CARDIOVASCULAR DISEASE PREDICTION

By

Charith Govardhanam

Aditya Nayak

Presented to the

Data Science Department

Columbian College of Arts & Sciences

George Washington University

Under the Supervision of

Prof. Abdi Awl, D.Eng

Table of Contents

I. Introduction

a. Background

The human heart is only the size of a fist, but it is the hardest working muscle in the body. With every heartbeat, the heart pumps blood, carrying oxygen and nutrients to every part of the body. The heart is a muscular organ that pumps blood throughout the body via the circulatory system, which is made up of the heart, blood and blood vessels. The pumped blood carries oxygen and nutrients to tissues and organs through the blood vessels, while carrying away metabolic waste such as carbon dioxide. In humans, the heart is around the size of a large fist and sits between the lungs, in the middle compartment of the chest, slightly to the left of center. The heart beats around 100,000 times and pumps up to 7,500 liters of blood every day.

Cardiovascular disease (CVD) is a class of diseases that affects the heart or blood vessels (veins and arteries). It can be caused by a combination of socio-economic, behavioral, and environmental risk factors, including high blood pressure, unhealthy diet, high cholesterol, diabetes, air pollution, obesity, tobacco use, kidney disease, physical inactivity, harmful use of alcohol and stress. Family history, ethnic background, sex, and age can also affect a person’s risk of cardiovascular disease.

b. Problem Statement

Thus by taking all the issues at hand in consideration, we came up with a problem statement of Identifying individuals at risk of CVD that remains challenging due to limitations in current risk prediction models, which often rely on a narrow set of predictors and fail to accurately assess lifetime risk.

c. Problem Elaboration

Identifying people at risk of CVD is a cornerstone of preventive cardiology. Risk prediction models currently recommended by clinical guidelines are typically based on a limited number of predictors with sub-optimal performance across all patient groups. In recent decades, clinical and public health efforts to reduce the burden of cardiovascular disease have emphasized the importance of calculating global, short-term (generally

10-year) risk estimates. However, the majority of adults in the United States and even worldwide who are considered to be at low risk for cardiovascular disease in the short term are actually at high risk across their remaining lifespan.

Data-driven techniques based on ML would improve the performance of risk predictions by agnostically discovering novel risk predictors and learning the complex interactions between them. However, only a few studies have investigated the potential advantages of using ML approaches for CVD risk prediction, focusing only on a limited number of ML methods or a limited number of risk predictors. Here, we aim to assess the potential value of using ML approaches to derive risk prediction models for CVD.

d. Motivation

The motivation of the project would be to enhance precision in risk assessment by identifying CVD risk crucial for timely prevention and better outcomes along with addressing limitations of current approaches that sometimes overlook factors, needing more sophisticated methodologies.

e. Project Scope

Through this project, we aim to develop a classification Machine Learning (ML) model for CVD risk prediction, aiming to improve accuracy by identifying novel risk factors and understanding their complex relationships.

II. Literature Review

a. Relevant Research

1. Study of cardiovascular disease prediction model based on random forest in eastern China

Study Objective: The objective was to construct a CVD prediction model tailored to the population in eastern China, aiming to assess the 3-year risk of CVD in a large cohort.

Key Findings: The study identified almost 30 indicators related to CVD risk, incorporating demographics, lifestyle habits, and clinical biomarkers, with the Random Forest algorithm demonstrating superior performance with an Area Under the Curve (AUC) of 0.787 compared to benchmark models.

Implications: The findings underscore the potential of machine learning, particularly Random Forest, in improving predictive accuracy for personalized risk assessment, which holds implications for targeted interventions and more effective management strategies in CVD prevention within the Chinese population.

2. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants

Study Objective: Evaluate the effectiveness of machine learning (ML) techniques, particularly AutoPrognosis, in improving CVD risk prediction compared to traditional methods.

Key Findings: The ML-based model using AutoPrognosis significantly outperformed traditional risk prediction algorithms, achieving an AUC-ROC of 0.774. It incorporated non-traditional variables like walking pace and overall health rating, improving risk prediction especially in underserved patient subgroups like those with a history of diabetes.

Implications: ML techniques, especially when integrated with automated frameworks like AutoPrognosis, offer a promising avenue for enhancing CVD risk prediction by incorporating novel predictors and capturing complex interactions among variables,

benefiting traditionally underserved patient populations and individuals with specific health conditions like diabetes.

3. Lifetime Risks of Cardiovascular Disease

Study Objective: Investigating lifetime CVD risks across age groups in black and white adults based on their health factors like blood pressure and cholesterol, using data from many studies.

Key Findings: People with healthier habits have much lower chances of heart disease compared to those with more risk factors, like high blood pressure or smoking. These findings are the same across different ages and races.

Implications: It's crucial to focus on preventing risk factors like high blood pressure and smoking early on to reduce the chances of heart disease later in life. This research can help doctors and policymakers make better decisions to improve heart health for everyone.

III. Methodology

a. Dataset Description

These datasets include data for the estimation of obesity levels in individuals from the countries of Mexico, Peru and Colombia, based on their eating habits and physical condition. The data contains 17 attributes which are as follows:

Table 1

Description of Data variables

Data Variable	Description
Age	Age of the individual
Height	Height (in meters)
Weight	Weight (in kilograms)
family_history_with_overweight	If family member suffered or suffers from overweight
FAVC	Frequent consumption of high caloric food
FCVC	Frequency of consumption of vegetables
NCP	Number of main meals
CAEC	Consumption of food between meals
SMOKE	Whether individual is a smoker
CH2O	Daily consumption of water
SCC	Calories consumption monitoring
FAF	Physical activity frequency
TUE	Time using technology devices
CALC	Consumption of alcohol
MTRANS	Transportation used
NObeyesdad	Obesity level deducted (Target variable)

b. Data Collection

The data consist of the estimation of obesity levels in people from the countries of Mexico, Peru and Colombia, with ages between 14 and 61 and diverse eating habits and physical condition. Data was collected using a web platform with a survey where anonymous users answered each question, then the information was processed obtaining 17 attributes.This data was then further passed through a deep learning model which generated additional records of the 17 collected variables.

c. Data Preprocessing and/or Feature Engineering

The data provided was clean, accurate and error free and did not require any preprocessing techniques.

d. Data Modeling

In this phase of the project, we focused on building machine learning models to classify obesity risk, a key factor associated with cardiovascular disease. Our approach consisted several key steps:

i. Model Selection and Training

We carefully selected appropriate machine learning models for classification tasks. Considering factors such as accuracy and performance. Our model selection included Random Forest, XGboost, Multilayer Perceptron (MLP), Logistic Regression, K-Nearest Neighbor, Gradient Boosting and Naive Bayes. These models were trained on the cleaned and preprocessed datasets (original dataset and deep learning model dataset) to learn patterns and relationships between features and target variable, which is the level of obesity.

ii. Model Evaluation

After model training, we rigorously evaluated the performance of each model using accuracy as the primary metric. However, we did not solely rely on accuracy; we also analyzed the confusion matrix to get deeper into prediction results to ensure our chosen model not only achieved high accuracy but also demonstrated high performance.

iii. Model Comparison

We compared the performance of different models to identify the top performers. This involved analyzing metrics such as accuracy, interpretability and robustness to determine which models were most suitable for deployment in real-life scenarios.

iv. Final Model Selection

Based on our evaluation results, we selected the best performing model(s) for our project. This decision was made considering the trade-offs between accuracy, interpretability and other factors relevant to our specific objectives.

IV. Results & Analysis

Original Dataset Analysis

In our analysis of the original dataset, the Multilayer Perceptron (MLP) model emerged as the top performer, achieving an impressive accuracy of 90%. We selected MLP for its exceptional ability to understand complex patterns in data, particularly when relationships between different variables are complex. Its multi-layer architecture enables it to grasp these relationships effectively. Additionally, MLP demonstrates versatility in handling various types of data and can be trained efficiently using complex techniques. Although models like Gradient Boosting and XGBoost performed well, MLP outshines them in accuracy.

Table 2

Accuracies for the Original dataset

Model	Accuracy
Random Forest	86.5%
XGBoost	89%
Multilayer Perceptron	90%
Logistic Regression	76.2%
K-Nearest Neighbor	89.1%
Gradient Boosting	89.5%
Naive Bayes	61.3%

Deep Learning Dataset Analysis

In the assessment of the deep learning model dataset, XGBoost showed superior performance with an accuracy of 88%, making it our top choice. We opted for XGBoost due to its efficiency, robustness in handling missing data, and automatic handling of feature interactions. Its ability to learn from past mistakes and enhance predictions over time further solidified its position. Despite decent performances from models like MLP and Gradient Boosting, none could surpass XGBoost's accuracy.

Table 3

Accuracies for the Deep Learning model dataset

Model	Accuracy
Random Forest	83%
XGBoost	88%
Multilayer Perceptron	86%
Logistic Regression	76.2%
K-Nearest Neighbor	77.5%
Gradient Boosting	84%
Naive Bayes	69.9%

Comparison:

Comparing both datasets in terms of model performance leads to the conclusion that, for capturing intricate patterns, the original dataset is preferred, as indicated by the superior accuracy of the MLP model. Conversely, for tasks involving structured data analysis and precise predictions, the deep learning dataset paired with the XGBoost model emerges as the more suitable choice. This highlights the importance of aligning the dataset and model selection with the specific objectives of the analysis, whether it be pattern recognition or accurate prediction in structured data scenarios.

V. Conclusion

a. Conclusion

In conclusion, our project aimed to develop a classification Machine Learning (ML) model for Cardiovascular Disease (CVD) risk prediction. Through rigorous analysis, we explored various machine learning algorithms and datasets to achieve this goal. The findings from our study explains the potential of ML techniques in enhancing predictive accuracy for personalized risk assessment, thereby aiding in targeted interventions and more effective management strategies in CVD prevention.

Our analysis revealed that for the original dataset, the Multilayer Perceptron (MLP) model emerged as the top performer, while the deep learning dataset, XGBoost, showed superior performance. These results highlight the importance of aligning dataset characteristics with appropriate machine learning models to achieve optimal performance.

b. Project Limitations

Despite the promising results, our project had several limitations. Firstly, the datasets used were limited to certain regions and may not be fully representative of diverse populations. Additionally, the datasets focused on obesity levels as a proxy for CVD risk, which may not capture the full spectrum of risk factors associated with cardiovascular disease. Moreover, while we explored various machine learning algorithms, our study did not delve into advanced techniques such as ensemble learning or neural architecture search, which could potentially further improve model performance.

c. Future Research

In the future, researchers could improve this study by using more varied and detailed information from different groups of people. They could also try more advanced machine learning algorithms to analyze the data, which might make predictions even better. Also, they could look into combining different methods to make the predictions more accurate. Furthermore, integrating real-time data streams and wearable sensor

data could enable continuous monitoring and personalized risk assessment, paving the way for proactive interventions and personalized healthcare approaches in cardiovascular disease prevention.

VI. References

[1] World Heart Federation. (2023, August 10). Cardiovascular Disease (CVD) | World Heart Federation. https://world-heart-federation.org/what-is-cvd/

[2] Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X., Yu, W., & Yan, J. (2020). Study of cardiovascular disease prediction model based on random forest in eastern China. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-62133-5

[3] 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. https://doi.org/10.1371/journal.pone.0213653

[4] Jarett D. Berry, M.D., Alan Dyer, Ph.D., Xuan Cai, M.S., Daniel B. Garside, B.S., Hongyan Ning, M.D., Avis Thomas, M.S., Philip Greenland, M.D., Linda Van Horn, R.D., Ph.D., Russell P. Tracy, Ph.D., and Donald M. Lloyd-Jones, M.D. Lifetime Risks of Cardiovascular Disease. https://www.nejm.org/doi/full/10.1056/NEJMoa1012848

[5] EliteDataScience. (2022, July 6). Overfitting in Machine learning: What it is and how to prevent it. EliteDataScience. https://elitedatascience.com/overfitting-in-machine-learning

[6] What is Overfitting? - Overfitting in Machine Learning Explained - AWS. (n.d.). Amazon Web Services, Inc. https://aws.amazon.com/what-is/overfitting/

[7] *adinayak99/GW-Capstone-Project: Files related to the capstone*. (n.d.). GitHub. https://github.com/adinayak99/GW-Capstone-Project