4. Infecções Sexualmente Transmissíveis. I. Engenharia da Computação - Dissertação. II. Endo, Patricia Takako (orient.). III. Universidade de Pernambuco, Escola Politécnica, Mestrado em Engenharia da Computação. IV. Título

CDD: 006.31

I dedicate this work to God and to everyone who has helped me on this journey, especially my beloved wife, who has always motivated and supported me in the most difficult moments.

Acknowledgements

First, I thank God for giving me strength and wisdom to complete this journey. I would like to thank my parents for always investing in my studies, and for always motivating me to pursue my goals.

I thank my beloved wife, who has always been my safe haven on this journey, giving me the strength to continue and finish this master's degree.

I also thank all of the DotLAB Brazil research group, who have always helped me in the development of this work, especially Sebastião Rogério da Silva Neto, Thomás Tabosa de Oliveira, Morgana Thalita da Silva Leite, Flávio Leandro de Morais Melo and Elisson da Silva Rocha. And a special thanks to my supervisor, Prof. PhD. Patricia Takako Endo, who has always guided me with patience, partnership, and wisdom.

I would like to thank all *Mãe Coruja Pernambucana* Program (PMCP) health experts who provided access to the databases used in this work and for having acted as stackholders of this work, providing support during all stages of its development.

Finally, I would like to thank *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) for the AWS credits that made this work possible.

"All we have to decide is what to do with the time that is given us."

Gandalf

Abstract

Among the sexually transmitted infections (STIs), the incidence of syphilis has increased in the last 10 years, in spite of the presence of an effective and available treatment. The epidemiological implications are especially important during pregnancy since it can lead to complications related to prematurity, stillbirth, and miscarriage, in addition to congenital syphilis, characterized by multisystem involved in the newborn. In light of this situation, this work proposes a performance evaluation of different machine learning models to classify possible outcomes of congenital syphilis, using clinical and sociodemographic data from pregnant women that were assisted by a social program in Pernambuco, Brazil, named *Mãe Coruja Pernambucana* Program (PMCP), enabling better monitoring and care during gestation. Based on a rigorous methodology, we propose six experiments using three feature selection techniques to select the most relevant attributes, pre-process and clean the data, apply hyperparameter optimization to tune the machine learning models, and train and test models to have a fair evaluation and discussion. The AdaBoost-BODS-Expert model, an Adaptive Boosting (AdaBoost) model that used attributes selected by health experts, presented the best results in terms of evaluation metrics and acceptance by health experts from PMCP. This can give more confidence and allow adoption in daily usage to classify possible outcomes of congenital syphilis using clinical and sociodemographic data.

Key-words: syphilis, congenital syphilis, machine learning, sexually transmitted infections.

Resumo

Entre as infecções sexualmente transmissíveis (IST), a incidência de sífilis tem aumentado nos últimos 10 anos, apesar da presença de um tratamento eficaz e disponível. As implicações epidemiológicas são especialmente importantes durante a gravidez, pois pode levar a complicações relacionadas à prematuridade, natimorto e aborto espontâneo, além da sífilis congênita, caracterizada pelo envolvimento multissistêmico no recém-nascido. Diante dessa situação, este trabalho propõe uma avaliação de desempenho de diferentes modelos de *machine learning* para classificar possíveis desfechos da sífilis congênita, utilizando dados clínicos e sociodemográficos de gestantes atendidas por um programa social em Pernambuco, Brasil, denominado Programa Mãe Coruja Pernambucana (PMCP), possibilitando melhor acompanhamento e cuidado durante a gestação. Com base em uma metodologia rigorosa, propomos seis experimentos usando três técnicas de *feature selection* para selecionar os atributos mais relevantes, pré-processar e limpar os dados, aplicar otimização de hiperparâmetros para ajustar os modelos de *machine learning* e treinar e testar modelos para ter uma avaliação justa e discussão. O modelo AdaBoost-BODS-Expert, modelo *Adaptive Boosting* (AdaBoost) que utilizou atributos selecionados por especialistas em saúde, apresentou os melhores resultados em termos de métricas de avaliação e aceitação por especialistas em saúde do PMCP. Isso pode dar mais confiança e permitir a adoção no uso diário para classificar possíveis desfechos da sífilis congênita usando dados clínicos e sociodemográficos.

Key-words: sífilis, sífilis congênita, *machine learning*, infecções sexualmente transmissíveis.

Contents

List of Figures

List of Tables

List of abbreviations and acronyms

AdaBoost Adaptive Boosting

BDS Balanced Data Set

BODS Balanced with One-hot Encoding Data Set

BODDS Balanced with One-hot Encoding with Column Drop Data Set

CAAE Certificate of Presentation of Ethical Appreciation

CEP Research Ethics Committee

GBM Gradient Boosting Machines

ICD-10 International Classification of Diseases 10th Revision

IDS Imbalanced Data Set

IODS Imbalanced with One-hot Encoding Data Set

IODDS Imbalanced with One-hot Encoding with Column Drop Data Set

KNN K-Nearest Neighbors

OAS Organization of American States

PHC Primary Health Care

PMCP *Mãe Coruja Pernambucana* Program

SFA Sequential Feature Algorithm

SFS Sequential Forward Selection

SBS Sequential Backward Selection

SIH *Sistema de Informações Hospitalares*

SIM *Sistema de Informação Sobre Mortalidade*

SINAN *Sistema de Informação de Agravos de Notificação*

SIS-MC *Sistema de Informação do Mãe Coruja*

STIs Sexually Transmitted Infections

SUS *Sistema Único de Saúde*

SDG Sustainable Development Goal

SVM Support Vector Machine

TP True Positive

TN True Negative

FP False Positive

FN False Negative

UFPE *Universidade Federal de Pernambuco*

UN United Nations

VDRL Venereal Disease Research Laboratory

XGBoost eXtreme Gradient Boosting

Chapter 1

Introduction

Sexually Transmitted Infections (STIs) are considered a public health problem and are among the most common transmissible diseases [1], negatively affecting people's quality of life and health. Among them, syphilis is a systemic infection exclusively caused by the bacterium *Treponema pallidum*, transmitted in three ways: sexual, congenital, or via blood transfusion. Sexual transmission is predominant [2], followed by congenital, which is the result of the transmission of *Treponema pallidum* present in the bloodstream of the pregnant woman to the conceptus via the placenta or, occasionally, through direct contact with the syphilitic lesion at the time of delivery. Transmission via blood transfusion has become very rare due to the testing processes carried out by blood centers.

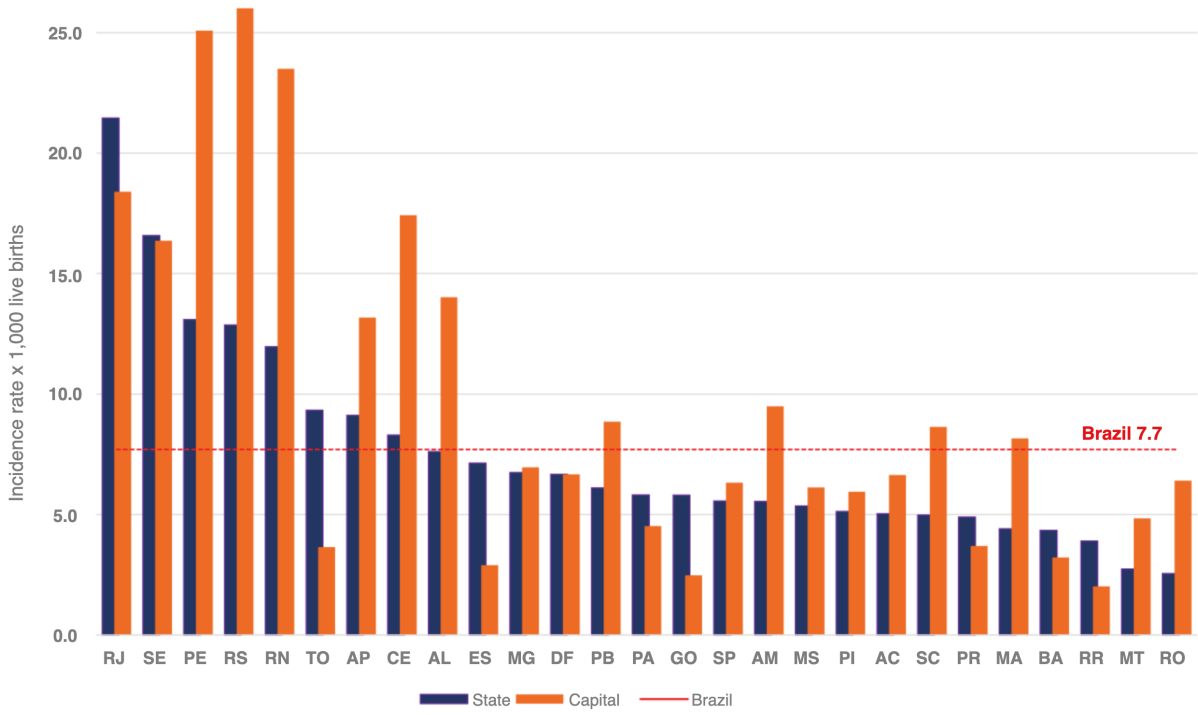
The adverse outcomes of untreated gestational syphilis are early pregnancy loss (40%), fetal death (11%), and preterm or low weight (12% to 13%). Furthermore, at least 20% of newborns have signs suggestive of early congenital syphilis [3]. In newborns, congenital syphilis can manifest either early or late. The main consequences of early congenital syphilis are preterm and low birth weight, in addition to various skin lesions, periostitis, radiographic abnormalities, limp pseudoparalysis, and respiratory distress. On the other hand, in late congenital syphilis, symptoms appear after the second year of birth and, in this case, owing to the longer development time of treponemas, more tissue and bone damage and cognitive losses may occur, such as saber blade tibia, Clutton's joints, Olympic forehead, saddle nose, deformed upper middle incisor teeth (Hutchinson's teeth), blackberry molars, short jaw, raised palatal arch, interstitial keratitis, sensory hearing loss, and learning difficulties [2].

Congenital syphilis must be precisely combated because of these severe consequences in newborns. In addition to being a condition with easy and accessible treatment, it is objectively eradicable. The increase in the number of tests, especially in Primary Health Care (PHC), after 2017, allowed a better understanding of the scenario of gestational and congenital syphilis in Brazil.

According to the Syphilis Epidemiological Bulletin 2021 of the Ministry of Health of Brazil [1], between 2011 and 2020, it was registered an unbridled growth of the incidence rate of

congenital syphilis, rising from 3.3 cases per 1,000 live births to 7.7 cases, reaching a peak in 2018 with the incidence rate of 9/1,000 live births. In 2020, 20,065 cases of congenital syphilis were recorded, and an infant mortality coefficient of 6.5/100,000 live births with 186 deaths due to congenital syphilis were reported in Brazil. In 2020, complying actively with the national epidemiological scenario, according to the bulletin [1], the state of Pernambuco presented an incidence rate of 13.1 cases of congenital syphilis in children under one year per 1,000 live births, thus the 3rd highest, a rate 1.7 times higher than that in Brazil (Figure 1). Based on the rate of mortality because of congenital syphilis in children under one year per 1,000 live births, Pernambuco appears in the 7th place, at a rate of 7.5 (Figure 2).

Figure 1 – Incidence rates of congenital syphilis (per 1,000 live births). Brazil, 2020.

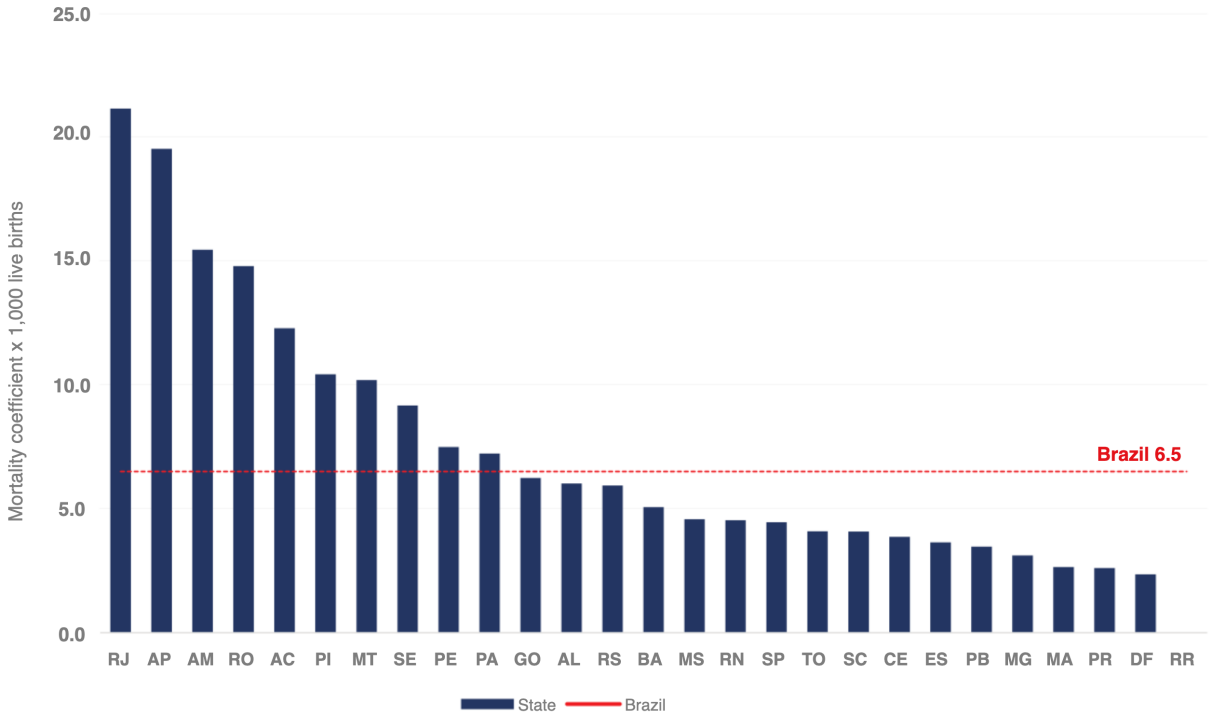


Source: *Sistema de Informação de Agravos de Notificação* (SINAN), updated on 06/30/2021.

The lack of interest in the treatment of syphilis by the pregnant woman's sex partners may be impacting the incidence rates of congenital syphilis, enabling the discontinuity of the syphilis chain of infection, and avoiding possible reinfections after a successful treatment. According to the Brazilian Protocol for Sexually Transmitted Infections [4], one-third of the sexual partners of people with recent syphilis will develop the infection within 30 days of exposure.

Despite the numbers presented and unlike many neonatal infections, congenital syphilis is a preventable condition, as long as the pregnant woman and her sex partners are identified (through effective antenatal care) and appropriate treatment is carried out. With simple and oriented interventions for the mothers, sex partners, and newborns, it is possible to obtain a great reduction in congenital syphilis cases [5].

Figure 2 – Infant mortality coefficient due to congenital syphilis (per 1,000 live births). Brazil, 2020.



Source: *Sistema de Informação Sobre Mortalidade* (SIM), updated on 12/31/2020.

In this context, it is necessary to create and update public policies in the area of maternal and child health to improve this scenario. There have already been several efforts by public administrators at different levels to reduce maternal and child deaths. Among these efforts, there are specific guidelines for assistance to pregnant women with syphilis (and other STIs) in PHC. An example is the *Mãe Coruja Pernambucana* Program (PMCP) [6], a Brazilian social program of reference in the maternal and child area, implemented in October 2007, recognized and awarded by the United Nations (UN) and the Organization of American States (OAS) as a public policy management model. The PMCP is a priority program of the government of Pernambuco, which aims to ensure comprehensive care for pregnant women using the *Sistema Único de Saúde* (SUS) and their children up to 5 years of age, creating a solidarity network to reduce maternal and child mortality, in addition to contributing to the improvement of social indicators. Currently, the PMCP is present in the 105 most vulnerable municipalities in the state of Pernambuco[1].

However, the scarcity of communication and information channels, and the existence of stigmatization by the society about STIs, especially in the context of vulnerable populations, weakens awareness and access to preventive measures and, consequently, to treatment for these diseases. According to Macedo et al. [5], "*among the sociodemographic factors, low education,*

[1] <https://maecoruja.pe.gov.br/cantos-mae-coruja/>

low income, and material status (stable or non-stable union) are identified as risk situations and an expression in which syphilis is related to poverty, although it is not limited to it". Additionally, although the rapid test for syphilis is available free of charge to the entire Brazilian population [7], its scarcity in PHC is common and can contribute to masking the real number of positive diagnoses and, consequently, to a decrease in the number of people who should be under treatment and follow-up [8]. According to Santos et al. [8], knowing the trends in syphilis and identifying the main factors related to PHC and the sociodemographic structure of a locality can guide new strategies for health promotion and disease prevention, as well as direct resources that significantly influence the reduction of a possible epidemic.

Due to the diversity of the profiles of the population attended by the PHC, such as the pregnant women assisted by the PMCP, the detection of the main clinical and sociodemographic factors related to congenital syphilis has proved to be a challenge. In addition, the scarcity of resources for public health in Brazil [9] has highlighted the need for innovative solutions, such as the use of machine learning techniques, which can assist in pattern recognition through the clinical and sociodemographic data of the pregnant women assisted by the PMCP, helping health professionals to make a decision for better monitoring of pregnant women. After being trained, the computational models have a low operating cost, thus facilitating the implementation and use of these resources.

The use of machine learning techniques to assist health professionals in monitoring and caring for pregnant women has a great social impact, contributing to reaching Sustainable Development Goal (SDG) 3 of the 2030 Agenda for Sustainable Development from the UN [10], which aims to ensure healthy lives and promote well-being for all at all ages, ending the epidemics of communicable diseases, such as syphilis.

Some works in the literature have proposed solutions using machine learning techniques to predict the incidence of syphilis cases in the population using data from Twitter [11], and profiling syphilis patients [12]. Differently from the current literature, this work aims to present a performance evaluation of different machine learning models to classify possible undesirable outcomes of congenital syphilis, using data from pregnant women assisted by the PMCP. The model uses clinical and sociodemographic data as input, enabling better monitoring and care during gestation.

1.1 Main goal

The objective of this work is to present a performance evaluation of different machine learning models to classify possible outcomes of congenital syphilis, using clinical and sociodemographic, enabling better monitoring and care during gestation.

1.1.1 Specific objectives

1. Conduct an exploratory analysis of data about pregnant women assisted by the PMCP;
2. Conduct a state-of-the-art survey on machine learning techniques for classifying congenital syphilis;
3. Database acquisition and pre-processing;
4. Implement and quantitatively evaluate machine learning models for the classification of congenital syphilis.

1.2 Dissertation organization

This work is presented in 6 chapters. Chapter 1 gives a brief overview of congenital syphilis and how it affects people, as well as the main and specific goals. Chapter 2 presents the related works found in the literature about the classification of congenital syphilis using machine learning. Chapter 3 presents the theoretical concepts that were used in this work. Chapter 4 presents the methodology used in the development and evaluation steps. Chapter 5 presents the results of the performance evaluation of the machine learning models. Lastly, chapter 6 concludes the work, presents contributions and future work.

Chapter 2

Related Work

This chapter presents related works found in the literature on machine learning models related to congenital syphilis case classification. But, the majority of the studies found in the literature concentrated on the incidence and risk factors for syphilis.

Some works in the literature have already presented studies on the incidence and the risk factors for syphilis, such as Santos et al. [8] that presented a study to understand the factors related to the trends of syphilis in Brazil, analyzing their association with sociodemographic aspects and the PHC in the period between 2011 and 2019. Authors concluded that there are some important predictors of the upward trends of acquired syphilis, such as the quality of the PHC service, the availability of penicillin in the PHC, the availability of the female condom, and the influence of population size. The need for training of the health professionals in the care of STI is also reinforced as an indispensable action to control syphilis cases.

Lima et al. [13] described the incidence of congenital syphilis in the city of Belo Horizonte, capital of Minas Gerais a Brazilian state, between the years 2001 and 2008, aiming to identify the risk factors associated with the diagnosis of the disease. A multivariate logistic regression analysis was performed to identify maternal and antenatal characteristics independently associated with the occurrence of congenital syphilis. It was identified as an independent risk factor for congenital syphilis characteristics such as black-skinned mothers, maternal educational level, and the absence of antenatal care, which is the main risk factor, presenting an 11 times greater chance of congenital syphilis than pregnancies that the mother attended at least one time to the antenatal follow-up.

Melo et al. [14] analyzed the association between morbidity from congenital syphilis and biological, socioeconomic, and antenatal care indicators using data from the city of Recife, capital of the state of Pernambuco, between 2004 and 2006. According to their results, factors such as fewer than four antenatal appointments, mothers under 20 years of age, and black and mixed skin color were associated with cases of congenital syphilis. The increase in risk was proportional to the deterioration of socioeconomic and biological indicators and antenatal care, and the worst situation occurred in more peripheral areas of the cities.

Young et al. [11] conducted a study to explore whether data from a social network (Twitter) could be used to identify trends in cases of syphilis, primary and secondary, and late latent syphilis in the United States of America. They sought to identify associations between tweets related to sexual risks, that would possibly be reported in the following year. For the years 2012 and 2013, weekly disease data from counties and the capital were collected. According to the results, counties and capitals with the highest number of risk-related tweets in 2012 were associated with a 2.7% increase in primary and secondary syphilis cases and a 3.6% increase in latent syphilis cases in 2013. Their results were consistent in all models run in the analysis, suggesting a relationship between syphilis and risk-related tweets.

Silva [12] developed a research close to the objective of our work and, therefore, could be used for later comparison. A predictive analysis was performed to create a profile of risk groups using historical data generated by systems used by SUS, SIM, the Brazilian death registration system, SINAN, and *Sistema de Informações Hospitalares* (SIH), the Brazilian hospital admissions system. The data related to cases of acquired, gestational, and congenital syphilis in Brazil between 2010 and 2019 were used. Due to difficulties in linking all the syphilis-related attributes available in the databases, only data related to sex, color, educational level, and age group were selected. After applying clustering techniques, 130 profiles were found, obtaining an accuracy of 97.41%.

2.1 Considerations

Our work differs from others in the current literature by conducting and discussing a performance evaluation of machine learning models using clinical and sociodemographic data presented in an integrated system that records data from antenatal care, birth, and child's development monitoring. This data allows us to focus on the classification of possible cases of congenital syphilis in a full range of relationships and evaluate which machine learning models are more efficient in our scenario.

Chapter 3

Theoretical Concepts

This chapter presents the fundamental concepts utilized in this work. Section 3.1 describes the seven machine learning algorithms used in this work. The techniques for data balancing and encoding the data set are described in section 3.2. The techniques used to optimize the machine learning models are presented in section 3.3. Section 3.4 concludes with the metrics used to evaluate the machine learning models.

3.1 Machine Learning techniques

In this work, we evaluate the following machine learning models for the classification of congenital syphilis cases: Decision Tree, Random Forest, Adaptive Boosting (AdaBoost), Gradient Boosting Machines (GBM), eXtreme Gradient Boosting (XGBoost), K-Nearest Neighbors (KNN), and Support Vector Machine (SVM).

A Decision Tree [15] [16] is a supervised machine learning algorithm that may be used to classify categorical values (classification tree) or predict numerical values (regression tree) by making a statement followed by a decision depending on whether the statement is true or false. Basically, a Decision Tree consists of a root node, internal nodes, and leaf nodes, which are created by gradually splitting the data based on the target attribute. A root node is identified by the attribute that best splits the data, focusing on the target attribute. In other words, the attribute with the lowest impurity value. Then, to reduce the impurity, the root node is split, creating the internal nodes. The procedure of splitting the nodes is performed until it is no longer possible to reduce the impurity value, resulting in the formation of leaf nodes, which indicate the outputs of the Decision Tree.

Random Forest [17] consists of an ensemble of Decision Trees, each of which is trained by a subset of random attributes extracted from the training data set, a technique known as bootstrap. The results of each Decision Tree are then aggregated according to the classification of the class, and the class with the most votes is chosen by majority vote.

The AdaBoost is based on the idea of boosting, in machine learning this means creating

a highly accurate prediction by combining many weak predictors [18]. In the algorithm, given the training data set, a distribution is calculated over the examples using a weak learner with the objective of finding a classifier with a low error relative to the distribution. This process is repeated n times, and the final classifier is a weighted combination of the classifications of weak learners [19].

GBM [20] does a sequential procedure to predict values based on previous errors. Gradient refers to the error gained after building a model, and boosting relates to improvement. In the GBM algorithm, the predictions are started by a simple Decision Tree. The residual is calculated by subtracting the actual value from the predicted value. Another shallow Decision Tree is built that predicts residuals based on all the independent values. Then, the original prediction is updated with the new prediction multiplied by the learning rate. The last three steps are repeated for a certain number of iterations, known as the number of trees.

As its core, XGBoost has a Decision Tree boosting algorithm that attempts to correct or minimize the errors in the previous model. The XGBoost algorithm is based on a generalized GBM, but it also includes a regularization term to prevent overfitting and supports arbitrary differentiable loss functions. [21].

KNN is an algorithm that selects the k closest samples, or k nearest neighbors, in the training data set based on a distance metric such as Euclidean distance, and then predicts the class based on the major class from those samples in the k nearest neighbors [22].

SVM aims to find linearity between data by using hyperplanes, even if the data is not linearly separable [23]. For that, SVM uses the kernel concept to map the scenario data initially presented by the variables to a high-dimensional space to enhance separation between classes [24]. Due to specific aspects of understanding how to manipulate non-linear data, SVM has become a very popular model in the healthcare scenario [25].

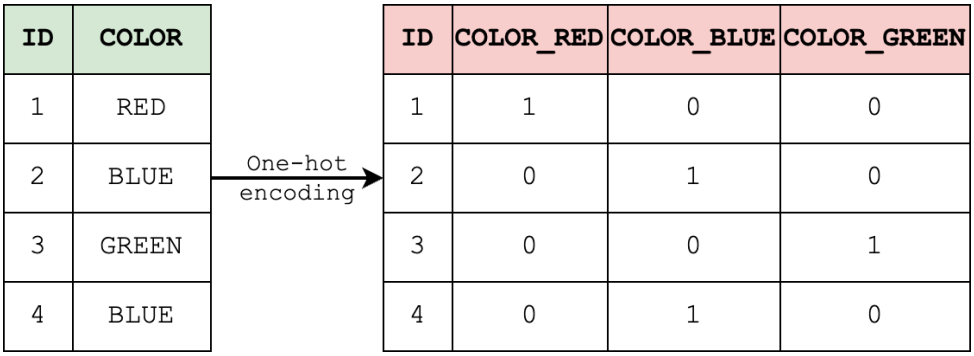
3.2 Data balancing and encoding techniques

The purpose of data balancing is to equalize the amount of data of the target attribute whose positive and negative classes are out of balance. According to [26] and [27], data imbalance is one of the obstacles that hinder the learning of classification algorithms, as it can lead to a learning bias in which the model learns more about the majority class than the minority, resulting in models with low performance due to the disparity between classes. One of the ways to get around the problem is the random undersampling technique, which [28] presents as a heuristic method that randomly eliminates instances of the majority class until the quantity is balanced with the minority class.

Categorical classes are typically encoded to binary or numeric values using computational coding techniques prior to training machine learning models. According to [29], one of the most

widely used encoding techniques used in the literature is the one-hot encoding technique, which seeks to transform each categorical attribute into new attributes with binary values where a value of one indicates the presence and a value of zero indicates the absence of the coded categorical value specific for the new attribute [30]. Figure 3 illustrates how one-hot encoding technique is applied to a data set with two attributes. In this example, the technique was only applied to the COLOR attribute, which has split into three new attributes, one for each of the original attribute's possible categories.

Figure 3 – One-hot encoding technique example.



Source: Produced by the author.

3.3 Hyperparameter optimization and feature selection techniques

Machine learning models might have several parameters, also known as hyperparameters, to configure, making manual configuration impractical. Automated hyperparameter optimization techniques, such as Grid Search, were used in this work to determine the best hyperparameter combination for each model. The Grid Search technique [31] performs an exhaustive search for training and evaluating models with all the combinations of hyperparameters in a given search space, returning the hyperparameters that achieve the best performance.

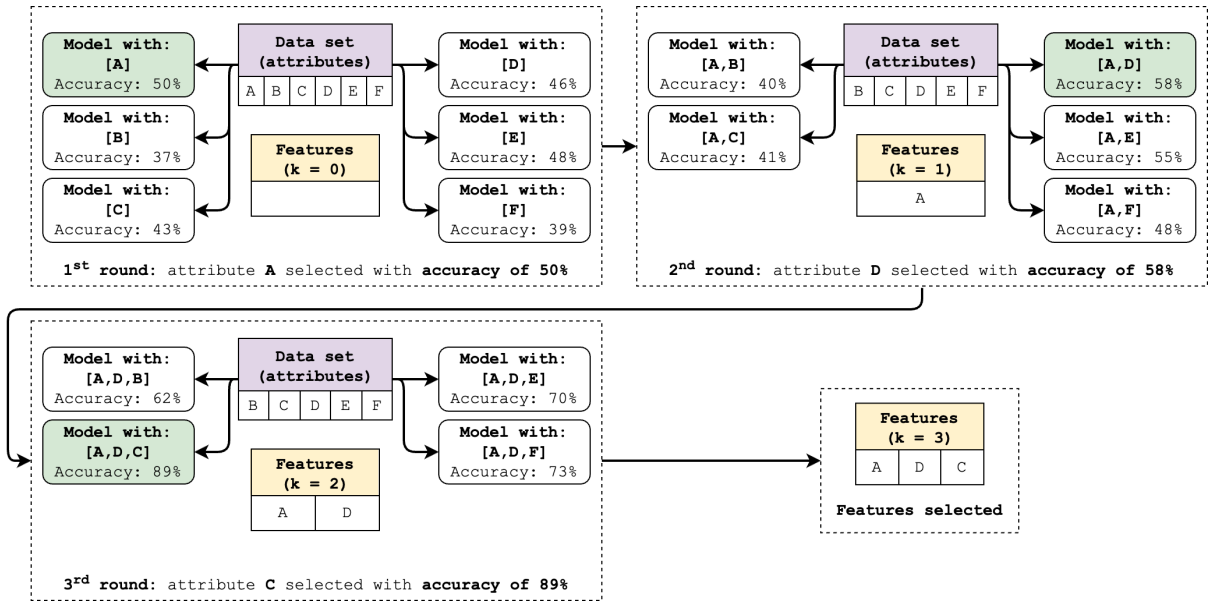
As is the case with this work, data sets with numerous attributes (high dimensionality) may result in the preponderance of noisy, irrelevant, and redundant data [32], thus impacting the machine learning models' learning. To deal with it, feature selection techniques could be used to reduce the dimensionality of the data set by selecting a relevant subset of attributes from the original data set. These techniques can be categorized into three approaches: filter, wrapper, and embedded [33]. In this work, we used a wrapper approach.

Computationally more expensive, the wrapper approach uses a machine learning model to select and evaluate the subset of attributes with the best performance, running until a stop condition is satisfied [32]. For this work, we set the SFA stop condition equal to the number of

attributes of the data set, which forces the technique to evaluate all possible subsets of attributes and select the one with the highest performance. We used the SFA technique to implement the wrapper approach, focusing on Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS), which are its two main types [34]. In order to better comprehend how SFS and SBS techniques work, an example data set containing six attributes was used to select a subset of three attributes ($k = 3$).

The SFS technique starts with an empty subset of features, and with each iteration, a new attribute is added, thereby selecting the attributes that increase the model's performance. Figure 4 illustrates the execution of the SFS technique. In the first round, all attributes were individually evaluated by four different machine learning models, where the model with attribute A outperformed, selecting the attribute A to compose the subset of selected features. In the second round, three models were evaluated with attribute A and one of the remaining attributes. The subset of attributes A and C had the best performance, so attribute C was added to the subset of selected features. In the third and final round, the best results came from the subset of attributes A, C, and B. Attribute B was chosen for the subset of selected attributes, and the desired number of features was reached.

Figure 4 – Example of execution of the SFS technique.

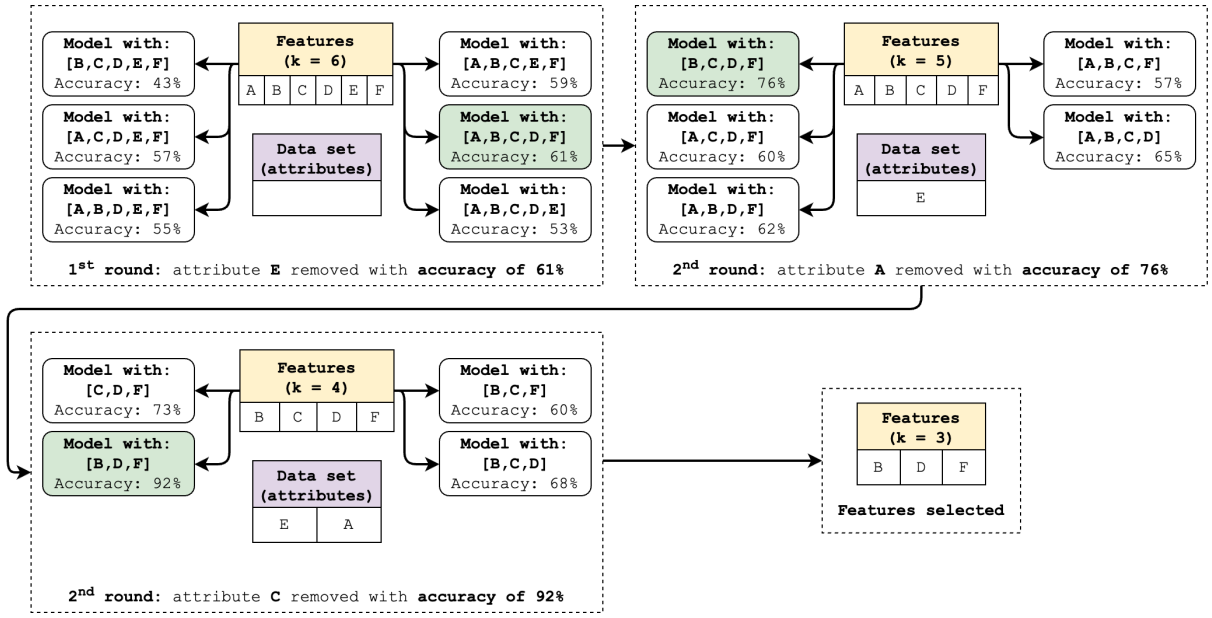


Source: Adapted from [35]

On the other hand, the SBS technique starts with a subset of features that includes all of the data set's attributes. With each iteration, one attribute is removed, and different combinations of attributes are evaluated. The combinations of attributes which generate the best model's performance is then chosen. Figure 5 illustrates the execution of the SBS technique. In the first round, the combination of attributes B, C, D, E, and F generated a model with the best

performance, so the attribute E was removed. In the second and third rounds, the procedure is repeated, but this time the attributes A and C are removed. Finally, the subsets B, D, and F are chosen.

Figure 5 – Example of execution of the SBS technique.



Source: Adapted from [35]

3.4 Evaluation metrics

In this work, we evaluated the proposed models using the accuracy, precision, sensitivity, specificity, and F1-Score metrics. All these evaluation metrics are based on the Confusion Matrix [36], which seeks to calculate:

- True Positive (TP) for classified as positive and really positive.
- False Positive (FP) for classified as positive but actually negative.
- True Negative (TN) for classified as negative and really negative.
- False Negative (FN) for classified as negative but actually positive.

Accuracy [37] measures the model's performance according to the total samples correctly classified, indicating how frequently the model was correct, and is defined as

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3.1}$$

Precision is a metric used to determine the proportion of positive classifications that are actually positive in reality [38], calculated as

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{3.2}$$

Sensitivity determines the proportion of real positives that were correctly classified [39], defined as

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3.3}$$

Specificity is the inverse of sensitivity, which seeks to determine the proportion of real negatives that were correctly classified [39], calculated as

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{3.4}$$

F1-Score [40] is a metric that calculates the harmonic mean of precision and sensitivity to summarize the predictive performance of the models, and is defined as

$$\text{F1} - \text{Score} = 2 \times \frac{precision \times sensitivity}{precision + sensitivity} \tag{3.5}$$

In the following chapter, we will describe the characteristics of the data sets used in this work, which demonstrate a significant data imbalance between the positive and negative classes of congenital syphilis. Due to this disparity, we chose the F1-Score metric as the primary metric for some evaluation steps of the models proposed, as it provides a harmonic mean between two extremely relevant metrics for health problems. We will be able to evaluate the proportion of positive classifications that were correctly classified (precision) and the proportion of congenital syphilis cases that were correctly classified by the proposed models (sensitivity).

Chapter 4

Materials and Methods

This chapter presents the methodology used in this work as well as the data set used. Section 4.1 presents the pre-processing steps used in the data set used in this work. Section 4.2 presents the methodology applied in the proposed machine learning experiments. Section 4.3 presents the technical resources used to perform all methodology.

4.1 Data set and pre-processing

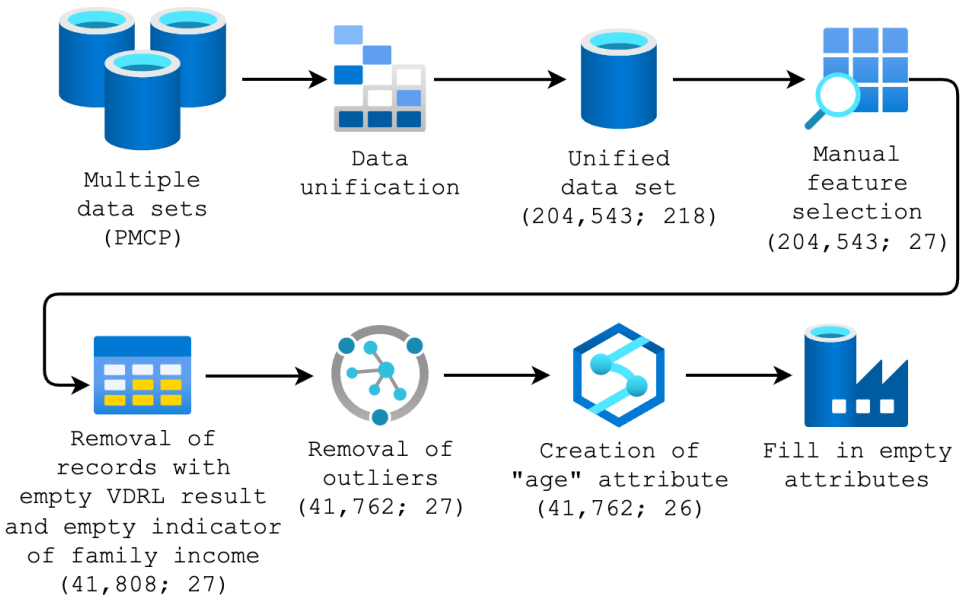
We used anonymized data sets provided by the PMCP, extracted from their information system, named *Sistema de Informação do Mãe Coruja* (SIS-MC). These data sets contain all the data from the SIS-MC from the cities served by the PMCP in the State of Pernambuco, Brazil, between the years of 2013 and 2021. The data set is publicly available at <https://data.mendeley.com/datasets/3zkcvybvkz/1>. The use of data from the SIS-MC was authorized by the Research Ethics Committee (CEP) of the *Universidade Federal de Pernambuco* (UFPE) with the Certificate of Presentation of Ethical Appreciation (CAAE) number 12438019.2.0000.5208 and authorized by the partner institution, the PMCP.

Figure 6 illustrates the data pre-processing methodology designed to unify the data sets provided by the PMCP and to perform manual feature selection, handle missing data, remove outliers, and create new attributes.

A unification step was applied as the data sets provided by the PMCP were shared across different data structures. At first, data sets were chosen that included any clinical and sociodemographic data regarding antenatal care, pregnant women's outcomes, and their children. Figure 7 presets the selected data sets. Since the goal of this work is to predict undesirable outcomes of congenital syphilis, the children's data set was chosen as the starting point for the unification. This is because it includes the result of the Venereal Disease Research Laboratory (VDRL) test, which is a screening test for congenital syphilis at birth and was chosen as the target attribute of the classification.

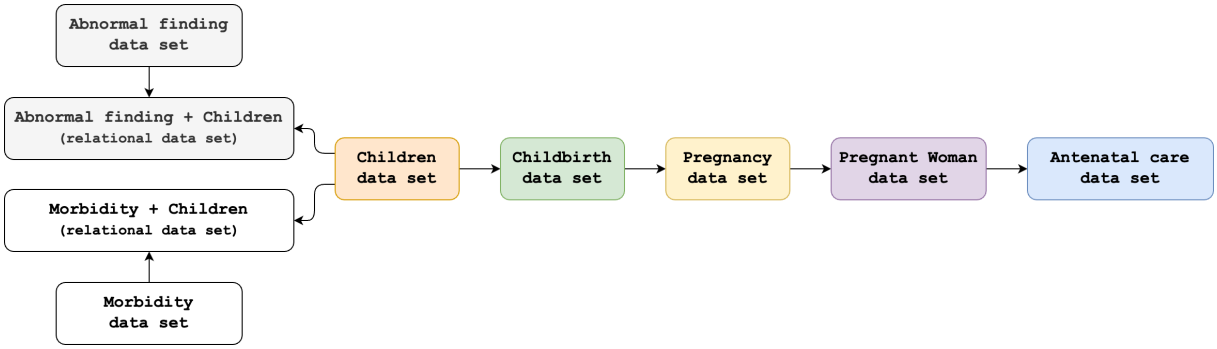
The target attribute had 807 positive cases of congenital syphilis, 40,995 negative cases,

Figure 6 – Data pre-processing methodology.



Source: Produced by the author.

Figure 7 – Original data sets mapping.



Source: Produced by the author.

and 162,741 empty records. We used the data sets related to abnormal findings and morbidities related to children's data to find other possible cases of congenital syphilis not recorded by the target attribute. The abnormal findings and morbidities selected are presented in Table 1 with their respective International Classification of Diseases 10th Revision (ICD-10). Therefore, only 17 records with empty VDRL test result and 3 negative cases were changed to positive. This resulted in a target attribute with 827 cases of congenital syphilis positive, 40,992 cases negative, and 162,724 empty records, which were removed from the data set.

The children's data set was unified with the data sets related to childbirth, pregnancies, pregnant women, and antenatal care. At the final stage of data unification, the unified data set contained 204,543 records and 218 attributes.

Table 1 – Abnormal findings and morbities related to syphilis.

Data set	**ICD-10**	**Description**
Abnormal findings	O98	Maternal infectious and parasitic diseases classifiable elsewhere but complicating pregnancy, childbirth, and the puerperium
Morbidity	A50.0	Early congenital syphilis, symptomatic
Morbidity	A50.1	Early congenital syphilis, latent
Morbidity	A50.2	Early congenital syphilis, unspecified
Morbidity	A50.4	Late congenital neurosyphilis [juvenile neurosyphilis]
Morbidity	A50.5	Other late congenital syphilis, symptomatic
Morbidity	A50.6	Late congenital syphilis, latent
Morbidity	A50.7	Late congenital syphilis, unspecified
Morbidity	A50.9	Congenital syphilis, unspecified
Morbidity	A51.0	Primary genital syphilis
Morbidity	A51.1	Primary anal syphilis
Morbidity	A51.2	Primary syphilis of other sites
Morbidity	A51.3	Secondary syphilis of skin and mucous membranes
Morbidity	A51.4	Other secondary syphilis
Morbidity	A51.5	Early syphilis, latent
Morbidity	A51.9	Early syphilis, unspecified
Morbidity	A52.0	Cardiovascular and cerebrovascular syphilis
Morbidity	A52.1	Symptomatic neurosyphilis
Morbidity	A52.2	Asymptomatic neurosyphilis
Morbidity	A52.3	Neurosyphilis, unspecified
Morbidity	A52.7	Other symptomatic late syphilis
Morbidity	A52.8	Late syphilis, latent
Morbidity	A52.9	Late syphilis, unspecified
Morbidity	A53.0	Latent syphilis, unspecified as early or late
Morbidity	A53.9	Syphilis, unspecified
Morbidity	A65	Nonvenereal syphilis

Source: Produced by the author.

We conducted a manual feature selection in order to select relevant clinical and sociodemographic attributes related to pregnancies, pregnancy outcomes, and pregnant women, and reduce the data dimensionality. To clean up the data set, we removed attributes with more than 70% missing data, except for the target attribute, reducing the dimensionality to 27 attributes.

We removed 11 records with an empty value for the family income indicator, as informed in the pregnant woman's record in SIS-MC, which resulted in a data set with 41,808 records. These records were related to negative cases of congenital syphilis, and as the data set has numerous negative cases, especially when compared to the number of positive cases, they were removed without negatively impacting on data quality.

After analyzing the baseline characteristics of the data set with the help of health experts (stakeholders) from the PMCP, some records considered as outliers were also removed, referring

to *(i)* pregnant women who were born before 1960 and after 2020; *(ii)* family income informed at antenatal care greater than 20,000; and *(iii)* number of residents in the household greater than 20.

The age of the pregnant women when they were attended by the PMCP attribute was created, and to calculate it, we subtracted the date of registration of the pregnancy in the SIS-MC from the date of birth of the pregnant woman. After that, these two attributes (date of registration and date of birth) were removed from the data set. We also converted attributes from numerical to categorical, such as the number of abortions, number of living children, number of pregnancies, and number of residents in the household. At the end, the data set was left with 41,762 records and 26 attributes.

At last, we created a new category to fill in the empty data. We use this strategy because all our attributes were binary or categorical (except for the age attribute), as can be seen in Table 2, which presents the data set attributes after all modifications.

Table 2 – Data set attributes.

Attribute	**Description**	**Type**	**Categorization**
VDRL_RESULT▲	VDRL result	Binary	(i) Positive and (ii) Negative
CONS_ALCOHOL	Consume alcohol	Categorical	(i) Positive, (ii) Negative, and (iii) Not informed
RH_FACTOR	RH factor	Categorical	(i) Positive, (ii) Negative, and (iii) Not informed
SMOKER	Smoker	Categorical	(i) Positive, (ii) Negative, and (iii) Not informed
PLAN_PREGNANCY★	Planned pregnancy	Categorical	(i) Positive, (ii) Negative, and (iii) Not informed
BLOOD_GROUP	Blood group	Categorical	(i) Positive, (ii) Negative, and (iii) Not informed
HAS_PREG_RISK★	Has pregnancy risk	Categorical	(i) O, (ii) A, (iii) B, (iv) AB, and (v) Not informed

Continued on next page

Table 2 - Continued from previous page

Attribute	Description	Type	Categorization
TET_VACCINE	Tetanus vaccination	Categorical	(i) Positive, (ii) Negative, and (iii) Not informed
IS_HEAD_FAMILY	Is head of family	Categorical	(i) Positive, (ii) Negative, and (iii) Not informed
MARITAL_STATUS★	Marital status	Categorical	(i) Single, (ii) Married, (iii) Widowed, (iv) Judicial separation, (v) Divorced, and (vi) Not informed
FOOD_INSECURITY★	Food insecurity	Categorical	(i) Positive, (ii) Negative, and (iii) Not informed
NUM_ABORTIONS★	Number of abortions	Categorical	(i) None, (ii) One, (iii) Two, (iv) More than two, and (v) Not informed
NUM_LIV_CHILDREN★	Number of living children	Categorical	(i) None, (ii) One, (iii) Two, (iv) More than two, and (v) Not informed
NUM_PREGNANCIES★	Number of pregnancies	Categorical	(i) None, (ii) One, (iii) Two, (iv) More than two, and (v) Not informed
FAM_PLANNING★	Received information about family planning	Categorical	(i) Positive, (ii) Negative, and (iii) Not informed

Continued on next page

Table 2 - Continued from previous page

Attribute	Description	Type	Categorization
TYPE_HOUSE	Type of house construction	Categorical	(i) Straw, (ii) Wood, iii) Clay, (iv) Plaster, (v) Masonry, and (vi) Not informed
HAS_FAM_INCOME	Has family income	Binary	(i) Positive and (ii) Negative
EDUC_LEVEL★	Educational level	Categorical	(i) Complete elementary school, (ii) Incomplete elementary school, (iii) Complete middle school, (iv) Incomplete middle school, (v) Complete high school, (vi) Incomplete high school, (vii) Complete superior school, (viii) Incomplete superior school, and (ix) Not informed
CONN_SEWER_NET	House connected to the sewer network	Categorical	(i) Positive, (ii) Negative, and (iii) Not informed
NUM_RES_HOUSEHOLD	Number of residents in the household	Categorical	(i) None, (ii) One, (iii) Two, (iv) Three, (v) More than three, and (vi) Not informed
HAS_FRU_TREE	Has fruit trees	Categorical	(i) Positive, (ii) Negative, and (iii) Not informed
HAS_VEG_GARDEN	Has a vegetable garden	Categorical	(i) Positive, (ii) Negative, and (iii) Not informed
FAM_INCOME★	Family income informed at antenatal cares	Categorical	(i) Less than 500, (ii) Between 501 and 1000, (iii) More than 1000, and (iv) Not informed

Continued on next page

Table 2 - Continued from previous page

Attribute	Description	Type	Categorization
HOUSING_STATUS	Housing status	Categorical	(i) Owned, (ii) Rented, (iii) Donated, and (iv) Not informed
WATER_TREATMENT	Type of water treatment used	Categorical	(i) Filtered, (ii) Boiled, (iii) Disinfected (chlorine), (iv) None, and (v) Not informed
AGE★	Age of the pregnant woman	Numerical	–

▲ Target attribute for classification; ★ Attribute manually selected by health experts from PMCP.

Source: Produced by the author.

After all pre-processing steps, the pre-processed data set contained 826 positive cases and 40,936 negative cases of congenital syphilis. The overall baseline characteristics of the pre-processed data set related to the pregnant women assisted by the PMCP are presented in Table 3.

Table 3 – Baseline characteristics of the data set.

Variables	Total	Positive	Negative
Total: n (%)	41,762	826	40,936
Consume alcohol: n (%)			
Yes	1,263 (3.0)	60 (7.3)	1,203 (2.9)
No	36,359 (87.1)	684 (82.8)	35,675 (87.1)
Missing	4,140 (9.9)	82 (9.9)	4,058 (9.9)
RH Factor: n (%)			
Positive	25,761 (61.7)	448 (54.2)	25,313 (61.8)
Negative	2,151 (5.2)	37 (4.5)	2,114 (5.2)
Missing	13,850 (33.2)	341 (41.3)	13,509 (33.0)
Smoker: n (%)			
Yes	1,479 (3.5)	78 (9.4)	1,401 (3.4)
No	37,105 (88.8)	684 (82.8)	36,421 (89.0)
Missing	3,178 (7.6)	64 (7.7)	3,114 (7.6)
Planned pregnancy: n (%)			
Yes	16,772 (40.2)	274 (33.2)	16,498 (40.3)

Continued on next page

Table 3 - Continued from previous page

Variables	Total	Positive	Negative
No	22,889 (54.8)	521 (63.1)	22,368 (54.6)
Missing	2,101 (5.0)	31 (3.8)	2,070 (5.1)
Blood group: n (%)			
O	13,100 (31.4)	234 (28.3)	12,866 (31.4)
A	10,350 (24.8)	169 (20.5)	10,181 (24.9)
B	3,457 (8.3)	56 (6.8)	3,401 (8.3)
AB	1,075 (2.6)	28 (3.4)	1,047 (2.6)
Missing	13,780 (33.0)	339 (41.0)	13,441 (32.8)
Has pregnancy risk: n (%)			
Yes	5,406 (12.9)	145 (17.6)	5,261 (12.9)
No	34,362 (82.3)	636 (77.0)	33,726 (82.4)
Missing	1,994 (4.8)	45 (5.4)	1,949 (4.8)
Got tetanus vaccine: n (%)			
Yes	36,726 (87.9)	743 (90.0)	35,983 (87.9)
No	3,185 (7.6)	69 (8.4)	3,116 (7.6)
Missing	1,851 (4.4)	14 (1.7)	1,837 (4.5)
Head of family: n (%)			
Yes	6,074 (14.5)	154 (18.6)	5,920 (14.5)
No	31,274 (74.9)	581 (70.3)	30,693 (75.0)
Missing	4,414 (10.6)	91 (11.0)	4,323 (10.6)
Marital status: n (%)			
Single	13,658 (32.7)	277 (33.5)	13,381 (32.7)
Married	10,102 (24.2)	115 (13.9)	9,987 (24.4)
Widowed	75 (0.2)	1 (0.1)	74 (0.2)
Judicial separation	93 (0.2)	4 (0.5)	89 (0.2)
Divorced	280 (0.7)	8 (1.0)	272 (0.7)
Other	17,554 (42.0)	421 (51.0)	17,133 (41.9)
Food insecurity: n (%)			
Yes	7,129 (17.1)	135 (16.3)	6,994 (17.1)
No	16,487 (39.5)	234 (28.3)	16,253 (39.7)
Missing	18,146 (43.5)	457 (55.3)	17,689 (43.2)
Number of abortions: n (%)			
None	13,421 (32.1)	231 (28.0)	13,190 (32.2)
One	4,962 (11.9)	126 (15.3)	4,836 (11.8)
More than one	1,333 (3.2)	28 (3.4)	1,305 (3.2)

Continued on next page

Table 3 - Continued from previous page

Variables	Total	Positive	Negative
Missing	22,046 (52.8)	441 (53.4)	21,605 (52.8)
Number of living children: n (%)			
None	4,559 (10.9)	80 (9.7)	4,479 (10.9)
One	9,685 (23.2)	187 (22.6)	9,498 (23.2)
Two	4,817 (11.5)	113 (13.7)	4,704 (11.5)
More than two	3,583 (8.6)	87 (10.5)	3,496 (8.5)
Missing	19,118 (45.8)	359 (43.5)	18,759 (45.8)
Number of pregnancies: n (%)			
None	4,477 (10.7)	86 (10.4)	4,391 (10.7)
One	11,533 (27.6)	196 (23.7)	11,337 (27.7)
Two	8,135 (19.5)	173 (20.9)	7,962 (19.4)
More than two	9,046 (21.7)	201 (24.3)	8,845 (21.6)
Missing	8,571 (20.5)	170 (20.6)	8,401 (20.5)
Received information about family planning: n (%)			
Yes	19,831 (47.5)	338 (40.9)	19,493 (47.6)
No	12,005 (28.7)	265 (32.1)	11,740 (28.7)
Missing	9,926 (23.8)	223 (27.0)	9,703 (23.7)
House construction: n (%)			
Straw	150 (0.4)	2 (0.2)	148 (0.4)
Wood	137 (0.3)	6 (0.7)	131 (0.3)
Clay	699 (1.7)	14 (1.7)	685 (1.7)
Plaster	70 (0.2)	5 (0.6)	65 (0.2)
Masonry	37,929 (90.8)	738 (89.3)	37,191 (90.9)
Missing	2,777 (6.6)	61 (7.4)	2,716 (6.6)
Has family income: n (%)			
Yes	31,261 (74.9)	609 (73.7)	30,652 (74.9)
No	10,501 (25.1)	217 (26.3)	10,284 (25.1)
Level of schooling: n (%)			
Illiterate	527 (1.3)	19 (2.3)	508 (1.2)
Complete elementary school	1,470 (3.5)	35 (4.2)	1,435 (3.5)
Incomplete elementary school	5,131 (12.3)	144 (17.4)	4,987 (12.2)
Complete middle school	2,521 (6.0)	49 (5.9)	2,472 (6.0)
Incomplete middle school	9,218 (22.1)	214 (25.9)	9,004 (22.0)
Complete high school	13,166 (31.5)	185 (22.4)	12,981 (31.7)
Incomplete high school	6,777 (16.2)	150 (18.2)	6,627 (16.2)

Continued on next page

Table 3 - Continued from previous page

Variables	Total	Positive	Negative
Complete superior school	1,063 (2.5)	7 (0.8)	1,056 (2.6)
Incomplete superior school	847 (2.0)	7 (0.8)	840 (2.1)
Missing	1,042 (2.5)	16 (1.9)	1,026 (2.5)
House connected to the sewer network: n (%)			
Yes	23,608 (56.5)	477 (57.7)	23,131 (56.5)
No	15,234 (36.5)	276 (33.4)	14,958 (36.5)
Missing	2,920 (7.0)	73 (8.8)	2,847 (7.0)
Number of residents in the household: n (%)			
None	3 (0.0)	-	3 (0.0)
One	415 (1.0)	12 (1.5)	403 (1.0)
Two	10,993 (26.3)	206 (24.9)	10,787 (26.4)
Three	11,105 (26.6)	207 (25.1)	10,898 (26.6)
More than three	15,404 (36.9)	330 (40.0)	15,074 (36.8)
Missing	3,842 (9.2)	71 (8.6)	3,771 (9.2)
Has fruit trees: n (%)			
Yes	7,545 (18.1)	119 (14.4)	7,426 (18.1)
No	27,282 (65.3)	555 (67.2)	26,727 (65.3)
Missing	6,935 (16.6)	152 (18.4)	6,783 (16.6)
Has a vegetable garden: n (%)			
Yes	3,716 (8.9)	49 (5.9)	3,667 (9.0)
No	31,323 (75.0)	626 (75.8)	30,697 (75.0)
Missing	6,723 (16.1)	151 (18.3)	6,572 (16.1)
Family income: n (%)			
Less than or equal to R$ 500.00	16,575 (39.7)	331 (40.1)	16,244 (39.7)
Between R$ 501.00 and R$ 1000.00	10,617 (25.4)	205 (24.8)	10,412 (25.4)
More than R$ 1001.00	3,381 (8.1)	49 (5.9)	3,332 (8.1)
Missing	11,189 (26.8)	241 (29.2)	10,948 (26.7)
Housing status: n (%)			
Owned	23,450 (56.2)	398 (48.2)	23,052 (56.3)
Rented	8,880 (21.3)	235 (28.5)	8,645 (21.1)
Donated	7,101 (17.0)	137 (16.6)	6,964 (17.0)
Missing	2,331 (5.6)	56 (6.8)	2,275 (5.6)
Type of water treatment used: n (%)			
Filtered	9,103 (21.8)	144 (17.4)	8,959 (21.9)
Boiled	834 (2.0)	23 (2.8)	811 (2.0)

Continued on next page

Table 3 - Continued from previous page

Variables	Total	Positive	Negative
Disinfected (chlorine)	22,503 (53.9)	441 (53.4)	22,062 (53.9)
None	5,795 (13.9)	133 (16.1)	5,662 (13.8)
Missing	3,527 (8.4)	85 (10.3)	3,442 (8.4)
Age: mean (SD)	25.2 (4.6)	24.9 (4.7)	25.2 (4.6)

n: number; %: proportional percentage value; SD: standard deviation.

Source: Produced by the author.

4.2 Experiments' methodology

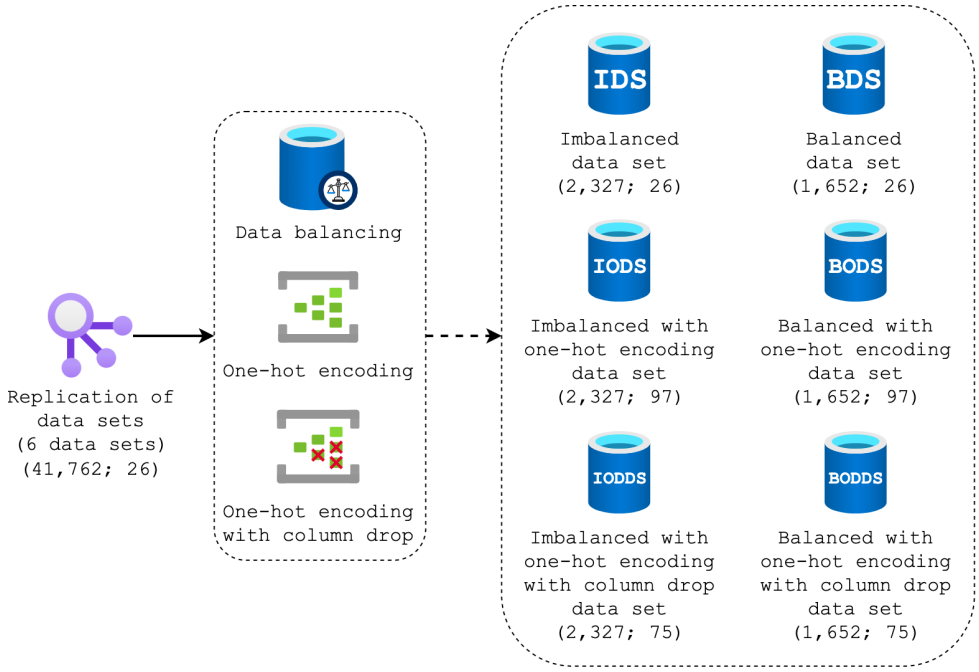
In this work, we defined six different experiments in order to compare the performance of machine learning models when handling different configurations of data sets. The main idea is to compare the impacts of using *(i)* imbalanced data and *(ii)* the one-hot encoding technique on the models' learning and performance. We evaluate the following machine learning techniques: Decision Tree, Random Forest, AdaBoost, GBM, XGBoost, KNN, and SVM. For each experiment, we built a data set according to the following characteristics:

- **Imbalanced Data Set (IDS)**: imbalanced data set with 2,327 records (826 positive cases and 1,501 negative cases) and 26 attributes. As the original data set contained numerous negative cases when compared to the number of positive cases (40,936 negative cases and 826 positive cases), we used the random undersampling technique to reduce the difference between the positive and negative congenital syphilis cases, setting a ratio of 55% of the number of samples in the minority class (positive cases) over the number of samples in the majority class (negative cases) after resampling.
- **Balanced Data Set (BDS)**: balanced data set using the random undersampling technique with 1,652 records (826 positive cases and 826 negative cases).
- **Imbalanced with One-hot Encoding Data Set (IODS)**: imbalanced data set with one-hot encoding technique applied to transform categorical data into binary data. In this case, the number of attributes increased to 97, since the one-hot encoding creates a new attribute for each class of a given categorical attribute. For example, the categorical attribute that indicates whether the pregnant woman is a smoker (SMOKER) has three possible categories: positive, negative, and not informed. Upon application of the one-hot encoding technique, this single attribute will be split into three binary attributes: (i) positive for smokers; (ii) negative for smokers; and (iii) not informed.
- **Balanced with One-hot Encoding Data Set (BODS)**: balanced data set with one-hot encoding technique applied.

- **Imbalanced with One-hot Encoding with Column Drop Data Set (IODDS)**: imbalanced data set with one-hot encoding technique applied with the column related to not informed by the patient removed from the data set. Some attributes have a class that represents the missing data. In this experiment, after applying the one-hot, we removed the column related to that, decreasing the number of attributes to 75.
- **Balanced with One-hot Encoding with Column Drop Data Set (BODDS)**: balanced data set and one-hot encoding with the column related to not informed by the patient removed from the data set.

Fig 8 illustrates how the data set was configured for each experiment.

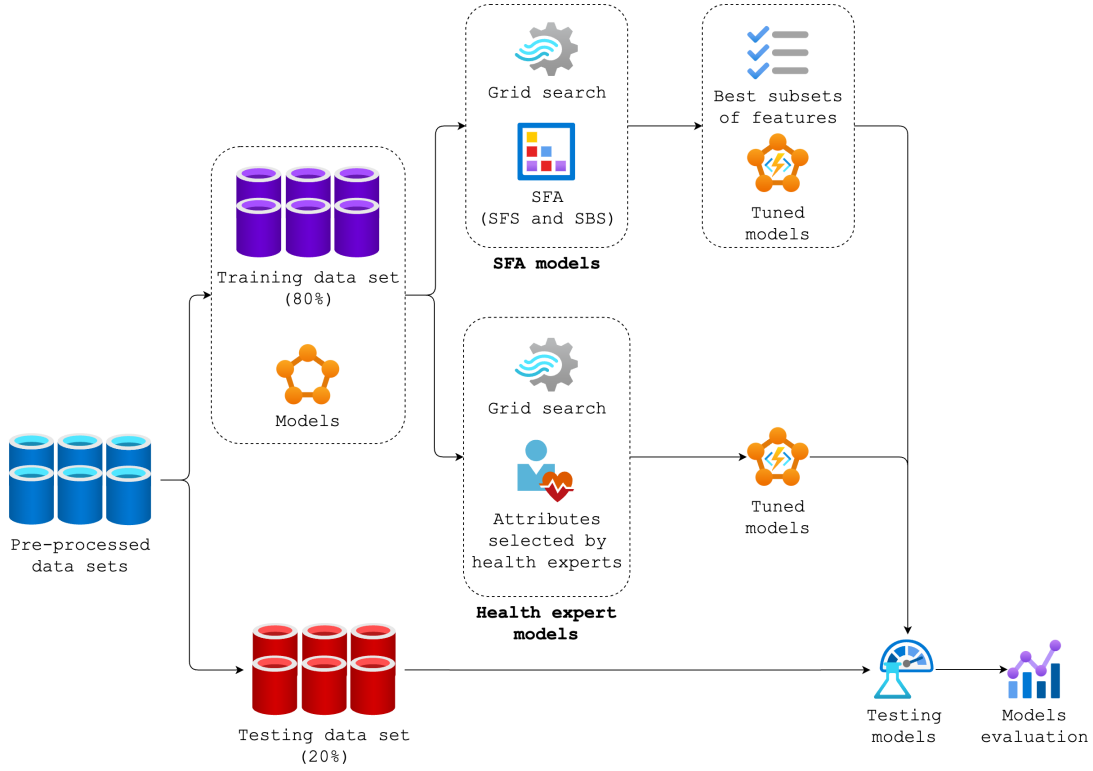
Figure 8 – Creation of data set for the six experiment using data balancing and one-hot encoding techniques.



Source: Produced by the author.

Fig 9 presents the methodology used to train and test our models. After the creation of data sets for each experiment, they were split into training (80%) and testing (20%) sets. In order to evaluate the best approach to selecting the most effective subset of attributes, we applied two different training processes to all experiments, using distinct feature selection techniques (SFA models and health expert models), resulting in 126 tuned models to be evaluated. We used the hold-out method to evaluate the models

Figure 9 – Design flowchart of models training and testing.



Source: Produced by the author.

4.2.1 SFA models

In order to automatically find the best subset of attributes (feature selection) and the best model configuration (hyperparameter optimization), in a given search space, two techniques were applied: (*i*) SFA and (*ii*) grid search, respectively. For each combination of the grid search technique (Table 4 presents the hyperparameters used in the grid search), two flavors of the SFA technique were performed (SFS and SBS).

All executions were based on the accuracy metric with cross-validation ($k = 10$). At the end, the testing data set had its dimensionality reduced based on the best subsets of features for each model, and it was used to evaluate them.

4.2.2 Health expert models

As a complement to our experiments, we consulted the health experts from PMCP and asked them to manually select a subset of attributes from the 26 described in Table 2. They analyzed those attributes and selected 11 of them, in addition to the target attribute, highlighted in Table 2 with a star icon (★). Then, the data sets from all experiments (IDS, BDS, IODS, BODS, IODDS, and BODDS) were filtered to reflect the attributes selected by the health experts.

On the basis of the accuracy metric with cross-validation ($k = 10$), the grid search

Table 4 – Parameters used in the grid search.

Model	Parameters	Values
Decision Tree	criterion splitter	[“gini", “entropy"] [“best", “random"]
Random Forest	n_estimators criterion	[50, 100, 150] [“gini", “entropy"]
AdaBoost	n_estimators learning_rate	[50, 100, 150] [0.5, 1]
GBM	n_estimators learning_rate loss	[50, 100, 150] [0.5, 1] [“deviance", “exponential"]
XGBoost	learning_rate max_depth	[0.3, 0.5] [5, 10]
KNN	n_neighbors p weights	[5, 10, 15] [1, 2] [“uniform", “distance"]
SVM	kernel gamma	[“rbf", “linear", “poly", “sigmoid"] [“scale", “auto"]

Source: Produced by the author.

technique was also used to identify the best model configuration in a given search space.

4.3 Technical resources

The proposed experiments were funded by Public Notice CNPq/AWS 032/2019 - Access to AWS Cloud Computing Platforms (Cloud Credits for Research), performed on an AWS EC2 instance of type c5a.4xlarge, with the following configuration:

- 2nd generation AMD EPYC 7002 series processors running at frequencies up to 3.3 GHz.
- 16 vCPUs.
- 32 GB of RAM.
- 30 GB SSD.

Chapter 5

Results

This chapter presents the results of this work. Section 5.1 presents the evaluation steps used to evaluate the machine learning models from all experiments, and section 5.2 presents the machine learning models evaluation.

5.1 Evaluation steps

In this work, we use clinical and sociodemographic data and two training processes (SFA models and health expert models) to evaluate seven machine learning models under six different experiments. Because there were so many tuned models to be evaluated, we set up three evaluation steps to figure out which model was more efficient for classifying congenital syphilis:

- **First step:** we identified the best models for each experiment from both training processes, considering only the F1-Score metric, thus defining the semifinalist models. This metric was chosen because it is a harmonic mean between precision and sensitivity, which are two relevant metrics for the analysis of health problems. Due to the large amount of data, we decided to not use all the evaluation metrics and compare the attributes chosen for each model, because that would make evaluating the results more complex.
- **Second step:** we analyzed the semifinalists from the SFA models and the health expert models and identified which model from each of these two training processes of experiments performed the best. Once again, the F1-Score metric was used, and the finalist models were defined.
- **Third step:** we compared the finalist models of each training process, considering all evaluation metrics to determine which model performed better. A comparison was also carried out between the attributes chosen by the SFA technique for the finalist SFA model and the ones chosen by the health experts used by the finalist expert model.

5.2 Evaluation of the machine learning models

Table 5 and 6 presents the top-3 best models of each experiment that used attributes selected by the SFA technique and by health experts, respectively. Appendices A and B presents all results from all experiments. The best model of each experiment (semifinalist) is highlighted with an upside down triangle icon (▼) and the best model among the experiments (finalist) are highlighted with a circle icon (●).

Table 5 – Results of the top-3 SFA models of each experiment.

Model	SFA	Qty. Att.	F1-Score	Accuracy	Precision	Sensitivity	Specificity
IDS							
Random Forest▼	SBS	15	34.85%	63.09%	47.42%	27.54%	82.94%
KNN	SBS	16	33.71%	62.45%	45.83%	26.35%	82.61%
AdaBoost	SBS	10	30.71%	64.16%	50.00%	22.16%	87.63%
BDS							
SVM▼●	SBS	13	63.04%	61.03%	60.11%	66.27%	55.76%
AdaBoost	SFS	9	61.33%	57.70%	56.63%	66.87%	48.48%
GBM	SBS	12	58.08%	57.70%	57.74%	58.43%	56.97%
IODS							
Random Forest▼	SBS	42	35.52%	64.16%	50.00%	27.54%	84.62%
KNN	SBS	34	31.91%	58.80%	39.13%	26.95%	76.59%
Decision Tree	SBS	11	30.53%	60.94%	42.11%	23.95%	81.61%
BODS							
Decision Tree▼	SBS	16	60.44%	56.50%	55.56%	66.27%	46.67%
SVM	SBS	37	60.41%	59.21%	58.86%	62.05%	56.36%
GBM	SFS	32	59.71%	58.01%	57.54%	62.05%	53.94%
IODDS							
XGBoost▼	SBS	16	43.92%	64.38%	50.39%	38.92%	78.60%
KNN	SBS	41	37.80%	61.16%	44.35%	32.93%	76.92%
Random Forest	SBS	39	35.48%	65.67%	54.32%	26.35%	87.63%
BODDS							
SVM▼	SFS	14	59.08%	59.82%	60.38%	57.83%	61.82%
AdaBoost	SFS	26	58.38%	56.50%	56.11%	60.84%	52.12%
Decision Tree	SFS	11	58.14%	56.50%	56.18%	60.24%	52.73%

▼ Best model of the experiment; ● Best model among the experiments.

Source: Produced by the author.

The results of the experiments presented a large variation of F1-Score values, ranging from 12.70% to 63.51%. In general, the experiments with imbalanced data set, which had a greater number of negative cases than positive cases of congenital syphilis, obtained inferior results than their respective experiment with balanced data set, creating models with a low rate of correct classification of truly positive cases (low sensitivity), directly affecting the F1-Score, and with a high rate of correct classification of truly negative cases (high specificity). Therefore, such models are not relevant to the purpose of this work, since we aim to predict positive cases

Table 6 – Results of the top-3 health expert models for each experiment.

Model	F1-Score	Accuracy	Precision	Sensitivity	Specificity
IDS					
Random Forest▼	43.83%	60.94%	45.22%	42.51%	71.24%
Decision Tree	38.41%	56.65%	39.13%	37.72%	67.22%
GBM	36.90%	63.30%	48.08%	29.94%	81.94%
BDS					
KNN▼	62.37%	57.70%	56.31%	69.88%	45.45%
AdaBoost	62.01%	58.91%	57.81%	66.87%	50.91%
Random Forest	61.99%	60.73%	60.23%	63.86%	57.58%
IODS					
Decision Tree▼	44.84%	59.87%	44.19%	45.51%	67.89%
GBM	35.84%	61.59%	44.64%	29.94%	79.26%
XGBoost	34.11%	57.73%	38.64%	30.54%	72.91%
BODS					
AdaBoost▼•	63.51%	60.42%	59.07%	68.67%	52.12%
SVM	62.98%	59.52%	58.16%	68.67%	50.30%
GBM	59.39%	59.52%	59.76%	59.04%	60.00%
IODDS					
Decision Tree▼	41.79%	58.15%	41.67%	41.92%	67.22%
Random Forest	38.41%	63.52%	48.62%	31.74%	81.27%
XGBoost	36.84%	58.80%	40.88%	33.53%	72.91%
BODDS					
SVM▼	62.01%	58.91%	57.81%	66.87%	50.91%
AdaBoost	61.80%	58.91%	57.89%	66.27%	51.52%
Random Forest	60.66%	60.42%	60.48%	60.84%	60.00%

▼ Best model of the experiment; • Best model among the experiments.

Source: Produced by the author.

of congenital syphilis.

It was possible to observe that the experiments that used the one-hot encoding technique, dropping or not dropping the attribute related to not informed data, produced results that were not significantly different from those that did not use it. The large amount of missing data may have impacted the results of these models, impairing the models' learning.

For the SFA models, only experiments with imbalanced data set benefited from the one-hot encoding technique. When we compared the best models from the IODS and IODDS experiments with the best model from the IDS experiment, we found a slight improvement in almost all metrics, particularly for the XGBoost model from the IODDS experiment, which benefited from the dropping of the attribute related to not informed data, obtaining a higher sensitivity (38.92%), resulting in a better F1-Score value of 43.92%. Nonetheless, the XGBoost model did not achieve relevant results.

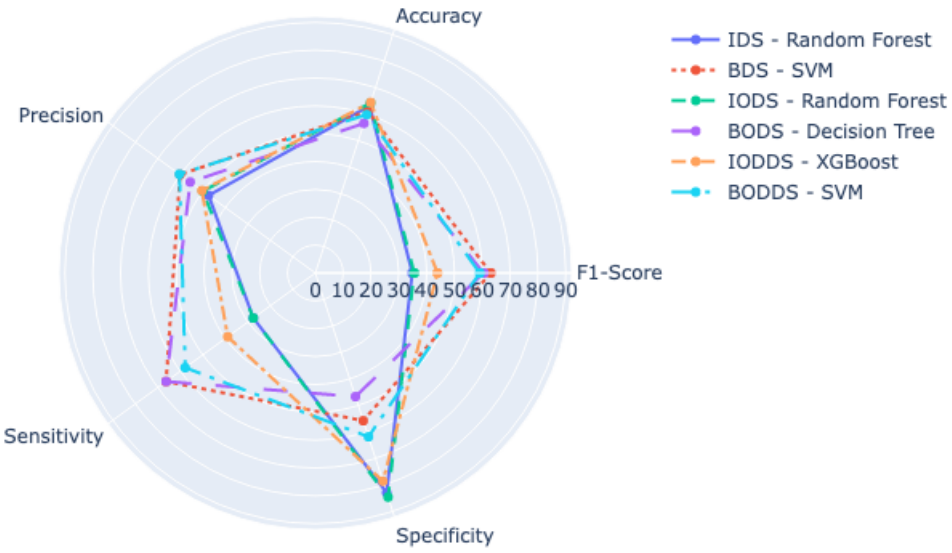
The one-hot encoding technique improved all health expert models. For experiments with

imbalanced data set, the F1-Score value of the Decision Tree model from the IODS experiment (44.84%) was slightly higher than the Random Forest model from the IDS experiment (43.83%). The Decision Tree model from the IODDS experiment, on the other hand, obtained a lower F1-Score value (41.79%), indicating that there was no benefit in dropping the attribute related to not informed data. As for the experiments with balanced data set, the AdaBoost model from the BODS experiment and the SVM model from the BODDS experiment performed a little better than the KNN from the BDS experiment, except for sensitivity.

The most effective results, among the models with the best performance from each experiment, were provided by the tree-based models (Random Forest, Decision Tree, AdaBoost, and XGBoost), standing out in four out of six experiments, both with attributes chosen through SFA technique and those chosen by health experts. Tree-based models are typically good choices for problems involving tabular data [41].

In the first evaluation step, the semifinalist SFA models of each experiment were selected, and they are compared in Figure 10, where it is possible to notice different results especially for F1-Score, sensitivity, and specificity metrics. Table X presents the results from the grid search technique of the semifinalist SFA models. These models presented subsets of attributes that ranged between 13 and 42 attributes, presented in Table X. The most common attributes were: EDUC_LEVEL, MARITAL_STATUS, FOOD_INSECURITY, WATER_TREATMENT, and SMOKER, which appeared in almost all experiments.

Figure 10 – Comparison of the semifinalist SFA models among all experiments.



Source: Produced by the author.

In the second evaluation step, the model that obtained the highest F1-Score was the SVM model when executing the BDS experiment, called SVM-BDS-SFA, using the SBS, with 13

attributes, $gama = scale$ and $kernal = rbf$. It reached a F1-Score of 63.04%, an accuracy of 61.03%, a precision of 60.11%, a sensitivity of 66.27%, and a specificity of 55.76%. On the other hand, the Decision Tree model in the IDS experiment, with 2 attributes selected by the SBS, $criterion = entropy$ and $splitter = best$, had the worst performance among the SFA models, with a F1-Score of 12.70%.

Figure 11 presents a comparison between the semifinalist health expert models selected in the first evaluation step. In the second evaluation step, the model with the best performance was the AdaBoost when executing the BODS experiment, called AdaBoost-BODS-Expert, with $learning\_rate = 0.5$ and $n\_estimators = 150$, which achieved a F1-Score of 63.51%, an accuracy of 60.42%, a precision of 59.07%, a sensitivity of 68.67%, and a specificity of 52.12%. Meanwhile, the SVM model in the IDS experiment, $gama = auto$ and $kernel = rbf$, reached a F1-Score of 20.37%, the worst one among the health experts experiments.

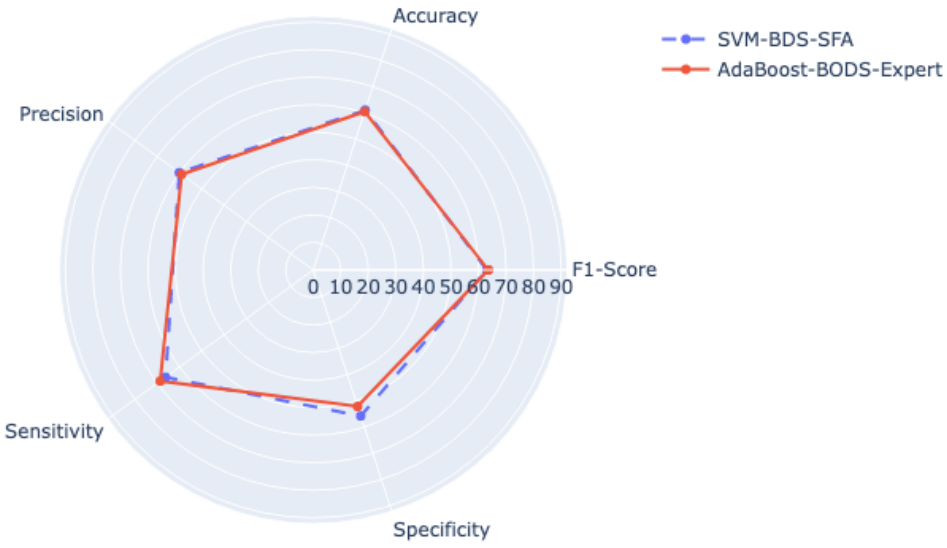
Figure 11 – Comparison of the semifinalist health expert models.

Source: Produced by the author.

Comparing the finalist models in the third evaluation step, we noticed that the SVM-BDS-SFA and the AdaBoost-BODS-Expert models achieved similar results, presenting major differences in the sensitivity, specificity, and attributes used. Figure 12 presents a comparison between them and Table 7 presents the grid search results and the attributes selected for both models.

Seven of the 11 attributes selected by health experts used by the AdaBoost-BODS-Expert model were also chosen for the SVM-BDS-SFA via the SFA technique, demonstrating that more than half of the attributes chosen by health experts are still relevant even when using SFA technique.

Figure 12 – Finalist models comparison: SVM-BDS-SFA and AdaBoost-BODS-Expert.



Source: Produced by the author.

The AdaBoost-BODS-Expert model is more interpretable by health experts and, as a result, would be more favored for usage by the PMCP, despite having a slightly lower performance when compared to the SVM-BDS-SFA model. In this sense, the AdaBoost-BODS-Expert model has a significant advantage over the SVM-BDS-SFA model, with approximately 2% difference in results for each metric. Even if it does not produce the greatest outcomes, a model that provides interpretability and knowledge of all attributes used might enable higher acceptance by health experts, producing more confidence and allowing adoption in daily usage [42].

None of the models presented an accuracy greater than 70%, ranging from 56.50% to 64.38%, demonstrating the difficulty of classifying possible outcomes of congenital syphilis using only clinical and sociodemographic data. The reason that may explain this fact is the abundance of missing data that reduces the data quality and model learning [43]. In order to cover this issue, we created a categorical value that represented that one that was not informed by the patient. However, this may have caused: (*i*) difficulty in finding patterns in the data that would allow a more accurate classification and (*ii*) relevance for categories related to not informed data, which may hinder the classification of more realistic data, when the attributes are correctly filled in. As potential improvements, we recommend (*i*) the use of data imputation technique to reduce the proportion of missing data in the existing data set, as well as (*ii*) a better completeness of the data by the PMCP, which is a significant improvement that allows the data to more precisely reflect reality.

Although, according to the health experts from the PMCP, this work is very applicable in the daily monitoring of pregnant women, being relevant for the development of new strategies

Table 7 – Grid search and feature selection results for the finalist models.

Model	Hyperparameters	Qty. Att.	Attributes
SVM-BDS-SFA	gamma: scale kernel: rbf	13	EDUC_LEVEL, HAS_FRU_TREE, WATER_TREATMENT, RH_FACTOR, PLAN_PREGNANCY, HAS_PREG_RISK, TET_VACCINE, IS_HEAD_FAMILY, MARITAL_STATUS, FOOD_INSECURITY, NUM_ABORTIONS, NUM_PREGNANCIES, and HAS_FAM_INCOME
AdaBoost-BODS-Expert	learning_rate: 0.5 n_estimators: 150	11	AGE, LEVEL_SCHOOLING, FAM_INCOME, PLAN_PREGNANCY, HAS_PREG_RISK, MARITAL_STATUS, FOOD_INSECURITY, NUM_ABORTIONS, NUM_LIV_CHILDREN, NUM_PREGNANCIES, and FAM_PLANNING

Source: Produced by the author.

in the PMCP, enabling more qualified monitoring and enabling the creation of protocols that today do not exist in the PMCP. Also, the use of AdaBoost-BODS-Expert model will raise the alert level of PMCP health professionals. This will allow better monitoring of all pregnancies, not just those that are thought to be possible cases of congenital syphilis, which will improve the quality of service. In addition, the low data quality reported in this work shows the need to rethink the data collection process, emphasizing the importance of correctly filling in the data in the SIS-MC. The feedback from health experts from PMCP regarding the results obtained in this work is available at <https://youtu.be/a80XlyrTH0M>.

In this chapter, were presented the results of this work. After applying all the evaluation steps, the SVM-BDS-SFA model and AdaBoost-BODS-Expert model were selected as the finalist models from the SFA models and health expert models, respectively. And the AdaBoost-BODS-Expert model was chosen as the favored model for usage by the PMCP to classify possible outcomes of congenital syphilis.

Chapter 6

Final considerations and next steps

Congenital syphilis has become a serious public health problem, with a significant increase of cases in Brazil [1]. The PHC has tests and treatments for congenital syphilis, but the test availability and the full application of the treatment, specially including the sex partners of the pregnant woman [4], have made it a significant issue. Although in 2020 there was a decrease in the incidence rate of congenital syphilis and in the rate of mortality due to congenital syphilis in children under one year, part of this reduction may be related to problems in sharing data between the layers of SUS management [1]. In addition, the decline in the number of cases may also result from the underreporting of cases due to the local mobilization of health professionals caused by the COVID-19 pandemic [1].

In this work, machine learning models were evaluated for classification of possible congenital syphilis outcomes in pregnancies assisted by the PMCP. We used data sets provided by the PMCP extracted from their information system (SIS-MC), with clinical and sociodemographic data regarding antenatal care, pregnant women's outcomes, and their children from the cities served by the PMCP in the State of Pernambuco, Brazil, between the years of 2013 and 2021. After applying the pre-processing methodology, the pre-processed data set showed a great imbalance between positive and negative cases of congenital syphilis and also a great number of not informed data.

We evaluated seven machine learning techniques for the prediction of congenital syphilis cases: Decision Tree, Random Forest, AdaBoost, GBM, XGBoost, KNN, and SVM. To deal with data imbalance and not informed data, we used the random undersampling and one-hot encoding techniques to propose six experiments (IDS, BDS, IODS, BODS, IODDS, and BODDS), executing a grid search for each model along with the SFA technique (SFA models). We also executed all experiments only with the grid search, using 11 attributes selected by health experts from PMCP (health expert models).

The SVM model from the BDS experiment that used the SFA technique to select its 13 attributes, called SVM-BDS-SFA, and the AdaBoost model from the BODS experiment with 11 attributes chosen by health experts, named AdaBoost-BODS-Expert, produced the best results in

terms of F1-Score metric. When comparing metrics of SVM-BDS-SFA and AdaBoost-BODS-Expert models, both models obtained similar outcomes, although the SVM-BDS-SFA model outperformed the AdaBoost-BODS-Expert model in almost all metrics, except sensitivity and F1-Score.

6.1 Contributions

The present work developed intellectual contributions and products throughout the development of the master's project. All the code developed in this work are available in <https://github.com/dotlab-brazil/congenital-syphilis>. The AdaBoost-BODS-Expert model is registered with the National Institute of Industrial Property, from Portuguese *Instituto da Propriedade Industrial* (INPI) under procedure number **BR 51 2022 002788-7**. The data set used in the experiment is public available in Mendeley Data [44]. Two different papers were written:

- Predicting congenital syphilis cases: a performance evaluation of different machine learning models [45]: Teixeira, Igor Vitor and Leite, Morgana Thalita da Silva and Melo, Flavio Leandro de Morais and Rocha, Elisson da Silva Rocha and Sadok, Sara and Carrarine, Ana Sofia Pessoa da Costa and Santana, Marilia, Rodrigues, Cristina Pinheiro and Oliveira, Ana Maria de Lima and Gadelha, Keduly Vieira and Morais, Cleber Matos de and Kelner, Judith and Endo, Patricia Takako. PLOS ONE, Public Library of Science (PLoS) - Under reviw.
- Syphilis Trigram: a domain-specific visualisation to combat syphilis epidemic and improve the quality of maternal and child health in Brazil [46]: Morais, Cleber Matos de and Teixeira, Igor Vitor and Sadok, Sara and Endo, Patricia Takako and Kelner, Judith. BMC Pregnancy and Childbirth, Springer Science and Business Media LLC - Published.

6.2 Challenges and Limitations

As reported in this work, no studies using machine learning techniques to classify congenital syphilis cases were found in the literature. This research explored this limitation through the application of optimization techniques in the data set used, as well as in the proposed models and experiments. Our results showed that, despite being a challenge, it is possible to predict congenital syphilis during pregnancy only using clinical and sociodemographic data.

At the same time, we also identified the limitations generated by the large amount of missing data, which may have precluded more accurate results, indicating the need for the PMCP to improve data acquisition quality.

6.3 Future works

As future work, we intend to apply different data balancing and imputation techniques to investigate alternative methods for dealing with data set imbalances and missing data. It is planned to acquire new data from the SIS-MC of the PMCP for the year 2022, enabling the increase of records referring to positive cases of congenital syphilis in the data set. We plan to integrate the SIS-MC data with public databases made available by the Brazilian government, in order to increase the number of positive cases of congenital syphilis and retrain the proposed models with national data.

Finally, we also plan to expand our work to predict gestational syphilis. The timely identification of syphilis cases in pregnant women can help to reduce the incidence of congenital syphilis. We also plan to apply different techniques to handle the missing data.

Bibliography

[1] BRAZIL, M. of Health of. *Boletim Epidemiológico Sífilis*. 2021. Disponível em: <https://www.gov.br/aids/pt-br/centrais-de-conteudo/boletins-epidemiologicos/2021/sifilis/boletim_sifilis_2021_internet.pdf/@@download/file/boletim_sifilis_2021_internet.pdf>.

[2] BRASIL, M. da Saúde do. *Guia de Vigilância em Saúde*. 2020. Accessed 12 Dec 2020. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/guia_vigilancia_saude_3ed.pdf>.

[3] DOMINGUES, C. S. B.; DUARTE, G.; PASSOS, M. R. L.; SZTAJNBOK, D. C. d. N.; MENEZES, M. L. B. Brazilian protocol for sexually transmitted infections, 2020: congenital syphilis and child exposed to syphilis. *Revista da Sociedade Brasileira de Medicina Tropical*, SciELO Brasil, v. 54, 2021.

[4] BRASIL, M. da Saúde do. *Protocolo clínico e diretrizes terapêuticas para atenção integral às pessoas com infecções sexualmente transmissíveis (IST)*. 2020. Accessed 7 Dec 2022. Disponível em: <https://www.gov.br/aids/pt-br/centrais-de-conteudo/pcdts/2022/ist/pcdt-ist-2022_isbn-1.pdf/view>.

[5] MACÊDO, V. C. d.; LIRA, P. I. C. d.; FRIAS, P. G. d.; ROMAGUERA, L. M. D.; CAIRES, S. d. F. F.; XIMENES, R. A. d. A. Fatores de risco para sífilis em mulheres: estudo caso-controle. *Revista de Saúde Pública*, SciELO Brasil, v. 51, 2017.

[6] SAÚDE, S. de. *Programa Mãe Coruja Pernambucana*. 2021. Disponível em: <https://maecoruja.pe.gov.br/>. Acesso em: 09 out. 2021.

[7] BRASIL, M. da Saúde do. *Portaria No 77, de 12 de janeiro de 2021*. 2020. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/media/pdf/2020/outubro/29/BoletimSfilis2020especial.pdf>. Acesso em: 12 jul. 2021.

[8] SANTOS, M. M. d.; ROSENDO, T. M. S. d. S.; LOPES, A. K. B.; RONCALLI, A. G.; LIMA, K. C. d. Weaknesses in primary health care favor the growth of acquired syphilis. *PLoS neglected tropical diseases*, Public Library of Science San Francisco, CA USA, v. 15, n. 2, p. e0009085, 2021.

[9] TERRA. *Bolsonaro corta investimentos em Educação, Saúde e Segurança*. 2020. Disponível em: <https://www.terra.com.br/economia/bolsonaro-corta-investimentos-em-educacao-saude-e-seguranca,a0c81ff72f5ab50614d67ac1bd1b057a392c245i.html>. Acesso em: 11 nov. 2020.

[10] NATIONS, U. *Ensure healthy lives and promoting well-being for all at all ages*. 2022. Accessed 8 Dec 2022. Disponível em: <https://sdgs.un.org/topics/health-and-population>.

[11] YOUNG, S. D.; MERCER, N.; WEISS, R. E.; TORRONE, E. A.; ARAL, S. O. Using social media as a tool to predict syphilis. *Preventive Medicine*, Elsevier BV, v. 109, p. 58–61, abr. 2018. Disponível em: <https://doi.org/10.1016/j.ypmed.2017.12.016>.

[12] SILVA, R. D. d. *Análise preditiva baseada em dados para criação de perfil de grupos de risco no SUS: um estudo de caso aplicado a sífilis no Brasil*. Dissertação (Mestrado) — Universidade Federal do Rio Grande do Norte, 2020.

[13] LIMA, M. G.; SANTOS, R. F. R. d.; BARBOSA, G. J. A.; RIBEIRO, G. d. S. Incidência e fatores de risco para sífilis congênita em belo horizonte, minas gerais, 2001-2008. *Ciência & Saúde Coletiva*, SciELO Public Health, v. 18, p. 499–506, 2013.

[14] MELO, N. G. D. O.; FILHO, D. A. d. M.; FERREIRA, L. O. C. Diferenciais intraurbanos de sífilis congênita no recife, pernambuco, brasil (2004-2006). *Epidemiologia e Serviços de Saúde*, Geral de Desenvolvimento da Epidemiologia em Serviços/Secretaria de . . . , v. 20, n. 2, p. 213–222, 2011.

[15] ESMAILY, H.; TAYEFI, M.; DOOSTI, H.; GHAYOUR-MOBARHAN, M.; NEZAMI, H.; AMIRABADIZADEH, A. A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes. *Journal of research in health sciences*, Hamadan University of Medical Sciences, v. 18, n. 2, p. 412, 2018.

[16] PRASAD, B. *A Gentle Introduction to Decision Tree in Machine Learning - Life With Data*. 2022. Accessed 26 September 2022. Disponível em: <https://lifewithdata.com/2022/07/14/a-gentle-introduction-to-decision-tree-in-machine-learning/>.

[17] BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.

[18] YING, C.; QI-GUANG, M.; JIA-CHEN, L.; LIN, G. Advance and prospects of adaboost algorithm. *Acta Automatica Sinica*, Elsevier, v. 39, n. 6, p. 745–758, 2013.

[19] SCHAPIRE, R. E. Explaining adaboost. In: *Empirical inference*. [S.l.]: Springer, 2013. p. 37–52.

[20] AYYADEVARA, V. K. Gradient boosting machine. In: *Pro machine learning algorithms*. [S.l.]: Springer, 2018. p. 117–134.

[21] MITCHELL, R.; FRANK, E. Accelerating the xgboost algorithm using gpu computing. *PeerJ Computer Science*, PeerJ Inc., v. 3, p. e127, 2017.

[22] ZHANG, S.; CHENG, D.; DENG, Z.; ZONG, M.; DENG, X. A novel knn algorithm with data-driven k parameter computation. *Pattern Recognition Letters*, Elsevier, v. 109, p. 44–54, 2018.

[23] LORENA, A. C.; CARVALHO, A. C. D. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, 2007.

[24] NOBLE, W. S. What is a support vector machine? *Nature biotechnology*, Nature Publishing Group, v. 24, n. 12, p. 1565–1567, 2006.

[25] BHAVSAR, H.; GANATRA, A. A comparative study of training algorithms for supervised machine learning. *International Journal of Soft Computing and Engineering (IJSCE)*, Citeseer, v. 2, n. 4, p. 2231–2307, 2012.

[26] PRATI, R. C.; BATISTA, G. E.; MONARD, M. C. Class imbalances versus class overlapping: an analysis of a learning system behavior. In: SPRINGER. *Mexican international conference on artificial intelligence*. [S.l.], 2004. p. 312–321.

[27] BATISTA, G. E.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, ACM New York, NY, USA, v. 6, n. 1, p. 20–29, 2004.

[28] HE, H.; MA, Y. *Imbalanced learning: foundations, algorithms, and applications*. [S.l.]: John Wiley & Sons, 2013.

[29] POTDAR, K.; PARDAWALA, T. S.; PAI, C. D. A comparative study of categorical variable encoding techniques for neural network classifiers. *International journal of computer applications*, v. 175, n. 4, p. 7–9, 2017.

[30] SEGER, C. *An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing*. 2018.

[31] WU, J.; CHEN, X.-Y.; ZHANG, H.; XIONG, L.-D.; LEI, H.; DENG, S.-H. Hyperparameter optimization for machine learning models based on bayesian optimization. *Journal of Electronic Science and Technology*, Elsevier, v. 17, n. 1, p. 26–40, 2019.

[32] VENKATESH, B.; ANURADHA, J. A review of feature selection and its methods. *Cybernetics and Information Technologies*, v. 19, n. 1, p. 3–26, 2019. Disponível em: <https://doi.org/10.2478/cait-2019-0001>.

[33] MIAO, J.; NIU, L. A survey on feature selection. *Procedia Computer Science*, Elsevier, v. 91, p. 919–926, 2016.

[34] SEQUENTIALFEATURESELECTOR: The popular forward and backward feature selection approaches (including floating variants). <http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/>. Accessed 26 September 2022.

[35] OLIVEIRA, T. T. d. *Development of machine learning models to aid the diagnosis of arboviruses using clinical data*. Dissertação (Mestrado) — Universidade de Pernambuco, 2022.

[36] SUSMAGA, R. Confusion matrix visualization. In: *Intelligent information processing and web mining*. [S.l.]: Springer, 2004. p. 107–116.

[37] HOSSIN, M.; SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015.

[38] OLSON, D. L.; DELEN, D. *Advanced data mining techniques*. [S.l.]: Springer Science & Business Media, 2008.

[39] PARIKH, R.; MATHAI, A.; PARIKH, S.; SEKHAR, G. C.; THOMAS, R. Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, Wolters Kluwer–Medknow Publications, v. 56, n. 1, p. 45, 2008.

[40] CHICCO, D.; JURMAN, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, Springer, v. 21, n. 1, p. 1–13, 2020.

[41] SHWARTZ-ZIV, R.; ARMON, A. Tabular data: Deep learning is not all you need. *Information Fusion*, Elsevier BV, v. 81, p. 84–90, maio 2022. Disponível em: <https://doi.org/10.1016/j.inffus.2021.11.011>.

[42] OLIVEIRA, T. T. de; NETO, S. R. da S.; TEIXEIRA, I. V.; OLIVEIRA, S. B. A. de; RODRIGUES, M. G. de A.; SAMPAIO, V. S.; ENDO, P. T. A comparative study of machine learning techniques for multi-class classification of arboviral diseases. *Frontiers in Tropical Diseases*, Frontiers Media SA, v. 2, fev. 2022. Disponível em: <https://doi.org/10.3389/fitd.2021.769968>.

[43] EHSANI-MOGHADDAM, B.; MARTIN, K.; QUEENAN, J. A. Data quality in healthcare: A report of practical experience with the canadian primary care sentinel surveillance network data. *Health Information Management Journal*, SAGE Publications, v. 50, n. 1-2, p. 88–92, dez. 2019. Disponível em: <https://doi.org/10.1177/1833358319887743>.

[44] TEIXEIRA, I. V. *Clinical and sociodemographic data on congenital syphilis cases, Brazil, 2013-2021*. Mendeley, 2022. Disponível em: <https://data.mendeley.com/datasets/3zkcvybvkz/1>.

[45] TEIXEIRA, I. V.; LEITE, M. T. d. S.; MELO, F. L. d. M.; ROCHA, E. d. S. R.; SADOK, S.; CARRARINE, A. S. P. d. C.; SANTANA MARILIA, R. C. P.; OLIVEIRA, A. M. d. L.; GADELHA, K. V.; MORAIS, C. M. d.; KELNER, J.; ENDO, P. T. Predicting congenital syphilis cases: a performance evaluation of different machine learning models. *PLOS ONE*, Public Library of Science (PLoS), 2022.

[46] MORAIS, C. M. de; TEIXEIRA, I. V.; SADOK, S.; ENDO, P. T.; KELNER, J. Syphilis trigram: a domain-specific visualisation to combat syphilis epidemic and improve the quality of maternal and child health in brazil. *BMC Pregnancy and Childbirth*, Springer Science and Business Media LLC, v. 22, n. 1, maio 2022. Disponível em: <https://doi.org/10.1186/s12884-022-04651-w>.

Appendix A

Results of the SFA models of each experiment

Table 8 – Results of the SFA models of each experiment.

Model	SFA	Qty. Att.	F1-Score	Accuracy	Precision	Sensitivity	Specificity
IDS							
Decision Tree	SBS	2	12,70%	64,59%	54,55%	7,19%	96,66%
Random Forest▼	SBS	15	34,85%	63,09%	47,42%	27,54%	82,94%
AdaBoost	SBS	10	30,71%	64,16%	50,00%	22,16%	87,63%
GBM	SBS	4	24,00%	67,38%	72,73%	14,37%	96,99%
XGBoost	SBS	5	25,78%	64,16%	50,00%	17,37%	90,30%
KNN	SBS	16	33,46%	62,45%	45,83%	26,35%	82,61%
SVM	SFS	20	15,08%	63,73%	46,88%	8,98%	94,31%
BDS							
Decision Tree	SFS	4	54,81%	53,17%	53,11%	56,63%	49,70%
Random Forest	SFS	4	56,47%	55,29%	55,17%	57,83%	52,73%
AdaBoost	SFS	9	61,33%	57,70%	56,63%	66,87%	48,48%
GBM	SBS	12	58,08%	57,70%	57,74%	58,43%	56,97%
XGBoost	SBS	12	56,97%	58,01%	58,60%	55,42%	60,61%
KNN	SBS	22	53,18%	51,06%	51,11%	55,42%	46,67%
SVM▼•	SBS	13	63,04%	61,03%	60,11%	66,27%	55,76%
IODS							
Decision Tree	SBS	11	30,53%	60,94%	42,11%	23,95%	81,61%
Random Forest▼	SBS	42	35,52%	64,16%	50,00%	27,54%	84,62%
AdaBoost	SBS	13	29,07%	65,45%	55,00%	19,76%	90,97%
GBM	SBS	39	30,04%	65,02%	53,03%	20,96%	89,63%

Continued on next page

Table 8 - Continued from previous page

Model	SFA	Qty. Att.	F1-Score	Accuracy	Precision	Sensitivity	Specificity
XGBoost	SBS	9	20,39%	64,81%	53,85%	12,57%	93,98%
KNN	SBS	34	31,91%	58,80%	39,13%	26,95%	76,59%
SVM	SBS	25	27,23%	63,30%	47,06%	19,16%	87,96%
BODS							
Decision Tree▼	SBS	16	60,44%	56,50%	55,56%	66,27%	46,67%
Random Forest	SFS	29	55,59%	55,59%	55,76%	55,42%	55,76%
AdaBoost	SBS	21	59,49%	56,80%	56,15%	63,25%	50,30%
GBM	SFS	32	59,71%	58,01%	57,54%	62,05%	53,94%
XGBoost	SFS	24	57,57%	56,80%	56,73%	58,43%	55,15%
KNN	SBS	51	54,38%	54,38%	54,55%	54,22%	54,55%
SVM	SBS	37	60,41%	59,21%	58,86%	62,05%	56,36%
IODDS							
Decision Tree	SBS	5	31,90%	66,09%	56,92%	22,16%	90,64%
Random Forest	SBS	39	35,48%	65,67%	54,32%	26,35%	87,63%
AdaBoost	SBS	18	29,07%	65,45%	55,00%	19,76%	90,97%
GBM	SBS	11	27,03%	65,24%	54,55%	17,96%	91,64%
XGBoost▼	SBS	16	43,92%	64,38%	50,39%	38,92%	78,60%
KNN	SBS	41	37,80%	61,16%	44,35%	32,93%	76,92%
SVM	SFS	23	27,43%	64,81%	52,54%	18,56%	90,64%
BODDS							
Decision Tree	SFS	11	58,14%	56,50%	56,18%	60,24%	52,73%
Random Forest	SFS	18	57,31%	54,98%	54,64%	60,24%	49,70%
AdaBoost	SFS	26	58,38%	56,50%	56,11%	60,84%	52,12%
GBM	SFS	16	56,46%	56,19%	56,29%	56,63%	55,76%
XGBoost	SBS	35	57,57%	56,80%	56,73%	58,43%	55,15%
KNN	SFS	11	53,69%	52,57%	52,60%	54,82%	50,30%
SVM▼	SFS	14	59,08%	59,82%	60,38%	57,83%	61,82%

▼ Best model of the experiment; • Best model among the experiments.

Source: Produced by the author.

Appendix B

Results of the health expert models of each experiment

Table 9 – Results of the health experts models of each experiment.

Modelo	F1-Score	Accuracy	Precision	Sensitivity	Specificity
IDS					
Decision Tree	38,41%	56,65%	39,13%	37,72%	67,22%
Random Forest▼	43,83%	60,94%	45,22%	42,51%	71,24%
AdaBoost	24,32%	63,95%	49,09%	16,17%	90,64%
GBM	36,90%	63,30%	48,08%	29,94%	81,94%
XGBoost	34,69%	58,80%	40,16%	30,54%	74,58%
KNN	30,45%	63,73%	48,68%	22,16%	86,96%
SVM	20,37%	63,09%	44,90%	13,17%	90,97%
BDS					
Decision Tree	53,53%	52,27%	52,30%	54,82%	49,70%
Random Forest	61,99%	60,73%	60,23%	63,86%	57,58%
AdaBoost	62,01%	58,91%	57,81%	66,87%	50,91%
GBM	59,17%	58,31%	58,14%	60,24%	56,36%
XGBoost	55,49%	53,47%	53,33%	57,83%	49,09%
KNN▼	62,37%	57,70%	56,31%	69,88%	45,45%
SVM	57,65%	56,50%	56,32%	59,04%	53,94%
IODS					
Decision Tree▼	44,84%	59,87%	44,19%	45,51%	67,89%
Random Forest	32,82%	62,23%	45,26%	25,75%	82,61%
AdaBoost	21,30%	63,52%	46,94%	13,77%	91,30%
GBM	35,84%	61,59%	44,64%	29,94%	79,26%

Continued on next page

Table 9 - Continued from previous page

Modelo	F1-Score	Accuracy	Precision	Sensitivity	Specificity
XGBoost	34,11%	57,73%	38,64%	30,54%	72,91%
KNN	31,50%	62,66%	45,98%	23,95%	84,28%
SVM	28,22%	62,88%	45,95%	20,36%	86,62%
BODS					
Decision Tree	54,17%	53,47%	53,53%	54,82%	52,12%
Random Forest	58,01%	58,01%	58,18%	57,83%	58,18%
AdaBoost▼•	63,51%	60,42%	59,07%	68,67%	52,12%
GBM	59,39%	59,52%	59,76%	59,04%	60,00%
XGBoost	57,40%	57,40%	57,58%	57,23%	57,58%
KNN	53,41%	52,57%	52,63%	54,22%	50,91%
SVM	62,98%	59,52%	58,16%	68,67%	50,30%
IODDS					
Decision Tree▼	41,79%	58,15%	41,67%	41,92%	67,22%
Random Forest	38,41%	63,52%	48,62%	31,74%	81,27%
AdaBoost	25,22%	63,09%	46,03%	17,37%	88,63%
GBM	36,49%	61,16%	44,07%	31,14%	77,93%
XGBoost	36,84%	58,80%	40,88%	33,53%	72,91%
KNN	34,92%	64,81%	51,76%	26,35%	86,29%
SVM	27,35%	63,52%	47,76%	19,16%	88,29%
BODDS					
Decision Tree	58,36%	55,59%	55,08%	62,05%	49,09%
Random Forest	60,66%	60,42%	60,48%	60,84%	60,00%
AdaBoost	61,80%	58,91%	57,89%	66,27%	51,52%
GBM	59,57%	59,82%	60,12%	59,04%	60,61%
XGBoost	55,28%	56,50%	57,05%	53,61%	59,39%
KNN	54,05%	53,78%	53,89%	54,22%	53,33%
SVM▼	62,01%	58,91%	57,81%	66,87%	50,91%

▼ Best model of the experiment; • Best model among the experiments.

Source: Produced by the author.