Cardiovascular Disease Prediction by Machine Learning: Predictive Models and Optimization

Navid Azimi*
Emory University
navid.azimi@emory.edu

Masoud Nateghi*
Emory University
masoud.nateghi@emory.edu

October 13, 2023

Abstract

Cardiovascular disease (CVD) is characterized as a malfunction of the human heart and blood vessels, potentially resulting in physical impairment or loss of life. While the exact causes of CVD remain unknown, it is acknowledged that it may be linked to various risk factors. Machine learning has emerged as a promising tool within the healthcare domain, offering the potential to advance our comprehension of CVD risk factors and enhance predictive precision. This project seeks to tackle the challenges of predicting CVD by leveraging the capabilities of machine learning models. In our research, we will thoroughly investigate and assess an extensive array of machine learning models, employing various metrics and evaluation techniques.

Keywords: Cardiovascular Disease, Machine Learning, Binary Classification

1 Introduction

The prevalence of Cardiovascular diseases (CVD) has been rising alarmingly as people's living standards rise and their stress levels increase. Cardiovascular diseases continue to be one of the leading causes of death in the world [1]. By 2030, Cardiovascular diseases are expected to be the cause of death for about 23 million people, according to the latest predictions [2] [3]. Several factors, such as age, gender, height, weight, the BMI index, as well as the results of blood tests that evaluate renal function, liver function, and cholesterol levels, can have an impact on the incidence of cardiovascular disease [4] [5]. The development of a wide variety of health problems can be influenced by the complex interactions that take place between such factors.

Therefore, Machine Learning (ML) algorithms can help to predict diseases at an early stage and treat the patient accordingly. They are more adept than standard statistical models at capturing the complex interactions and nonlinear linkages that exist between the variables and the results [6]. Various machine learning techniques, including Support Vector Machines [7], Neural Networks, Decision Trees [8], and K-Nearest Neighbour (K-NN)] [9], show distinct advantages and limitations. The results of each technique differ owing to several constraints. Although these techniques have found applications in various domains like human heart diseases (echocardiogram signals) [10] [11], observations from related studies reveal further scope for the development of CVD prediction models using other ML algorithms, which can result in improved performance.

2 Dataset

In this project, we will use an openly accessible dataset [12]. This dataset contains 68,205 data entries and comprises 16 attributes, with 2 of them being categorical and the remaining 14 being numerical. More specifically, the categorical attributes are denoted as "bp_category" and "bp_category_encoded". The numerical attributes encompass "ID", "age", "age_years", "gender", "height", "weight", "ap_hi", "ap_lo", "cholesterol", "gluc",

*Equal Contribution

"smoke", "alco", "active", and "bmi". The target attribute, referred to as "cardio," is binary and indicates the presence or absence of cardiovascular disease. Additional details about the attributes can be found in Table 1.

Attributes	Description
ID	Unique identifier for each patient
age	Age of the patient (Days)
age_years	Age of the patient (Years)
gender	Gender of the patient (1: Female, 2: Male)
height	Height of the patient (cm)
weight	Weight of the patient (kg)
ap_hi	Systolic blood pressure
ap_lo	Diastolic blood pressure
cholesterol	Cholesterol levels (1: Normal, 2: Above Normal, 3: Well Above Normal)
gluc	Glucose levels (1: Normal, 2: Above Normal, 3: Well Above Normal)
smoke	Smoking status (0: Non-smoker, 1: Smoker)
alco	Alcohol intake (0: Does not consume alcohol, 1: Consumes alcohol)
active	Physical activity (0: Not physically active, 1: Physically active).
cardio	Presence or absence of cardiovascular disease (0: Absence, 1: Presence)
bmi	Body Mass Index
bp_category	Blood pressure category
bp_category_encoded	Encoded form of bp_category for Machine Learning purposes

Table 1: An overview of the dataset attributes

Our initial data examination revealed the absence of null and duplicated rows in the dataset. The dataset includes continuous attributes such as "age," "height," "weight," "ap_hi," "ap_lo," "age_years," and "bmi." The statistical characteristics of these continuous attributes are outlined in Table 2. Notably, Table 2 highlights that the minimum and maximum values of the "height" and "weight" attributes significantly deviate from the rest of the data, suggesting the need for further investigation to address such outliers.

	age	age_years	height	weight	ap_hi	ap_lo	bmi
Mean	19462.66	52.82	164.37	74.10	126.43	81.26	27.51
Standard Dev.	2468.38	6.76	8.17	14.28	15.96	9.14	6.02
Minimum	10798.00	29.00	55.00	11.00	90.00	60.00	3.47
Maximum	23713.00	64.00	250.00	200.00	180.00	120.00	298.66

Table 2: An overview of the dataset's continuous attributes

3 Methods

This project's objective is to conduct binary classification on a CVD dataset, allowing us to train various models and gather valuable insights regarding their performance in addressing the given problem. Additionally, we will explore various preprocessing techniques and feature selection methods for further improvement of the final results.

3.1 Data Preprocessing

Preprocessing stands out as a pivotal stage in every machine learning pipeline. As we observed in one of our previous assignments, the way we preprocess and transform data can significantly impact the outcomes and the quality of classification. We will delve into the impacts of various preprocessing methods and document which one works best for each model. Moreover, we will assess the influence of Principal Component Analysis (PCA) as a preprocessing technique to provide a more comprehensive analysis.

3.2 Feature and Model Selection

In this project, we will dive deeper into the effect of various feature selection methods. These methods can be categorized as unsupervised, such as analyzing feature correlations to eliminate highly correlated features due to their similarity, or supervised, such as ranking features based on correlations (Spearman, Pearson, Kendall Tau), mutual information, Fisher score, forward/backward feature addition/elimination, and evolutionary algorithms.

Throughout this project, we will assess different classification methods, including Logistic Regression, K-Nearest Neighbors (K-NN), Gradient Boosting algorithms, Decision Trees, Random Forests, Neural Networks, and Support Vector Machines (SVM). If our schedule allows, we will also explore the potential benefits of ensemble models. The concept of combining weak learners may lead to a more robust classifier, and we will investigate the feasibility and advantages of this approach.

3.3 Model Assessment and Optimization

For this phase, we begin by meticulously splitting the data into distinct training and testing datasets. We adopt a robust approach to assessing our model's performance on the training dataset and fine-tuning the hyperparameters by employing k-Fold Cross-Validation. The choice of the Area Under the Curve (AUC) serves as our primary metric for comparing and contrasting the performance of various models, offering a comprehensive evaluation. We further supplement our model assessments with comprehensive reporting of key metrics, including the F1-score, Recall, Precision, and Accuracy.

Subsequently, our focus shifts to evaluating the model's real-world applicability and reliability. We conduct this assessment using the test dataset, under the assumption that it accurately mirrors the characteristics of the target population in a real-world scenario. This step is crucial in ensuring that the models' performance remain consistent and reliable beyond the training data.

4 Conclusion

The primary objective of our research is to provide valuable insights into the predictive capabilities of machine learning algorithms for CVD. By systematically evaluating these models, we aim to identify which approaches demonstrate the highest predictive accuracy and reliability. Ultimately, the findings of this research can serve as a foundation for the development of more effective CVD prediction tools, enabling early intervention and personalized healthcare strategies to mitigate the impact of cardiovascular diseases. Figure 1 illustrates our project schedule in a Gantt chart.

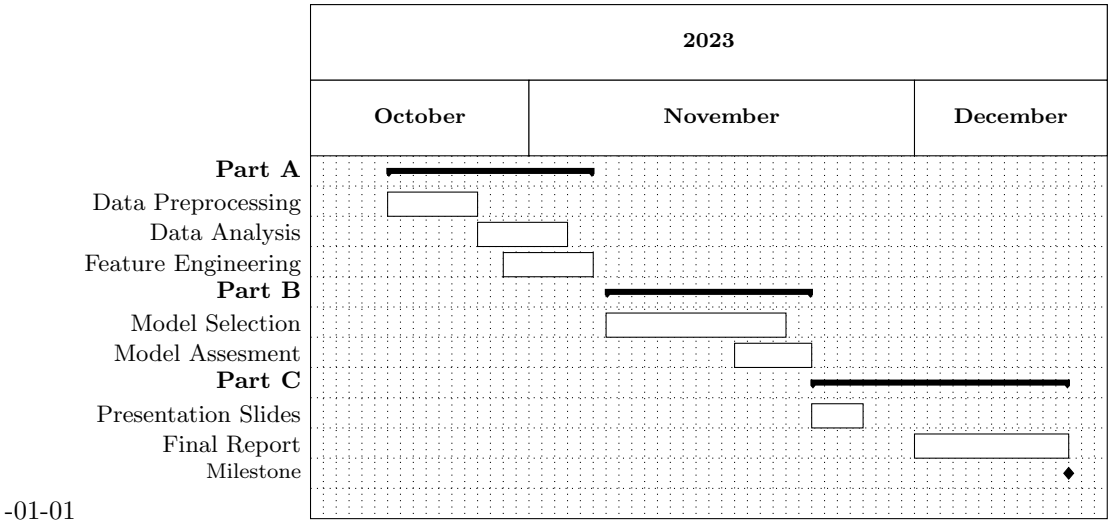

Figure 1: Project Schedule

References

[1] Aryal S, Alimadadi A, Manandhar I, Joe B, Cheng X. (2020). Machine learning strategy for gut microbiome-based diagnostic screening of cardiovascular disease. *Hypertension, 76(5), 1555-1562.*

[2] Juhola M, Joutsijoki H, Penttinen K, Aalto-Setälä K. (2018). Detection of genetic cardiac diseases by Ca2+ transient profiles using machine learning methods. *Scientific reports, 8(1), 9355.*

[3] Maheshwari, V., Mahmood, M. R., Sravanthi, S., Arivazhagan, N., ParimalaGandhi, A., Srihari, K., ... & Sundramurthy, V. P. (2021). Nanotechnology-based sensitive biosensors for COVID-19 prediction using fuzzy logic control. *Journal of Nanomaterials, 2021, 1-8.*

[4] Maiga, J., & Hungilo, G. G. (2019, October). Comparison of machine learning models in prediction of cardiovascular disease using health record data. In *2019 international conference on informatics, multimedia, cyber and information system (ICIMCIS) (pp. 45-48). IEEE.*

[5] Sivasankari, S. S., Surendiran, J., Yuvaraj, N., Ramkumar, M., Ravi, C. N., & Vidhya, R. G. (2022, April). Classification of diabetes using multilayer perceptron In *2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) (pp. 1-5). IEEE.*

[6] Alaa, A. M., Bolton, T., Di Angelantonio, E., Rudd, J. H., & Van der Schaar, M. (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. *PloS one, 14(5), e0213653.*

[7] Yan, H., Ye, Q., Zhang, T. A., Yu, D. J., Yuan, X., Xu, Y., & Fu, L. (2018) Least squares twin bounded support vector machines based on L1-norm distance metric for classification. In *Pattern recognition, 74, 434-447.*

[8] Jaworski, M., Duda, P., & Rutkowski, L. (2017). New splitting criteria for decision trees in stationary data streams. In *IEEE transactions on neural networks and learning systems, 29(6), 2516-2529.*

[9] Zhang, S., Cheng, D., Deng, Z., Zong, M., & Deng, X. (2018). A novel kNN algorithm with data-driven k parameter computation. *Pattern Recognition Letters, 109, 44-54.*

[10] Pławiak, P. (2018). Novel genetic ensembles of classifiers applied to myocardium dysfunction recognition based on ECG signals. *Swarm and evolutionary computation, 39, 192-208.*

[11] Pławiak, P. (2018). Novel methodology of cardiac health recognition based on ECG signals and evolutionary-neural system. In *Expert Systems with Applications, 92, 334-349.*

[12] Cardiovascular Disease Dataset, predict the presence or absence of cardiovascular disease. URL `https://www.kaggle.com/datasets/colewelkins/cardiovascular-disease/data`.