Lincoln University, Malaysia

NATIONAL COLLEGE OF MANAGEMENT AND TECHNICAL SCIENCE

Samakhusi, Kathmandu, Nepal

Documentation

On

CVD Prediction Model

Submitted By:

Alex Yonjan Lama (LC00021001001)

Submitted To:

Department of Information and Technology

National College of Management and Technical Science

Aug 28, 2025

ACKNOWLEDGEMENT

We would like to extend our sincere gratitude to all those who have directly or indirectly supported us in the completion of this project. We are especially grateful to our course instructor, Bal krishna Shrestha, for his valuable time, guidance, and support throughout the project. His mentorship and encouragement have been instrumental in shaping the direction of our work.

We also appreciate the feedback and encouragement from our classmates and peers, which have helped us refine and improve our project. Finally, we acknowledge the reference materials and resources that contributed to the successful completion of this work.

Alex Yonjan Lama (LC00021001001)

Program Coordinator

Project Coordinator

ABSTRACT

The prevalence of cardiovascular disease (CVD) is a major global health concern, and early detection can significantly improve patient outcomes. This project focuses on developing a predictive model for cardiovascular disease using machine learning techniques. The dataset contains patient information such as age, height, weight, blood pressure, cholesterol, glucose levels, lifestyle factors, and more.

We performed data preprocessing, feature engineering, and exploratory data analysis to clean and understand the data. Several machine learning models, including Logistic Regression, Decision Tree, and Random Forest, were trained and evaluated. Among these, Random Forest was selected as the best-performing model due to its higher accuracy and ability to handle complex relationships between features.

The developed system can predict the likelihood of a patient having cardiovascular disease based on input features. This project demonstrates the application of machine learning in healthcare and highlights the most influential factors affecting cardiovascular health, such as BMI, blood pressure, age, cholesterol, and glucose levels.

Keywords: *Cardiovascular Disease, Machine Learning, Prediction Model, Random Forest, Feature Engineering, Data Analysis*

.

Table of Contents

1. INTRODUCTION

Cardiovascular disease (CVD) is one of the leading causes of death globally. Predicting the risk of CVD at an early stage can help individuals take preventive measures and improve overall health outcomes. With the growth of data availability and machine learning techniques, it is possible to analyze health-related data and identify high-risk individuals.

This project focuses on developing a machine learning model to predict the likelihood of cardiovascular disease based on patient health data. By using algorithms such as Logistic Regression, Decision Tree, and Random Forest, the model aims to provide accurate predictions that can assist healthcare professionals in early intervention.

1.1 Background

CVD includes conditions such as coronary artery disease, heart attacks, and strokes. Key risk factors include high blood pressure, cholesterol, obesity, diabetes, smoking, and physical inactivity. Traditional risk assessment relies on medical check-ups and manual analysis, which can be time-consuming and less efficient for large populations.

Machine learning offers a data-driven approach to analyze patient records and predict the risk of developing CVD. Early prediction can lead to timely medical advice, lifestyle changes, and targeted healthcare interventions, reducing the burden of cardiovascular diseases on individuals and healthcare systems.

1.2 Objective of the Project

The primary objective of this project is to develop a machine learning model that can accurately predict the likelihood of cardiovascular disease (CVD) in individuals based on their health data. The specific objectives include:

- **Data Analysis:** Understand the dataset, explore features, and identify key factors affecting cardiovascular disease.
- **Feature Engineering:** Create new meaningful features, encode categorical variables, and prepare the data for modeling.
- **Model Development:** Build and compare multiple machine learning models, including Logistic Regression, Decision Tree, and Random Forest.
- **Model Evaluation:** Assess the performance of the models using metrics such as accuracy, precision, recall, F1-score, and confusion matrix.
- **Feature Importance Analysis:** Identify the most influential health indicators contributing to cardiovascular disease prediction.

The ultimate goal is to provide a reliable tool that can support early detection and preventive healthcare for cardiovascular disease.

2. LITERATURE REVIEW

Cardiovascular disease (CVD) is one of the leading causes of death worldwide, making early detection and prevention extremely important. Over the years, researchers have explored various machine learning techniques to predict the likelihood of CVD in individuals based on their health data and lifestyle factors.

Logistic Regression is commonly used for predicting CVD because it is simple, interpretable, and effective for binary classification problems. Decision Trees are also popular because they allow easy visualization of decision-making rules and can capture complex, non-linear relationships between different health features. Random Forest, which is an ensemble of multiple decision trees, has been shown to improve prediction accuracy and reduce overfitting, making it highly effective for datasets with many features.

Previous studies emphasize the importance of key features such as age, blood pressure, cholesterol levels, glucose levels, and body mass index (BMI) in predicting cardiovascular risk. Proper data preprocessing, such as handling missing values and encoding categorical variables, as well as feature engineering, like creating new features from existing data, can significantly improve model performance.

Overall, the literature shows that combining good data preparation with strong machine learning models can lead to reliable prediction of cardiovascular disease. This project follows these insights by applying Logistic Regression, Decision Tree, and Random Forest models to a real-world dataset and identifying the most important features influencing the risk of CVD.

3. DATASET DESCRIPTION

The dataset used in this project contains health-related information of individuals, which is aimed at predicting the likelihood of cardiovascular disease (CVD). It consists of several attributes that capture personal, medical, and lifestyle factors. The dataset is structured in a tabular format where each row represents one individual and each column corresponds to a specific feature.

3.1 Features and Attributes

The main features in the dataset include:

- **Age**: Age of the individual in years (converted from days).
- **Gender**: Male or Female.
- **Height**: Height of the individual in centimeters.
- **Weight**: Weight of the individual in kilograms.
- **Systolic Blood Pressure (ap_hi)**: The higher number in a blood pressure reading.
- **Diastolic Blood Pressure (ap_lo)**: The lower number in a blood pressure reading.
- **Cholesterol**: Cholesterol level categorized as Normal, High, or Very High.
- **Glucose (Gluc)**: Glucose level categorized as Normal, High, or Very High.
- **Smoking Status (smoke)**: Indicates whether the person smokes (Yes/No).
- **Alcohol Intake (alco)**: Indicates whether the person consumes alcohol (Yes/No).
- **Physical Activity (active)**: Indicates whether the person is physically active (Yes/No).
- **BMI (Body Mass Index)**: Derived from weight and height to assess body fat.
- **Overweight**: Derived binary feature; 1 if BMI > 25, else 0.

.

3.2 Data Source

The dataset used for this project was obtained from **Kaggle** and is titled *Cardiovascular Disease Dataset* (https://www.kaggle.com/datasets/mahmudulhaqueshawon/cardiovascular-disease).

It contains a total of 69,997 rows with key columns including age, gender, height, weight, systolic blood pressure (ap_hi), diastolic blood pressure (ap_lo), cholesterol, glucose, smoking, alcohol intake, and physical activity.

For this project, additional features such as BMI and Overweight were derived from the original data.

The target variable is cardio, where a value of 1 indicates the presence of cardiovascular disease and 0 indicates no disease.

4. METHODOLOGY

This section explains the approach taken to predict cardiovascular disease using the dataset. The methodology consists of three main steps: data preprocessing, feature engineering, and exploratory data analysis.

4.1 Data Preprocessing

Data preprocessing involved cleaning the raw dataset to remove invalid or inconsistent values. For example, unrealistic values in height, weight, and blood pressure were filtered out. Age was converted from days to years, and missing or duplicate records were handled. This ensures that the data fed into the models is accurate and reliable.

4.2 Feature Engineering

Feature engineering was performed to create new meaningful features that can improve model performance. The BMI (Body Mass Index) was calculated using height and weight, and a binary Overweight feature was derived from BMI. Categorical variables like cholesterol and glucose levels were encoded into numeric form. Additional columns like age groups and blood pressure categories were also created for better analysis.

4.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to understand patterns and relationships in the dataset. Visualizations were created to examine the distribution of cardiovascular disease across age groups, gender, smoking habits, cholesterol levels, glucose levels, and blood pressure categories. This analysis helped identify important trends and guided feature selection for model building.

5. MODEL BUILDING

This section explains the machine learning models used to predict cardiovascular disease and the process of selecting the best-performing model.

5.1 Logistic Regression

Logistic Regression is a linear model used for binary classification. It estimates the probability of a person having cardiovascular disease based on input features. The model is simple, interpretable, and provides insight into how each feature affects the outcome.

5.2 Decision Tree

Decision Tree is a non-linear model that splits the dataset into subsets based on feature values. It creates a tree-like structure of decision rules that can classify patients with or without cardiovascular disease. It is easy to visualize and understand but can overfit on the training data.

5.3 Random Forest

Random Forest is an ensemble model that combines multiple decision trees to improve accuracy and reduce overfitting. It is robust, handles both numerical and categorical features well, and was found to be the best-performing model in this project. Feature importance can also be extracted from the model to understand which factors contribute most to cardiovascular risk.

5.4 Model Selection

After training and evaluating all three models, Random Forest was selected as the final model because it achieved higher overall accuracy and provided stable predictions. Logistic Regression was useful for linear insights, and Decision Tree helped understand feature-based decision rules, but Random Forest offered the best performance for this dataset.

6. Model Evaluation

This section describes how the trained models were evaluated using various performance metrics and how the important features were identified.

6.1 Accuracy, Precision, Recall, F1-Score

The models were assessed using accuracy, precision, recall, and F1-score. Accuracy shows the overall percentage of correct predictions. Precision measures how many of the predicted positive cases were actually positive. Recall indicates how many actual positive cases were correctly identified. F1-score is the harmonic mean of precision and recall, providing a balanced measure of model performance.

6.2 Confusion Matrix

A confusion matrix was generated for each model to visualize the number of true positives, true negatives, false positives, and false negatives. This helps to understand where the model makes errors and how it differentiates between patients with and without cardiovascular disease.

6.3 Feature Importance

Random Forest provides a ranking of feature importance, showing which factors most influence the prediction of cardiovascular disease. In this project, features like BMI, Overweight, systolic blood pressure (ap_hi), age groups, cholesterol, and glucose were the most influential in determining cardiovascular risk. This insight can guide preventive measures and further research.

7. RESULT AND DISCUSSION

After training and evaluating the models, the following observations were made:

- **Logistic Regression** achieved the highest accuracy of 73%, providing good linear insights into the data. It showed a balanced performance in predicting both patients with and without cardiovascular disease.
- **Decision Tree** had lower accuracy at around 64% and tended to overfit on certain patterns, making it less reliable for unseen data.
- **Random Forest** performed well with an accuracy of 70% and offered robust predictions by combining multiple decision trees, reducing overfitting compared to a single tree.

The most important features identified by Random Forest included BMI, Overweight status, systolic blood pressure (ap_hi), age groups, cholesterol, and glucose levels. These features had the highest impact on predicting cardiovascular disease and can serve as key indicators for risk assessment.

The results highlight the effectiveness of machine learning models in predicting cardiovascular risk using easily measurable patient attributes. This model can assist healthcare providers in identifying high-risk patients and prioritizing preventive measures.

8. CONCLUSION

In this project, a cardiovascular disease prediction model was developed using machine learning techniques. The dataset was preprocessed, features were engineered including BMI and Overweight status, and exploratory data analysis was performed to understand patterns in the data. Three models, Logistic Regression, Decision Tree, and Random Forest, were trained and evaluated.

Random Forest was identified as the most effective model due to its robustness and ability to handle complex patterns in the data. Key features influencing cardiovascular risk included BMI, Overweight, systolic blood pressure (ap_hi), age groups, cholesterol, and glucose levels.

The project demonstrates that machine learning can be a valuable tool for early detection of cardiovascular disease, potentially aiding healthcare professionals in preventive care and risk management.

References:

Mahmudul Haque Shawon. Cardiovascular Disease Dataset. Kaggle. Retrieved from: https://www.kaggle.com/datasets/mahmudulhaqueshawon/cardiovascular-disease

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 2nd Edition. O'Reilly Media.

Raschka, S., & Mirjalili, V. (2019). Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2. 3rd Edition. Packt Publishing.

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Chollet, F. (2018). Deep Learning with Python. Manning Publications.

Brownlee, J. (2020). Machine Learning Mastery with Python. Machine Learning Mastery.

World Health Organization. Cardiovascular Diseases (CVDs). Retrieved from: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)