Early Cardiovascular Risk Prediction in Quilombola Afro-descendants: A Data-Driven Approach

J.S.L. Figuerêdo[1,*], R.S. Rosa[1,+], R.N.S.O. Boery[2], J.B. Júnior[1,+], R.T. Calumby[1,*]

[1] University of Feira de Santana, Brazil
[*]Postgraduate Program in Computer Science
[+]Postgraduate Program in Collective Health
jslfigueredo@ecomp.uefs.br, enfrandson@gmail.com, bessa@uefs.br, rtcalumby@uefs.br
[2] University of Southwest Bahia
Postgraduate Program in Nursing and Health
rboery@gmail.com

Abstract. Cardiovascular diseases (CVD) are the leading cause of global mortality. People from different social groups can be affected. However, socially vulnerable groups, such as the Quilombola communities in Brazil, may have an increased risk. Recently, sample data from this population were used to predict metabolic syndrome with machine learning(ML). Although metabolic syndrome is a risk factor for CVD, directly predicting cardiovascular risk itself might be more effective for implementing preventive strategies. Therefore, this study developed and assessed ML models to estimate CVD risk, including a variable importance analysis. Most models achieved over 80% effectiveness, with logistic regression achieving the best result. Considering the variable importance analysis, sex, age and income were identified as the most important variables, along with other socioeconomic and anthropometric data.

CCS Concepts: • **Computing methodologies** → **Machine learning algorithms**.

Keywords: Cardiovascular Risk Prediction; Computer aided prognosis; Quilombola; Framingham score.

1. INTRODUCTION

Cardiovascular diseases (CVD) are the leading cause of death worldwide, claiming about 17.9 million lives each year. They comprise a group of disorders of the heart and blood vessels and include coronary heart, cerebrovascular, rheumatic heart diseases, heart attack, and other conditions [World Health Organization]. In Brazil, CVD also stand as the leading cause of death. Every year, thousands of Brazilians die due to these diseases [Malta et al. 2021]. The growth in prevalence numbers of CVD has engaged the healthcare system in developing actions to control and prevent risk factors [Costa and Thuler 2012]. Several factors influence the genesis of CVD, usually categorized into modifiable and non-modifiable risk factors [Day and Goldlust 2010]. In the former, include dyslipidemia, smoking, high blood pressure, diabetes, obesity, physical inactivity, unhealthy diets, stress, and the use of contraceptives; in contrast, the latter encompasses family history of CVD, age, sex, and race.

The complexity inherent in the etiology of CVD is characterized by risk factors associated with genetic, social, cultural, and economic aspects. Nevertheless, specific populations, such as the Brazilian quilombola communities, may be more susceptible to these diseases due to social vulnerabilities. Quilombola communities originated from the escape of enslaved individuals who formed socially organized groups in opposition to the abuses endured during the slavery period [Dorigny 2017]. However, even after the abolition of slavery in Brazil, immediate public policies addressing the needs of the Black population were not implemented, leaving them without social or legal assistance to facilitate their equitable integration into society [Tanure 2021]. This contributed to the persisting inequities and social and health disparities that impact the lifestyles, mortality, and morbidity of this population group, mainly to quilombola Afro-descendants communities [Torres et al. 2023].

Individuals in Quilombola communities commonly share their daily living and health conditions, as well as inherited knowledge, attitudes, beliefs, cultures, and alternative health practices. They often face social vulnerability and healthcare access challenges, creating barriers to diagnostic, clinical, therapeutic, and rehabilitative services, which exacerbate chronic illnesses like CVD [Rosa et al. 2021]. For instance, studies indicate that African-American populations, who share ancestral origins with Brazilian quilombola communities, have a higher incidence and risk of premature death from coronary artery disease compared to the general population [Mozaffarian et al. 2015; Oshunbade et al. 2021]. In order to reduce the CVD effects, identifying the patient's cardiovascular risk has been used. Risk prediction stands out as a primary factor in prevention, with several risk estimation approaches have been developed over the years, such as the Framingham Risk Score (FRS) [Dawber 1980]. The FRS is a gender-specific algorithm used to estimate an individual's cardiovascular risk over 10 years. Once the risk is identified, strategies such as dietary changes and encouragement of physical activity can be devised to prevent or minimize its impact. To further enhance the risk estimation process, approaches utilizing artificial intelligence (AI) techniques can be applied.

AI systems have been utilized to aid in the diagnosis and prognosis of a wide range of medical conditions [Yu et al. 2018; Ventura et al. 2022]. The primary objective of these systems is to enhance healthcare decision-making in disease diagnosis, treatment, and maintenance [Lobo 2017], mainly with ML models. For CVD, these models can estimate risk based on socio-economic, demographic, lifestyle, and health condition factors. Several studies have addressed this problem [Li et al. 2020; Bhardwaj et al. 2023; Torres et al. 2023]. Torres et al. (2023) conducted the first study involving the quilombola community and ML in Brazil. The objective was to forecast the probability of quilombolas developing metabolic syndrome. Despite being pioneering, the study has some limitations. For example, only one decision tree was utilized. While decision trees are simple, they can perform poorly with highly non-linear, imbalanced, or irrelevant-feature-heavy data. Additionally, the authors use an optimistic validation process, evaluating the model's effectiveness with only one validation set. Although this approach is common, it can lead to misinterpretations, especially when applied to small datasets. Ultimately, due to the stochastic nature of data partitioning, the model's effectiveness may be limited to that specific subset. Therefore, new approaches should be proposed to enable better predictions and validated with rigorous protocols for reliable conclusions in similar scenarios.

Considering the aforementioned challenges, we developed and assessed ML models to predict cardiovascular risk of the Quilombola population based on the Framingham score. Although Torres et al. (2023) also studied the quilombola population, their objectives and methods were different, as they evaluated metabolic syndrome, which is a risk factor, but not the risk score itself. In this sense, for our knowledge, our work establishing it as a pioneering investigation in Brazil. For this purpose, it was utilized data from a Quilombola community in the interior of Bahia, Brazil. To conducted this data-driven study, we considered basic individual information such as gender and age, socioeconomic data, anthropometric measurements, among others. We also performed a variable importance estimation to understand the overall profile of individuals with cardiovascular risk.

2. RELATED WORKS

Considering the complexity of cardiac diseases, a system based on ML was proposed to assist in their diagnosis [Li et al. 2020]. Specifically, the authors applied six techniques: support vector machine, logistic regression, artificial neural network, k-nearest neighbor, naïve bayes and decision tree. Additionally, several feature selection strategies were evaluated, including a new approach developed by the authors. In general, all the evaluated models achieved high effectiveness, especially the support vector machine, when combined with the feature selection proposed. In another study, Silva et al.

(2022) developed a method for detecting cardiac arrhythmias in electrocardiogram signals using a graph convolutional network (GCN). The authors modeled each heartbeat as a node in a graph and employed a GCN to classify these nodes. The database utilized mapped three types of beats: normal heartbeats (N), supraventricular ectopic heartbeats (S) and ventricular ectopic heartbeats (V). The experiments indicated promising results for classes N and S, both with 100.0% positive prediction. However, the model presented low effectiveness for class V (41.2%).

Bhardwaj et al. (2023) designed an ensemble-based ML model to predict the risk of CVD. The ensemble consisted of five models: neural network, random forest, bayesian network, and decision trees (C5.0 and QUEST). The authors' ensemble utilized a majority voting scheme for decision-making. The model was generated from a dataset comprising $70,000$ patient records with 11 features. The results indicated a higher accuracy of the ensemble model (99.1%) compared to individual ones. Based on these results, the authors emphasized the potential for ML algorithms to aid in early disease identification and enhance treatment outcomes. Finally, Torres et al. (2023) developed a model in order to predict the probability of quilombolas developing metabolic syndrome. To perform the prediction, the authors used a decision tree based on anthropometric data collected from a Brazilian quilombola community. The developed model achieved 75% of accuracy on the test set.

Despite recent advances, there are still some limitations to be addressed. Except for the work developed by Torres et al. (2023), the models were not specifically designed for quilombolas, a vulnerable population with unique characteristics requiring specific calibration. Although Torres et al. (2023) used ML for quilombolas, it focused on metabolic syndrome. Thus, prediction based on a specific risk score like the FRS need to be explored. Additionally, the evaluation used only one ML algorithm, a decision tree. While decision trees are simple and interpretable, exploring more robust algorithms is crucial. Ultimately, these algorithms are likely to discover patterns that could be missed by the decision tree. Furthermore, the authors used optimistic validation protocols, which could produce unreliable results. In contrast, we evaluated several algorithms, including decision tree, kNN, logistic regression, random forest, and XGBoost, to predict cardiovascular risk in quilombolas using the FRS and robust validation protocols. We also analyzed the variables that best characterize patients with cardiac risk.

3. EXPERIMENTAL PIPELINE

The experimental process used in this work is illustrated in Figure 1. There are four stages: Data Collection and preprocessing (Section 3.1), model training (including optimization and validation) Section 3.2), and, finally, the evaluation and analysis of the model (Section 3.2).

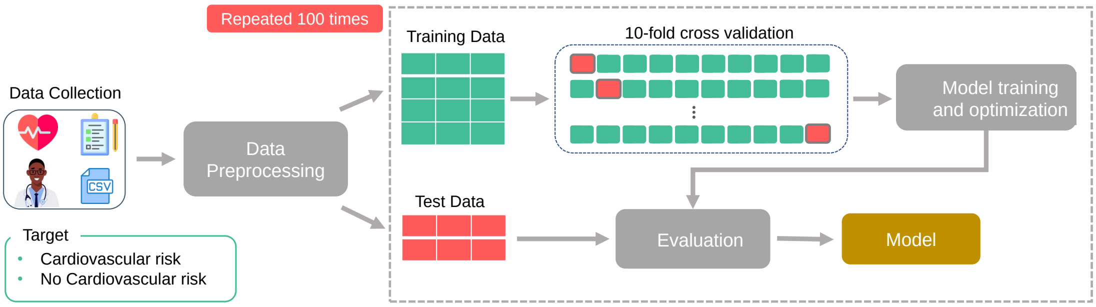


Fig. 1: Experimental Benchmark Workflow.

3.1 Dataset and Data Preprocessing

The dataset utilized consists of data collected from 303 quilombolas residing in the quilombola community of Barro Preto, located in the municipality of Jequié, Bahia, Brazil. The data were obtained through an epidemiological, census-based, cross-sectional, and community-based study [Rosa et al. 2021]. Three validated and/or recommended instruments for the Brazilian context were employed. The first instrument applied was a Primary Health Care Hypertension Questionnaire. The second corresponded to the eight-item Morisky Therapeutic Adherence Scale. Lastly, the third instrument was related to the FRS, utilizing a form for cardiovascular risk and associated factors, following guidelines from the specialized authorities [SBC et al. 2010]. Specifically, the FRS was calculated using the following variables: age, sex, diabetes, smoking, treated and untreated blood pressure, LDL and HDL cholesterol. They were summed according to their assigned points at each step of the score to estimate low risk, with values below 10%; medium risk, between 10% and 20%; and high risk, equal to or greater than 20%.

In our study, the FRS was used as a heuristic for determining cardiovascular risk, essentially defining the ground-truth. The FRS computation involved the variables of age, sex, diabetes, smoking, treated and untreated blood pressure, LDL and HDL cholesterol. Based on the calculated value, individuals were classified into one of three risk levels: low, medium, or high. Considering the medium and high levels as the most significant, we mapped these two categories as cardiovascular risk, while the low category was used as no cardiovascular risk. This step transformed the originally multiclass problem into a binary one. Binary classification in ML offers numerous advantages compared to multiclass problems. These include simplicity, efficiency, interpretability, and the capability to better address class imbalance, making it advantageous in many applications [Bishop 2006]. It is noteworthy that some of the attributes used in FRS calculation were excluded: diabetes, smoking, treated and untreated blood pressure, LDL and HDL cholesterol. These variables were removed to keep only low-cost variables and to reduce their dependency on the target attribute, as they were initially used to determine the target attribute.

The initial dataset comprised 48 attributes. However, a preliminary analysis showed that some of these variables do not add relevant predictive content, e.g., family composition. we also engaged the collaboration of an expert who selected the features deemed most relevant to our context. Thus, some variables considered non-relevant for the task were removed, resulting in a final set of 22 variables (including the outcome attribute)[1]. Additionally, during the preprocessing step, some attributes with missing data were identified, namely: income, number of medicines, amputation due to diabetes, alcoholism and triglycerides. From these attributes, we decided to remove the triglycerides, given the high percentage of missing data (60% higer). In contrast, imputation was performed for the remaining attributes. Mean imputation was applied to the "income" variable, mode imputation was used for "amputation due to diabetes". For the variables "number of medicines" and "alcoholism", imputation based on value was conducted. For these two cases, we considered a scenario in which the person do not take medication to control hypertension (value 0) and was not an alcoholic (value 1). An adjustment was also necessary to the heart rate variable, as it had been collected at two different time points, generating two features. Therefore, the mean was calculated, which resulted in a new feature named average heart rate.

3.2 Experimental Setup and Effectiveness Assessment

This study used five ML algorithms: decision tree, kNN, logistic regression, random forest and XGBoost. To carry out the experiment process, the dataset was partitioned into training and test sets. It was randomly partitioned using the 75/25 ratio, with 75% for training and the remaining 25% for the

[1] The list of the variables used can be found at `https://figshare.com/articles/conference_contribution/26490031?file=48157048`

test set. This process was performed 100 times for each algorithm. The training set was used to build the models that were optimized via hyperparameter optimization[2] through grid search, using cross-validation based on k-folds, with $k = 10$ and considering the F_1 as maximization criteria. The test set is used to evaluate the overall effectiveness of the models. As the experimental process presented in Figure 1 was executed 100 times for each algorithm, the result reported in this work corresponds to the average of these executions.

The effectiveness assessment was carried out using classical ML measures, such as *Precision*, *Recall* and F_1. *Precision* quantifies the portion of samples correctly predicted as belonging to the class of interest (Cardiovascular risk). On the other hand, *Recall* quantifies the portion of samples of the class of interest the were correctly predicted as belonging to that class. Finally, the F_1 measure is taken as the harmonic mean of *Precision* and *Recall*. For the strict comparison of the effectiveness results, the developed models were compared using Wilcoxon's Signed Rank Test in order to assess the statistical significance of the results.

4. RESULTS AND DISCUSSION

4.1 Effectiveness Analysis

Table I presents the effectiveness of the models developed, considering the *Precision*, *Recall* and F_1. Additionally, the standard deviation for each algorithm is provided. Except for the kNN, the models demonstrated promising results, achieving effectiveness over 75% across all evaluated measures. Among the evaluated algorithms, logistic regression achieved the best result, obtaining $F_1 = 0.8370$. Recognizing cardiovascular risk is crucial for improving clinical management and controlling chronic cardiovascular diseases. Thus, these models can aid in ensuring patients receive necessary healthcare services, especially in minority groups and lower socioeconomic populations like the quilombolas.

Since logistic regression achieved the best results, we selected it as the baseline for statistical tests. According the tests, logistic regression outperformed other algorithms in all measures, except for the decision tree (*Precision*) and the random forest (*Recall*). Notably, except for kNN and the decision tree, the models achieved *Recall* over 80%. These results are beneficial, as *Recall* indicates greater coverage of individuals with cardiovascular risk. This is relevant for developing strategies to minimize or prevent the onset or worsening of CVD. For instance, clinical interventions supported by multidisciplinary approaches could be implemented, incorporating patient education on risk awareness and diagnostic measures (laboratory tests), such as clinical monitoring of lipid profile parameters, glycemic control, and risk behavior variables. We emphasize that these predictive models should complement, not replace, basic patient care. Healthcare practitioners must still conduct personalized analyses and compare them with model outcomes to ensure accurate and relevant patient care decisions.

Table I: Prediction effectiveness of the developed models in terms of *Precision*, *Recall* and F_1. The statistically significant superiority of the evaluated algorithms in relation to the baseline (logistic regression) is highlighted in italic. In turn, statistically significant inferiority is marked with "*". The remainder indicates statistical equivalence

Algorithm/Classifier	Precision	Recall	F_1
Logistic Regression	0.8165 ±0.0547	0.8616 ±0.0470	0.8370 ±0.0385
Random Forest	0.8031* ±0.0571	0.8565 ±0.0535	0.8271* ±0.0400
XGBoost	0.8322* ±0.0553	0.8096* ±0.0665	0.8182* ±0.0416
Decision Tree	*0.8500*±0.0600	0.7620* ±0.0581	0.8014* ±0.0424
kNN	0.7190* ±0.0587	0.7343* ±0.0701	0.7237* ±0.0457

[2] The Hyperparameters tested can be found at `https://figshare.com/articles/conference_contribution/26405710`

4.2 Variable Importance Analysis

Figure 2(a-e) illustrates the ten most significant variables for predicting cardiovascular risk using the top-5 models trained from the XGBoost algorithm, based on *Recall*. We used the XGBoost due to its relatively lower interpretability among the evaluated algorithms. Considering the selected models, the attributes sex, age, and income were present in all of them. This result aligns with observations in the literature, especially concerning sex and age [Rodgers et al. 2019]. This result positions ML as a alternative strategy for predicting cardiovascular risk, alongside other approaches. Some studies also indicate the relationship between income and cardiac risk. For example, in a multi-ethnic study named as Jackson Heart Study, considered the largest cohort study on CVD among African-Americans, socio-economic conditions such as low income and residing in socioeconomically disadvantaged areas were considered factors associated with cardiovascular risk. These aspects affects the quilombola population, making it challenging to adhere to therapeutic measures and promote healthy lifestyles. It is noteworthy that characteristics such as family benefit[3], waist circumference, BMI, and educational background were also highlighted as significant variables for predicting risk.

Depending on the ML algorithm employed, the importance of features can vary, as each algorithm utilizes a different approach to build the predictive model. Different approaches may prioritize certain variables over others. It is emphasized that this discrepancy can occur even among models generated by the same algorithm when using different hyperparameters. However, this difference can be more evident when considering models generated by different algorithms. To illustrate this, we selected the best model obtained by the random forest, also based on *Recall* (Figure 2-f). Similar to XGBoost, age and sex emerge as the most important features, but factors related to socioeconomic conditions and anthropometric variables are also highlighted (Education, BMI, Weight, Stature, among others), presenting greater importance than in XGBoost.

4.3 Sex and Age Dependence Analysis

For the generation of predictive models, some of the variables used in the calculation of the FRS were removed. Sex and age were preserved due to their low-cost acquisition and importance to the cardiovascular risk prediction. However, given their prior utilization in the FRS computation, an initial analysis was conducted to assess their impact on the predictive model's effectiveness. For this analysis, a decision tree was chosen given to its simplicity and low training overhead. The effectiveness measures for this model were: *Precision* (0.6882), *Recall* (0.8086) and F_1(0.7380). Compared to the decision tree with age and sex, its ability to accurately identify individuals without risk decreased, impacting *Precision* by increasing false positives. Conversely, an improvement in *Recall* was observed, indicating that the model enhanced the estimation process for patient with cardiovascular risk.

To illustrate feature behavior, a variable importance was computed using the same data sample for both models. Figure 3-a shows that sex and age are the most influential variables. On the other hand, their exclusion accentuates the importance of other variables, notably *Education*. This suggests that without sex and age, other factors are used for risk assessment. Indeed, education level is recognized as correlated with cardiovascular risk. In Borhanuddin et al. (2018) was identified a greater risk of cardiovascular disease development among individuals with lower educational levels. Insights from the educational variable have significant implications for monitoring and evaluating the health status of these individuals, revealing disparities in health access based on race and ethnicity. This phenomenon can be attributed to individuals with lower educational backgrounds, who usually possess a more superficial understanding of their health and disease conditions. Consequently, it is plausible that these individuals engage with medical and health services less frequently, thereby incurring an elevated risk of cardiovascular diseases. While this initial analysis of sex and age impact has been done, future research with different algorithms is needed for deeper insights.

[3] A Brazilian governmental program for income distribution to low-income families

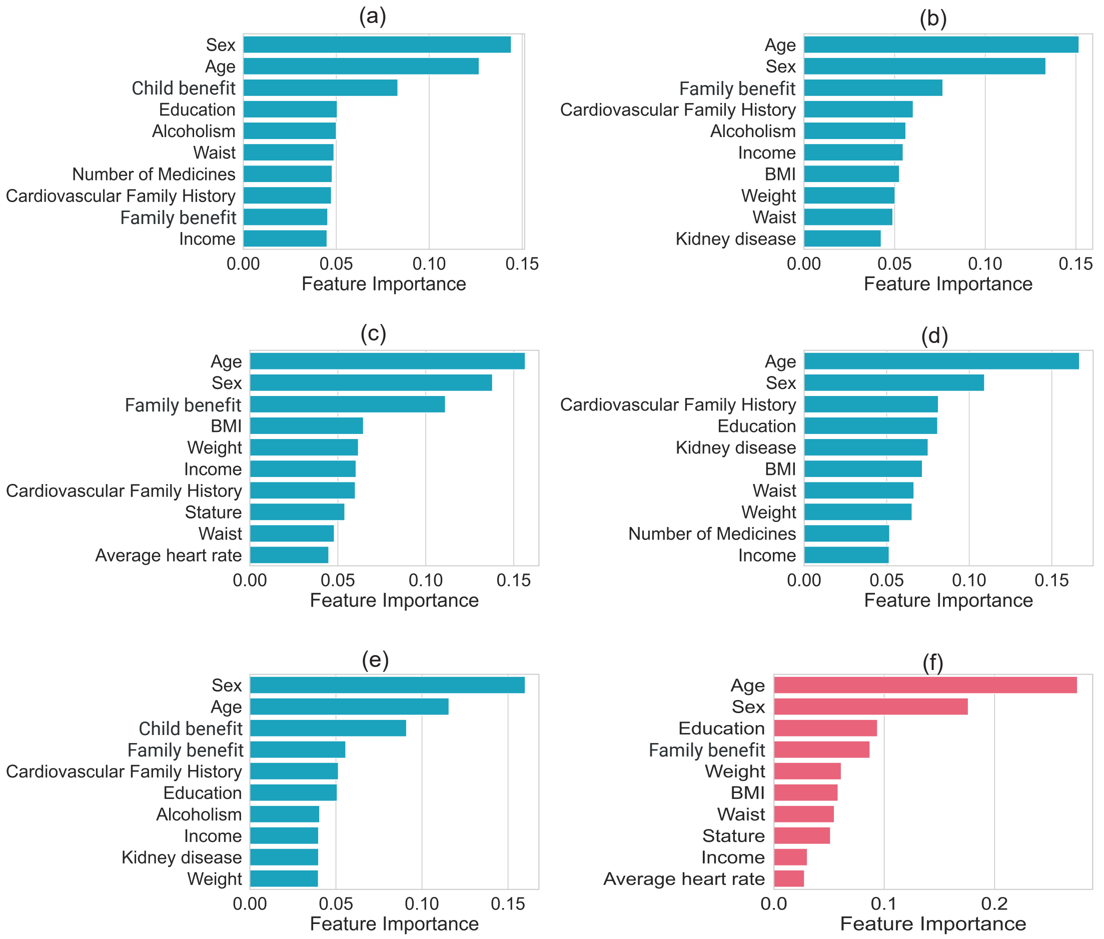
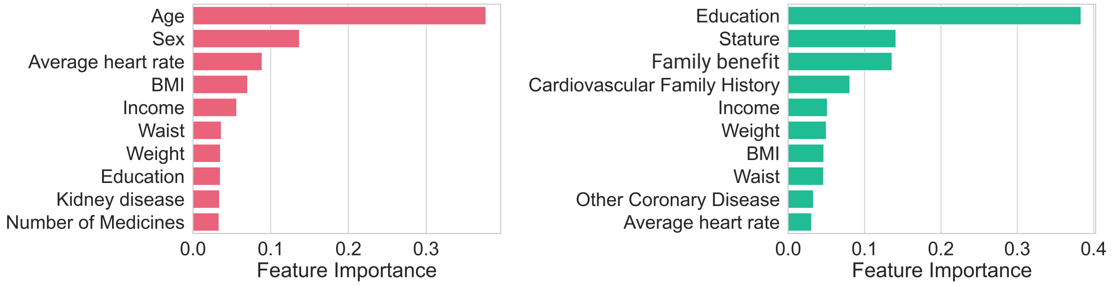


Fig. 2: Variables considered most important: from (a) to (e) considering XGBoost. In (f) considering Random Forest



(a)

(b)

Fig. 3: Variables considered most important: (a) Decision tree with age and sex included. (b) Decision tree after remove age and sex

5. CONCLUSION

In this study, we developed and experimentally evaluated models to assist in predicting the cardiovascular risk of the quilombola Afro-descendants population based on machine learning algorithms.

Specifically, we used decision tree, kNN, logistic regression, random forest and XGBoost. Additionally, a variable importance analysis was performed. Experimental findings identified logistic regression as the most effective model, achieving results over 80% for all considered measures. Among the evaluated models, kNN achieved lowest effectiveness. Through XGBoost, the variable importance analysis highlighted sex, age and income as the most significant factors for identifying individuals at risk of developing cardiovascular diseases. Furthermore, other socioeconomic and anthropometric factors were also highlighted. This became even more evident in an alternative experiment where a decision tree model was generated without considering age and sex. It was verified that, even without considering these two features, the model could estimate the risk, indicating that other variables, such as education level, are also relevant in cardiovascular risk estimation. In future work, we aim to conduct new experiments using a larger dataset. Moreover, we plan to undertake a more comprehensive analysis of the relationship between some variables and cardiovascular risk. Additionally, other variables, such as the marital status and laboratory tests, will be incorporated into the predictive model.

ACKNOWLEDGMENTS

This work was partially supported by UEFS AUXPPG 2024 and CAPES PROAP 2024 grants.

REFERENCES

Bhardwaj, A. et al. Application of machine learning for cardiovascular disease risk prediction. *Computational Intelligence and Neuroscience* vol. 2023, 2023.

Bishop, C. M. *Pattern Recognition and Machine Learning.* Springer-Verlag, Berlin, Heidelberg, 2006.

Borhanuddin, B. et al. 10-year cardiovascular disease risk estimation based on lipid profile-based and bmi-based framingham risk scores across multiple sociodemographic characteristics: The malaysian cohort project. *Scientific World Journal* vol. 2018, pp. 2979206, 2018.

Costa, L. C. and Thuler, L. C. S. Fatores associados ao risco para doenças não transmissíveis em adultos brasileiros: estudo transversal de base populacional. *Revista Brasileira de Estudos de População* 29 (1): 133–145, Jan, 2012.

Dawber, T. *The Framingham Study: The Epidemiology of Atherosclerotic Disease.* HUP, Cambridge, 1980.

Day, T. E. and Goldlust, E. Cardiovascular disease risk profiles. *American Heart Journal* 160 (1): e3, 2010.

Dorigny, M. *Atlas das escravidões: da Antiguidade até os nossos dias.* Vozes, Petrópolis (RJ), 2017.

Li, J. P. et al. Heart disease identification method using machine learning classification in e-healthcare. *IEEE Access* vol. 8, pp. 107562–107582, 2020.

Lobo, L. C. Inteligência artificial e medicina. *Revista Brasileira de Educação Médica* 41 (2): 185–193, Apr, 2017.

Malta, D. C. et al. Estimativas do risco cardiovascular em dez anos na população brasileira: Um estudo de base populacional. *Arquivos Brasileiros de Cardiologia* 116 (3): 423–431, Mar, 2021.

Mozaffarian, D. et al. Heart disease and stroke statistics–2016 update: A report from the american heart association. *Circulation* vol. 133, pp. e38–e360, 2015.

Oshunbade, A. A. et al. Cigarette smoking, incident coronary heart disease, and coronary artery calcification in black adults: The jackson heart study. *Journal of the American Heart Association* 10 (7): e017320, 2021.

Rodgers, J. L. et al. Cardiovascular risks associated with gender and aging. *JCDD* 6 (2): 19, 2019.

Rosa, R. S. et al. Cardiovascular Risk and Factors Associated to the Health in Hypertensive African Descent People Resident in Quilombola Community. *Revista Cuidarte* vol. 12, 00, 2021.

SBC, SBH, and SBN. Vi diretrizes brasileiras de hipertensão. *Arq Bras Cardiol* 95 (1 suppl.1): 1–51, 2010.

Silva, G. et al. Cardiac arrhythmia detection in ecg signals using graph convolutional network. In *Anais do XXII SBCAS.* SBC, Porto Alegre, RS, Brasil, pp. 25–35, 2022.

Tanure, R. G. A. Da necessidade da implementação de políticas públicas no combate ao racismo estrutural. *Boletim Científico ESMPU* 20 (57): 265–284, 2021.

Torres, R. C. et al. Modelo matemático para prever probabilidade de quilombolas desenvolverem síndrome metabólica com fluxograma de atendimento de saúde. *Scientia Plena* 19 (8), set., 2023.

Ventura, N., , et al. An interpretable classification model for identifying individuals with attention defict hyperactivity disorder. In *Anais do X KDMiLe.* SBC, Porto Alegre, RS, Brasil, pp. 66–73, 2022.

World Health Organization. Cardiovascular Diseases. `https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1`. Accessed: 2024-03-04.

Yu, K.-H. et al. Artificial intelligence in healthcare. *Nature Biomedical Engineering* 2 (10): 719–731, Oct, 2018.