CARDIOVASCULAR RISK PREDICTION

Noel Benny 20MIA1020
Department of Computer Science andEngineering
Vellore Institute of Technology Chennai, Tamil Nadu, India
noel.benny2020@vitstudent.ac.in

Diya Al Din 20MIA1048
Department of Computer Science andEngineering
Vellore Institute of Technology Chennai, Tamil Nadu, India
diyaal.din2020@vitstudent.ac.in

Antony George Mathew 20MIA1022
Department of Computer Science andEngineering
Vellore Institute of Technology Chennai, Tamil Nadu, India
antonygeorge.mathewk2020@vitstudent.ac.in

Hidesh Mathew Sebi 20MIA1064
Department of Computer Science andEngineering
Vellore Institute of Technology Chennai, Tamil Nadu, India
hideshmathew.sebi2020@vitstudent.ac.in

***Abstract*- This paper discusses the features of Cardiovascular Risk Prediction which gathers data related to the topic from multiple sources, analyzes the information, and compiles the result to produce an article with proper citations and references.**

I. INTRODUCTION

Cardiovascular disease (CVD) is a group of conditions that affect the heart and blood vessels, with coronary heart disease (CHD) being one of the main types. CHD occurs when the flow of oxygen-rich blood to the heart muscle is blocked or reduced, often due to a build-up of fatty deposits in the arteries. Understanding the risk factors associated with CHD and accurately predicting the likelihood of developing the disease is crucial for effective prevention and management. This report is based on an ongoing cardiovascular study conducted in Framingham, Massachusetts, which aims to predict the 10-year risk of future CHD in patients. The data has been cleaned and analyzed to identify patterns and insights, and machine learning algorithms such as logistic regression, Random Forest, XGBoost, and naive Bayes classifier have been applied for prediction. The results show that XGBoost has the highest accuracy in predicting the risk of CHD. The findings from this study can provide valuable information for healthcare professionals and patients to understand the factors that influence CHD risk and take proactive steps towards prevention and management of cardiovascular disease.

II. LITERATURE REVIEW

Cardiovascular disease (CVD) is a leading cause of morbidity and mortality worldwide, with coronary heart disease (CHD) being a major contributor to this burden. Numerous studies have been conducted to investigate the risk factors associated with CHD and develop predictive models to identify individuals at higher risk. The Framingham Heart Study, initiated in 1948, is a seminal and ongoing cardiovascular study that has made significant contributions to our understanding of CVD risk factors and their impact on population health.

The Framingham dataset used in this project includes over 4,000 records and 15 attributes, representing a diverse range of potential risk factors, including demographic, behavioral, and medical factors. The data cleaning process is a critical step in ensuring the accuracy and reliability of the findings. The identification of meaningful research questions and the use of appropriate visualization techniques, such as graphs and other visual entities, have been employed to gain insights from the data.

The findings of this project highlight the importance of several key risk factors in predicting the 10-year risk of future CHD. Gender is a significant factor, with males being more susceptible to CHD compared to females. Age is also a crucial determinant, with patients in the 45-55 age range showing a higher likelihood of developing CHD. Other risk factors such as smoking, diabetes, and prevalent hypertension also emerged as important predictors of CHD risk, consistent with existing literature. Machine learning algorithms, including logistic regression, Random Forest, XGBoost, and naive Bayes classifier, have been employed to develop predictive models. Hyperparameter tuning has been performed to optimize the model performance.

The results indicate that XGBoost achieved the highest accuracy in predicting the risk of CHD, with a training accuracy of 96% and a testing accuracy of 81%. No overfitting was observed, suggesting that XGBoost is a robust model for predicting CHD risk in this dataset. The findings of this project align with previous research on the risk factors associated with CHD and their predictive capabilities. The identification of high-risk individuals using accurate predictive models can inform targeted interventions and preventive measures to reduce the burden of CHD. The results can also serve as a valuable resource for healthcare professionals, policymakers, and patients to make informed decisions about cardiovascular health.

In conclusion, this project based on the Framingham dataset provides important insights into the risk factors associated with CHD and their predictive capabilities. The findings highlight the significance of factors such as gender, age, smoking, diabetes, and prevalent hypertension in determining CHD risk. The application of machine learning algorithms, particularly XGBoost, has demonstrated high accuracy in predicting CHD risk in this dataset. The findings of this project contribute to the existing literature on CVD risk prediction and can have implications for clinical practice, public health policy, and patient education. Further research and validation in diverse populations are warranted to strengthen the evidence base and improve the accuracy of CHD risk prediction models.

III. METHODOLOGY

A. PROBLEM STATEMENT

Cardiovascular disease (CVD) is a leading cause of mortality worldwide. Early prediction and identification of individuals at risk of developing CVD is crucial for effective prevention and intervention strategies. The goal of the ongoing cardiovascular study on residents of Framingham, Massachusetts, as described in the given dataset, is to predict whether a patient has a 10-year risk of future coronary heart disease (CHD). This project aims to develop a machine learning model that can accurately predict the 10-year risk of CHD based on demographic, behavioral, and medical risk factors.

B. DATA SUMMARY

The dataset used in this study contains over 3390 rows and 16 columns present in the dataset. Each attribute represents a potential risk factor for CHD, including demographic information such as age, sex, education, and marital status, behavioral factors such as smoking and alcohol consumption, and medical factors such as blood pressure, cholesterol levels, and history of diabetes and previous stroke. The dataset is cleaned and prepared for analysis.

C. DATA PREPARATION

Data cleaning and preprocessing techniques were applied to the raw dataset to handle missing values, outliers, and inconsistencies. Data normalization and encoding were performed to ensure that all features are in a consistent format for machine learning algorithms. The dataset was then divided into training and testing sets for model development and evaluation.

D. EXPLORATORY DATA ANALYSIS

EDA was performed to gain insights from the dataset and identify patterns or trends. Graphs, visualizations, and statistical analysis were used to explore the relationships between different risk factors and the target variable (10-year risk of CHD). EDA revealed that males had a higher risk of CHD compared to females, and age, smoking, diabetes, and hypertension were identified as significant risk factors for CHD. Some key findings from the EDA include:

- Over 85% of patients in the dataset do not have a 10-year risk of future CHD.
- Males have a higher percentage (18%) of 10-year risk of CHD compared to females (12%).
- Patients between the ages of 45-55 have a higher chance of having a 10-year risk of CHD.
- Smoking, diabetes, and history of stroke were identified as significant risk factors for CHD.
- Patients with prevalent hypertension (high blood pressure) were found to be at higher risk of CHD.

E. FEATURE ENGINEERING

Feature engineering techniques were applied to create new features or transform existing features to improve the predictive performance of the model. For example, the age variable was categorized into age groups to capture the non-linear relationship between age and CHD risk. Also, for heart rate and BP meds the number of null values is less. So, we can fill those values with zero. For education and cigs per day we got many numbers of null values so we can replace with the mean and total cholesterol, BMI, glucose having some outliers so mean cannot be the perfect choice, instead we can fill by median. Feature selection techniques were also applied to identify the most important features for model prediction.

VIF technique was used to remove highly correlated features. VIF determines the strength of the correlation between the independent variables. It is predicted by taking a variable and regressing it against every other variable. VIF score of an independent variable represents how well the variable is explained by other independent variables. Feature, dropping columns having multicollinearity and validate through VIF. Highly correlated features were treated by excluding them from dataset and checking the variance inflation factors. SMOTE technique was used for balancing the dataset.

F. MODEL TRAINING

Four machine learning algorithms, namely logistic regression, random forest, XGBoost, and naive Bayes classifier, were applied to train the model using the training dataset. The models were evaluated based on performance metrics such as accuracy, precision, recall, F1-score, and area under the curve (AUC) using cross-validation techniques to ensure robustness and generalizability of the model.

G. MODEL BUILDING

Based on the evaluation results Random Forest and XG boost models gave precise values compare to other models. XGBoost was selected as the final prediction model due to its high accuracy (96%) on the training set and reasonable accuracy (81%) on the testing set without overfitting. XGBoost is an ensemble-based gradient boosting algorithm that can handle imbalanced datasets, handle non-linear relationships, and provide high predictive accuracy.

H. HYPERPARAMETER TUNING

Hyperparameter tuning was performed to optimize the performance of the XGBoost model. Grid search and random search techniques were used to search for the optimal combination of hyperparameter values such as learning rate, maximum depth, number of estimators, and

subsample ratio. The best hyperparameter values were selected based on the performance metrics, and the model was retrained with the optimized hyperparameters.

IV. FUTURE SCOPE

The project on predicting the 10-year risk of coronary heart disease (CHD) using machine learning techniques has potential for further development and future research. Some of the possible future scope for this project includes:

- **Incorporating More Data:** The current project may have utilized a limited set of data features, but there may be additional relevant data that could be collected and incorporated into the model. For example, genetic data, additional medical history, lifestyle factors, or socioeconomic factors could provide valuable insights and improve the predictive accuracy of the model.
- **Exploring Advanced Machine Learning Techniques:** While the current project utilized popular machine learning algorithms such as logistic regression, random forest, XGBoost, and naive bayes classifier, there are many other advanced machine learning techniques that could be explored. For instance, deep learning algorithms like neural networks or ensemble methods like stacking or boosting could potentially yield better predictive performance.
- **Model Interpretability:** Interpretable machine learning models could provide insights into the factors driving the CHD risk predictions, which could aid in better understanding the underlying patterns and mechanisms. Techniques such as explainable AI or model-agnostic interpretability methods like LIME or SHAP could be applied to make the model more interpretable and understandable to clinicians and stakeholders.
- **Real-Time Risk Prediction:** Currently, the model predicts the 10-year risk of CHD, but there may be a need for real-time or short-term risk prediction for timely intervention. Developing models that can provide risk predictions for shorter time horizons, such as 1-year or 5-year risk, could be useful in clinical practice.
- **Model Validation and Deployment:** The current project may have utilized a specific dataset for model development, but further validation of the model using diverse datasets from different populations and settings would be important to assess its generalizability and real-world performance. Additionally, deploying the model in a real clinical setting would require considerations such as model integration with electronic health records (EHRs), validation in a clinical trial, or compliance with regulatory requirements.
- **Decision Support System:** The developed model could be integrated into a decision support system to assist clinicians in making informed decisions about CHD risk assessment and management. This could include providing personalized risk predictions, generating risk reports, and suggesting appropriate interventions based on individual patient profiles.
- **Risk Stratification and Intervention Strategies:** Further research can be conducted to explore risk stratification strategies using the developed model, such as identifying high-

risk groups or subpopulations that may benefit from targeted interventions. Additionally, evaluating the effectiveness of different intervention strategies, such as lifestyle modifications, medication, or other preventive measures, in reducing CHD risk in identified high-risk groups could be an important area of future research.

The project has significant potential for future research and development, including incorporating more data, exploring advanced machine learning techniques, improving model interpretability, real-time risk prediction, model validation and deployment, developing decision support systems, and evaluating risk stratification and intervention strategies. These areas of future scope could contribute to further enhancing the accuracy and usability of the CHD risk prediction model and its potential impact on clinical practice and public health.

V. CONCLUSION

In this project, we utilized a dataset from an ongoing cardiovascular study to predict the 10-year risk of future coronary heart disease (CHD) in patients. Through data preparation, exploratory data analysis (EDA), feature engineering, and model training, we developed a machine learning model to predict CHD risk using various demographic, behavioral, and medical risk factors. The EDA revealed important insights from the data, such as the higher risk of CHD in males, the influence of age on CHD risk, and the significance of factors such as smoking, diabetes, and history of stroke. Feature engineering techniques were applied to create new features or transform existing features to capture relevant information related to CHD risk.

Four machine learning algorithms - logistic regression, random forest, XGBoost, and naive bayes classifier - were trained and evaluated for their predictive performance. Based on the evaluation results, the XGBoost model was selected as the final model due to its high accuracy, efficiency, and effectiveness in handling complex datasets. Hyperparameter tuning was performed on the XGBoost model to optimize its hyperparameters and further improve its performance. The best combination of hyperparameter values was identified to enhance the accuracy of the model.

The developed model can be utilized to predict the 10-year risk of CHD in patients, which can help in identifying individuals who may be at higher risk and take preventive measures such as lifestyle modifications, medication, or other interventions to reduce the risk of CHD. This can potentially contribute to better patient outcomes and reduce the burden of CHD in the population. Overall, this project provides valuable insights into the prediction of CHD risk using machine learning techniques and highlights the importance of data preparation, EDA, feature engineering, and model training in building accurate and effective predictive models. Future research can be conducted to further enhance the model's performance and explore additional risk factors for CHD prediction.

VI. ACKNOWLEDGMENT

We express our earnest gratitude and utmost appreciation to Dr S. A. SAJIDHA, Associate Professor, School of Computer Science and Engineering, VIT Chennai, who has been a constant source of encouragement and whose guidance has been invaluable for the duration of our project.

We would also like to extend our thanks to the entire faculty of the school for their unwavering support and for imparting their extensive knowledge and expertise to us during the course of our project.

We are also deeply grateful to our parents, family members, and friends who have patiently supported us throughout our project journey and provided us with the opportunity to pursue our academic aspirations in such a renowned institution. Their unwavering support and encouragement have been instrumental in our success.

VII. REFERENCES

https://www.sciencedirect.com/science/article/abs/pii/S0002914904004370

https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-9215.2007.06350.x

https://olympias.lib.uoi.gr/jspui/bitstream/123456789/23693/1/Tzoulaki-2009-Assessment%20of%20claims.pdf

https://www.bmj.com/content/344/bmj.e3318.full

https://arxiv.org/pdf/2210.03154.pdf

http://www.medicine.mcgill.ca/epidemiology/hanley/bios601/CandH-ch0102/TruettCornfiledKannell1967.pdf