PREDICTIVE ANALYSIS OF CARDIOVASCULAR DISEASE USING RANDOM FOREST ALGORITHM

\

*Tanuja K T   
 Department of Biotechnology  
 Kumaraguru College of Technology  
Coimbatore, India* [tanuja.22bt@kct.ac.in](mailto:tanuja.22bt@kct.ac.in)

*Dhivya J  
 Department of Mathematics*  
*Kumaraguru College of Technology  
Coimbatore, India* [dhivya.j.sci@kct.ac.in](mailto:dhivya.j.sci@kct.ac.in)

*Shamyuktha V P  
 Department of Biotechnology*   
*Kumaraguru College of Technology  
Coimbatore, India* [shamyuktha.22bt@kct.ac.in](mailto:shamyuktha.22bt@kct.ac.in)

*Anushree N  
 Department of Biotechnology*   
*Kumaraguru College of Technology  
Coimbatore, India* [anushree.22bt@kct.ac.in](mailto:anushree.22bt@kct.ac.in)

*Abstract*—This research gives a comprehensive analysis of cardiovascular disease data obtained from Kaggle, to predict the progression and [commencement](https://www.bing.com/ck/a?!&&p=d5f42aa65aa595d7e2b4b77404c313d86a988bb983d23476d3e0e13f45e9a5deJmltdHM9MTczNzQxNzYwMA&ptn=3&ver=2&hsh=4&fclid=0631f691-112d-6fe9-1772-e5fd10706e6f&u=a1L3NlYXJjaD9xPWRlZmluZStjb21tZW5jZW1lbnQmRk9STT1EQ1RSUVk&ntb=1) of cardiovascular diseases. CVDs, which are leading concern of the world, covers a wide area of conditions affecting blood vessels including heart, causing significant risks to public health. This study investigates the impact of various factors that influence CVDs such as age, gender, fasting blood sugar, resting blood pressure, exercise-induced angina, maximum heart rate, and the number of major blood vessels affected. Using exploratory data analysis techniques, the research seeks to expose the patterns and relationships among these variables. Using Machine Learning techniques such as Random Forest Algorithm, the risk factors of CVDs have been evaluated. In the end, the study proves to be a promising impact on public health and raise awareness about the heart health.

Keywords— Cardiovascular disease, Machine Learning, Random Forest algorithm, Exploratory data analysis

# Introduction (*Heading 1*)

CVDs signifies a potential concern of world health , which is the major reason of death across the world. In 2021 alone, CVDs were responsible for over 20.5 million deaths, affecting more than half a billion people [7]. Major risk parameters for stroke and heart disease include lack of physical activity, an unhealthy diet, excessive alcohol consumption and tobacco use. These diseases often involve the fatty deposits accumulation in atherosclerosis (arteries) with an increasing risk of blood clots, potentially leading to organs damage such as the heart, brain, eyes and kidneys .

The premature deaths due to CVDs can be prevented by early identification of individuals at high risk and ensuring their appropriate treatment at right time. The World Health Organization emphasizes on the importance of providing access to the essential health technologies and non-communicable disease medications in basic healthcare facilities to ensure that those in need receive proper treatment and counseling [8].

Healthcare systems have amassed vast amounts of data on heart disease, including various medical parameters such as gender, blood pressure , age ,cholesterol levels, and chest pain type. These datasets, containing 14 different medical parameters, are now available for analysis to extract crucial information. Data Analytical techniques, such as naïve Bayes, support vector machine, logistic regression, and k-nearest neighbor (KNN), can be employed to predict heart diseases at its initial stages by classifying whether a person has cardiovascular disease depending on the parameters obtained using the datasets.

In this paper, the primary aim is to enhance the effectiveness of predicting the incidence of heart disease. The chosen machine learning technique for prediction is the Random Forest algorithm. Random Forest is a widely used supervised learning method that excels in both regression tasks and classification. Based on ensemble learning principles, it harnesses the collective strength of multiple classifiers to tackle complex problems and improve overall model performance.

# Dataset Description

Whenever there is a reduction or blockage in the flow of oxygen-rich blood to the heart muscle, which results in increased strain on the heart and potential complications such as chest pain(angina), heart failure and heart attacks, CVDS occurs. Similarly, strokes along with transient ischemic attacks (TIAs) are associated with compromised blood supply to the brain. This dataset on heart disease originates from a Indian multispecialty hospital and it has the information on 1000 subjects along with 12 features. This serves an important data for creating predictive machine learning models and early-stage heart disease detection systems.

Each patient is identified uniquely by a numerical code in the "Patient Identification Number" attribute. Age is recorded in years under the "Age" attribute, while gender is represented as a binary variable (0 for female, 1 for male) in the "Gender" attribute. The "Chest Pain Type" attribute indicates chest pain into four types: non- asymptomatic (3) ,anginal pain (2), atypical angina (1) and typical angina (0). "Resting Blood Pressure" (resting) provides baseline blood pressure measurements in mm Hg, ranging from 94 to 200. Serum cholesterol levels are denoted in mg/dl under the "Serum Cholesterol" attribute, ranging from 126 to 564.

The "Fasting Blood Sugar" attribute presents whether fasting blood sugar exceeds 120 mg/dl (0 for false 1 for true). Results of the resting electrocardiogram are captured by the "Resting Electrocardiogram Results" attribute coded as 2 for probable or definite left ventricular hypertrophy, 1 for the presence of ST-T wave abnormality and 0 for normal, based on Estes' criteria. "Maximum Heart Rate Achieved" (max heart rate) records the highest heart rate attained by the patient, with values ranging from 71 to 202.

"Exercise Induced Angina" (exercise angina) is binary, indicating the presence (1) or absence (0) of angina induced by exercise. The "Old peak = ST" (old peak) attribute measures ST depression induced by exercise relative to rest, ranging from 0 to 6.2. The "Slope of the Peak Exercise ST Segment" (slope) is nominal, representing the slope as 1 (upsloping), 2 (flat), or 3 (down sloping). "Number of Major Vessels" (no of major vessels) denotes the count of major vessels colored by fluoroscopy, with possible values of 0, 1, 2, or 3.

Finally, the "Classification" attribute (target) is binary, indicating the presence (1) or absence (0) of heart disease. Totally, the dataset provides a wide range of features suitable for complete analysis and development of predictive models for heart disease identification.

# Methodology

The dataset was cleaned and preprocessed using MS Excel, applying few techniques to fill up the missing values and other errors. Pictorial and Graphical comparisons of various attributes were conducted with the help of Excel-generated graphs, enabling the identification of correlations along with crucial insights within the dataset. These graphical tools played a major role in illustrating significant relationships and important patterns.

In place of machine learning analysis, the Random Forest algorithm was used. Random Forest (RF), established by Leo Breiman, is a fast and highly accurate classification method well known for its buoyancy against noise. The technique joins bagging and random feature selection to establish an ensemble of trees. Each tree in the forest is created based on randomly sampled vectors, with all trees sharing an identical distribution. This strategy enriches the strength and productivity of the classification process. Random Forest is broadly acknowledged as a top-tier supervised learning algorithm in machine learning, outshinning in both regression and classification tasks. It excellently addresses overfitting challenges and increases prediction accuracy by creating an group of decision trees with the help of bagging and feature randomness.

Random Forest gives a robust and uncorrelated forest by creating large amount of decision trees and combining their predictions. During the node-splitting process, it cleverly considers only random subsets of nodes and features, resulting in major accuracy improvements. Moreover, this algorithm is skilled at handling large and huge datasets and mitigating overfitting concerns.

Construction of decision trees for each sample, and aggregating predictions through a voting mechanism is done by the workflow that entails by randomly selecting samples from the dataset. This collective approach constantly produces more accurate and reliable output compared to individual trees.

In contrast to this, the Naïve Bayes theorem works under the assumption of feature independence within a class, that uses a formula to predict events based on conditional probabilities. The equation, P(S|T) = P(T|S) \* P(S) / P(T), evaluates the possibility of an event fitting to a specific class given the observed features, enabling classification in complex scenarios. Here, P represents the total attribute, T denotes the event to be predicted, and S signifies the class value for the event.

# Discussion

*A. Presence and absence of CVDs:*

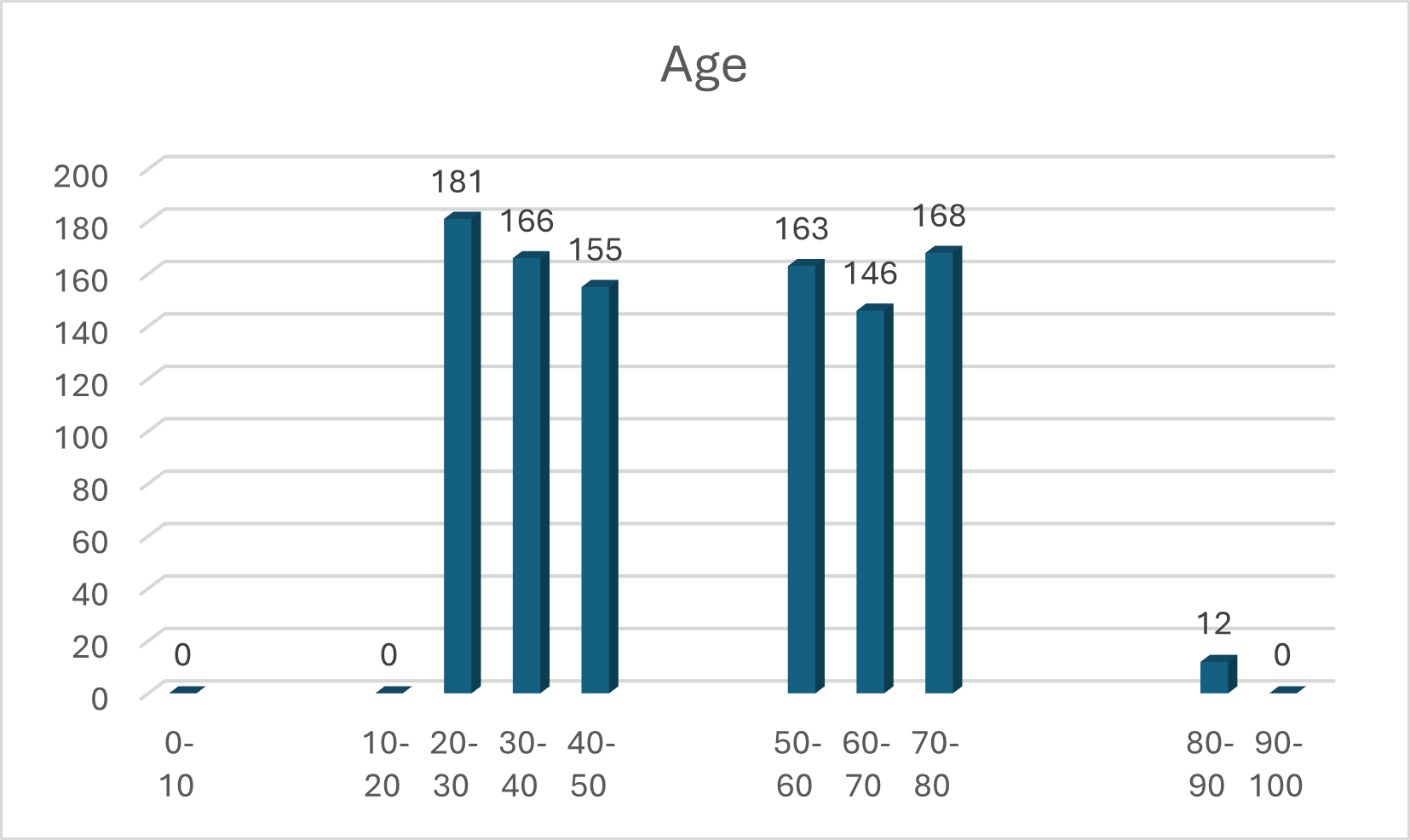
A blue pie chart with a white text

Description automatically generated

**Figure 1: This chart illustrates the absence or presence of cardiovascular disease (CVD) within a dataset comprising 1000 samples.**

Among these samples, 580 patients were identified as having confirmed cases of CVD. Consequently, this dataset proves valuable for predicting the factors contributing to the occurrence of cardiovascular diseases.

*B. Frequency of Target with respect to Age groups:*

****

**Figure 2: The bar graph provided illustrates the distribution of cardiovascular disease (CVD) across different age groups.**

Notably, the age bracket of 20-30 demonstrates the highest incidence of heart diseases, totaling 181 cases. This signifies a departure from previous trends, where CVDs were less prevalent among adults. The increase in coronary heart disease among young adults can be attributed to factors such as smoking, obesity, and insufficient physical activity, which impose strain on the cardiovascular system.

Conversely, the age group of 70-80 exhibits the second-highest rate of CVD, with 168 cases. This heightened risk is linked to age-related changes in older adults in the heart and blood vessels. As individuals age, there are modifications in cardiac function and vascular dynamics, leading to an augmented susceptibility to cardiovascular diseases. Age-related factors, including a diminished ability of the heart to beat rapidly during physical activity and stress, coupled with reduced metabolism, further contribute to the increased risk of coronary heart diseases.

*C. Frequency of Target with respect to Serum Cholesterol Level:*

A graph of a number of blue bars

Description automatically generated with medium confidence

**Figure 3: The histogram depicts the distribution of serum cholesterol levels in the blood.**

It represents the distribution of serum cholesterol level and the corresponding number of affected individuals, with the highest occurrence observed in the serum cholesterol range of 320-380. Conversely, there is minimal representation in the ranges of 64-128 and 576-640. Elevated cholesterol levels can lead to the accumulation of fatty deposits in blood vessels, impeding arterial blood flow and increasing the risk of clot formation, potentially resulting in heart attacks or strokes.

A person's serum cholesterol level encompasses the quantity of triglycerides, high-density lipoprotein (HDL) and low-density lipoprotein (LDL) in the blood. Research indicates that elevated quantity of HDL-C are significantly associated with a reduced risk of CVD, whereas LDL-C is identified as a common risk paramter for CVD. These findings underscore the complex nature of serum cholesterol levels and their relevance to cardiovascular risk.

.

*D. Presence of CVDs with respect to heart rate:*

A graph showing the growth of a number of people

Description automatically generated with medium confidence

**Figure 4: This chart represents the presence of CVDs with heart rate**

This chart implies a strong correlation between This the maximum heart rate and the presence or absence of cardiovascular disease (CVD). Remarkably, the absence of CVDs is observed for maximum heart rates as low as 96, and conversely, the absence of CVD is sometimes associated with maximum heart rates as high as 201.

*E. Chest pain type in relation to the presence of CVDs:*

A blue pie chart with white text

Description automatically generated

**Figure 5: The pie chart illustrates the types of chest pain**

The pie chart offers a comprehensive breakdown of the distribution of chest pain types concerning the presence of cardiovascular diseases (CVDs). The predominant category, non-anginal pain, commands the largest share, constituting 48% of the total. This indicates that a significant proportion of individuals experiencing chest pain during the presence of CVDs describe their pain as non-anginal.

Following closely, atypical angina represents 26% of the chart, signifying a substantial yet slightly smaller portion of cases. This suggests that a notable number of individuals with CVDs encounter chest pain that deviates from the typical angina pattern.

Typical angina, comprising 19% of the distribution, points to a distinct and recognizable pattern of chest pain in a considerable proportion of CVD cases. By comparing the asymptomatic category, which represents the smallest segment at 7%, this indicates that a small group of people with CVDs might not experience chest pain as a symptom.

This graphical representation highlights that non-anginal pain is the most prevalent chest pain type in the context of CVDs.

*F. Presence of CVDs with respect to fasting blood sugar:*

A graph of a graph showing the amount of blood sugar in the blood sugar level

Description automatically generated

**Figure 6: Grouped column chart demonstrating the CVDs with fasting blood sugar**

The given column chart describes the correlation between the occurrence of cardiovascular diseases (CVDs) and fasting blood sugar levels. In this column chart, value 0 indicates negative result for fasting blood sugar test, while value 1 denotes a positive result. This visual representation particularly reveals a pattern where CVDs are mostly absent during the fasting blood sugar representing the value 0. On the other hand, prevalence of CVDs is remarkably higher when the fasting blood sugar gives the value of 1. This statement recommends a potential relationship between elevated fasting blood sugar levels vs an increased likelihood of cardiovascular diseases. It highlights the prevalence of considering fasting blood sugar as a important factor in assessing the risk of CVDs.

*G. Presence of CVDs regarding Resting electrocardiogram:*

A graph showing the number of patients

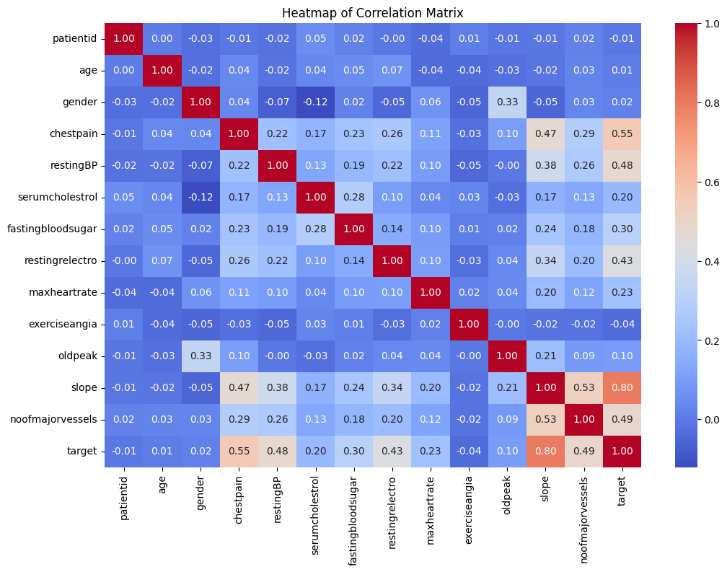
Description automatically generated with medium confidence

**Figure 7: Stacked Column chart representing presence of CVDs with Resting Electrocardiogram**

The above chart denotes the correlation between the presence of cardiovascular disease (CVD) and resting electrocardiogram (ECG) results. In the stacked column chart, values 0, 1, and 2 corresponds to ECG findings: 0 indicates a normal ECG, 1 represents the presence of ST-T wave abnormalities and 2 indicates ECG results indicating definite or probable left ventricular hypertrophy by Estes' criteria.

This chart successfully acknowledge the apparent pattern where the prevalence of CVD increases from 0 to 2. Precisely, most patterns and results exhibit the absence of CVD (value 0), while an increased presence of CVD is evident whenever the ECG result gives the value of 2, representing probable or definite left ventricular hypertrophy. This interpretation highlights the association between specific ECG abnormalities, espicially those that reveal left ventricular hypertrophy, and an elevated likelihood of cardiovascular diseases. The chart serves as an important visual tool for noticing the interplay between resting ECG and the presence of CVD.

*H. Performance Evaluation of the Classification Model*

**

**Figure 8: Heatmap of the Correlation Matrix**

The function of Random Forest classification model was estimated using of metrics derived from confusion matrix. This model depicts a strong predictive efficiency with True Positive count of 136 and True Negative count of 157, revealing its ability to specifically classify positive as well as negative commands. Misclassifications were likely to be low, with False Negative and False Positive counts at 2 and 5, respectively. The Sensitivity of this model is 98.60%, pointing its effectiveness in identifying positive cases,96.90% of specificity, depicting its precision in detecting negative cases. Moreover, the overall Accuracy of the Random Forest model was97.70%,which is quite impressive for its reliability and efficiency for classification tasks by maintaining a balanced approach to sensitivity and specificity.

|  |  |
| --- | --- |
| True Positive | 136 |
| True Negative | 157 |
| False Positive | 5 |
| False Negative | 2 |
| Sensitivity | 98.60% |
| Specificity | 96.90% |
| Accuracy | 97.70% |

**Table 1: Performance Metrics for Random Forest Classification Model**

# Conclusion

In this predictive analysis of the cardiovascular disease dataset that is cleaned and interpreted using both Excel and Random forest algorithm, through which we observed important and useful learnings about the complex interplay of factors impelling cardiovascular health. The addition of serum cholesterol, age, resting electrocardiogram (ECG), fasting blood sugar, gender, heart rate, and chest pain type allowed for a complete examination.  
Excel offers a deep foundation for visualizations and descriptive statistics, by revealing apparent trends such as the impact of gender, age, and cholesterol levels regarding cardiovascular risk. The addition of random forest techniques, increased our analysis point of view by helping machine learning which works by capturing intricate relationships within the data.  
The random forest algorithm has also helped in the development of a predictive model, by facilitating deep understandings with accuracy of 97.70%, like how numerous parameters contribute to cardiovascular outcomes. It addition to this ,it also identifies feature importance and uncovered complex interactions, offering a more precise and exact prediction of cardiovascular risk. This increased understanding can infer more targeted preventive strategies and interventions, as the model's performance metrics offer a quantitative assessment of its predictive capabilities.  
Finally , the aid of Random forest algorithm and Excel in our analysis enhanced and increased the depth of our insights into cardiovascular disease, offering a deep foundation for informed decision-making in healthcare and further research.

References

[1] United States. Public Health Service. Office of the Surgeon General. (2010). How tobacco smoke causes disease: the biology and behavioural basis for smoking-attributable disease: a report of the Surgeon General. US Department of Health and Human Services, Public Health Service, Office of the Surgeon General.

[2]Subasi, A., Alickovic, E., & Kevric, J. (2017). Diagnosis of chronic kidney disease by using random forest. In CMBEBIH 2017: Proceedings of the International Conference on Medical and Biological Engineering 2017 (pp. 589-594). Springer Singapore.

[3] Breiman, L. (2001). Random forests. Machine learning, 45, 5-32.

[4] Dudley, P. (2003). Exercise and physical activity in the prevention and treatment of atherosclerotic cardiovascular disease. Arterioscler Thromb Vasc Biol, 23, 1319-1321.

[5] Winkler, W. E. (2003, August). Data cleaning methods. In Proc. SIGKDD.

[6] de Silva, R. A. (2013). Heart disease. Bloomsbury Publishing USA.

[7] Di Cesare, M., Perel, P., Taylor, S., Kabudula, C., Bixby, H., Gaziano, T. A., McGhie, D. V., Mwangi, J., Pervan, B., Narula, J., Pineiro, D., & Pinto, F. J. (2024). The Heart of the World. *Global heart*, *19*(1), 11. <https://doi.org/10.5334/gh.1288>

[8] World Health Organization: WHO. (2021, June 11). *Cardiovascular diseases (CVDs)*. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).