Cardiovascular Disease Prediction using Machine Learning Models

Arnav Gupta 2021236 | Karan Gupta 2021258 | Shivesh Gulati 2021286 | Vishal Singh 2021575

1. Abstract

Cardiovascular diseases are one of the leading causes of fatalities worldwide. However, timely attention to specific parameters, such as blood pressure and blood sugar level, can help mitigate most of these risks. The primary motivation behind this study is to develop a machine learning model capable of predicting cardiovascular diseases(CVD) in an individual, using easy-to-determine parameters such as age, glucose levels, weight, and blood pressure indices. This can serve as a tool for early detection of the risk of cardiovascular diseases among individuals, allowing them to take preventive measures and seek medical attention at an early stage to reduce further risk.
In this study, classification models Naive Bayes, Logistic Regression, and Support Vector Machines were trained on a cleaned and standardized dataset after removing outliers. Dimensionality reduction techniques, such as PCA, were also used to optimize the model performance further. The optimal number of components to be used in PCA for each model were determined using the K-Fold Cross Validation method. The F1 score was used as an evaluation metric for evaluating the performance of the models, and the values obtained are as follows: Support Vector Machines (0.7305), Logistic Regression (0.7294), and Naive Bayes (0.7123)

2. Introduction

Cardiovascular disease (CVD) is a leading cause of morbidity and mortality worldwide. According to the WHO, CVDs account for approximately 31%(approx 17 million) of all global fatalities due to aging populations and changing lifestyles. It encompasses a range of conditions affecting the heart and blood vessels, including coronary artery disease, heart failure, stroke, and hypertension. Smoking remains a major contributor, with tobacco use responsible for nearly 8 million deaths yearly. Unhealthy diets, characterized by excessive saturated fats, salt, sugar, and inadequate fruits and vegetables, are associated with an increased risk of obesity, diabetes, and high glucose levels. High cholesterol levels affect about 38% of adults aged 25 years or older globally. High cholesterol and glucose levels have also been attributed as significant causes of CVD. [1]
The traditional techniques involve getting CT scans and electrocardiograms, which are often very expensive and time-consuming. It is therefore necessary to develop a cheaper, more accessible, and more efficient method for early diagnosing CVD. Machine learning models provide a more efficient and accurate approach to predicting cardiovascular diseases. This reinforces the diagnoses done by traditional methods. Early detection and precise cardiovascular disease risk prediction are critical for timely treatment. The algorithm proposed uses the parameters age, gender, weight, height, smoking, glucose level, systolic blood pressure, diastolic blood pressure, pulse pressure, and mean arterial pressure to predict if a person has CVD.

In our project, we test the effectiveness and accuracy of various algorithms to predict CVD. All the computations are done on Google Colab with Python and using a publicly available dataset on Kaggle.

3. Literature Survey

[1] Effective Heart disease prediction using machine learning techniques
This research paper aims to develop a machine-learning model to detect Cardiovascular diseases accurately. The study uses Decision trees, XGBoost, Random forest, and multilayer perceptron models. These were optimized using GridSearchCV, a library in scikit-learn for k-fold cross-validation. K-modes were applied to scale and preprocess the dataset. The dataset used for this study was obtained from Kaggle.com and consisted of 70,000 records.

The author used various techniques for outlier detection and removing errors in data input. Since most data is categorical, removing outliers would improve performance and accuracy. After outlier removal, the data had 57,155 instances. Feature engineering was done to improve accuracy further. One of the significant techniques was to convert continuous data into categorical data. The binning technique was applied to age, height, weight, systolic blood pressure, and diastolic blood pressure to convert them to categorical data. Two derived parameters, viz Body Mass index (BMI) and Mean Arterial Pressure (MAP) were derived from age, weight, and Systolic Blood Pressure, Diastolic Blood Pressure, respectively.

Clustering was initially used as a method for grouping similar instances. However, it was found unsuitable for categorical datasets; hence, a new algorithm, "K-Modes," was used. The optimal number of clusters for male and female datasets was determined using the elbow curve method. The dataset was split on a gender basis, making it possible to predict the risk of CVD for each gender. This resulted in better prediction. The dataset was split into 80:20 for training and testing. During various machine learning model testing, MLP was best to perform with an accuracy of 87.28% and other algorithms like XGBoost and random forest also showed comparable performances.

To conclude, this research paper applied models that achieved an accuracy of 85%, which could be further improved in the future to get more accurate predictions of CVD. .[1]

[2] Blood Pressure Variables and Cardiovascular Risk: New findings From ADVANCE

This research paper aims to analyze the effectiveness of various BP indices in predicting the risk of cardiovascular events, especially in people with type-2 diabetes. The dataset used for this study, although not disclosed in its entirety, was ob-

Feature	Description	Type of Feature	Data-Type	Unit of measurement	Value Range
age	Age	Objective Feature	integer	Days	Any Integer Value $>= 0$
height	Height	Objective Feature	integer	Centimeters (cm)	Any Integer Value > 0
weight	Weight	Objective Feature	float	Kilograms (kg)	Any Floating Value > 0
gender	Gender	Objective Feature	Categorical Code	-	1: Female 2: Male
ap_hi	Systolic Blood Pressure	Examination Feature	integer	Millimetre(s) of Mercury (mmHg)	Any Integer Value > 0
ap_lo	Diastolic Blood Pressure	Examination Feature	integer	Millimetre(s) of Mercury(mmHg)	Any Integer Value > 0
cholesterol	Cholesterol	Examination Feature	Categorical Code	-	1:Normal 2:Above Normal 3:Well Above Normal
gluc	Glucose	Examination Feature	Categorical Code	-	1:Normal 2: Above Normal 3:Well Above Normal
smoke	Smoking	Subjective Feature	Categorical Code	-	0: Non-Smoker 1: Smoker
alco	Alcohol Intake	Subjective Feature	Categorical Code	-	0: Does not drink alcohol 1: Drinks Alcohol
active	Lifestyle	Subjective Feature	Categorical Code	-	0: Sedentary Lifestyle 1: Active Lifestyle
PP	Pulse Pressure	Derived Feature	integer	Millimeter(s) of Mercury (mmHg)	Any Integer Value > 0
MAP	Mean Arterial Pressure	Derived Feature	float	Millimeter(s) of Mercury(mmHg)	Any Floating Value > 0

Table 1. Table describing the dataset features

tained from the ADVANCE study and mainly consisted of the four blood pressure indices: Systolic Blood Pressure (SBP), Diastolic Blood Pressure (DBP), Pulse Pressure (PP) (defined as mean (SBP)-mean (DBP)), and Mean Arterial Pressure (MAP) (defined as (mean (DBP)+1/3 (mean (PP)) along with demographic information like age, gender, total Cholesterol (Hb1Ac), and duration of mellitus. (type-2 diabetes) and information about the habits of a person, like smoking and alcohol. Each of the Blood Pressure indices used had two values viz the achieved (the values achieved throughout the study through controlled and the baseline values (values recorded at the start of the experiment).

COX proportional Hazard Regression Models were trained on the dataset to determine the Hazard Ratio (HR) and the 95% confidence interval (95% CI). For univariate analysis, AUC was used to assess the effectiveness of each BP index in predicting CVD. RIDI (Relative Integrated Discrimination Improvement) was used for multivariate analysis to analyze the model's discriminative capacity increase when new variables were introduced.

The conclusions and results obtained from the study were that if individual BP indices are considered, then for baseline BP measurements, the derived BP indices, like Pulse Pressure (PP) and Mean Arterial Pressure (MAP), were more effective in predicting Cardiovascular events. In the case of achieved BP estimates, it was found that the Systolic Blood Pressure (SBP) and Mean Arterial Pressure (MAP) were found to be better. Overall, if a combination of multiple BP indices is used, it was found that Systolic Blood Pressure (SBP) and Mean Arterial Pressure (MAP) were better at predicting cardiovascular events. It was also found that, among all Blood Pressure indices, Diastolic Blood Pressure (DBP) was the worst for predicting cardiovascular events, mainly because beyond 50 to 60 years, the Diastolic Blood Pressure either becomes constant or decreases.[2]

4. Datset Description

4.1. Size and Shape of the dataset

The dataset used for this project has been obtained from Kaggle.[3]. The original dataset consisted of 70,000 records and 13 columns. Two additional columns have been added for two derived features. Additionally, the first column of the dataset is the ID, which has been dropped, making the total number of columns in the modified dataset equal 14.

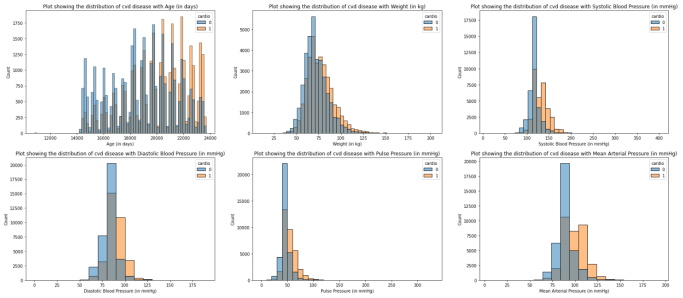
4.2. Description of the features of the dataset

Every feature in the dataset is divided into one of the four categories mentioned below:-

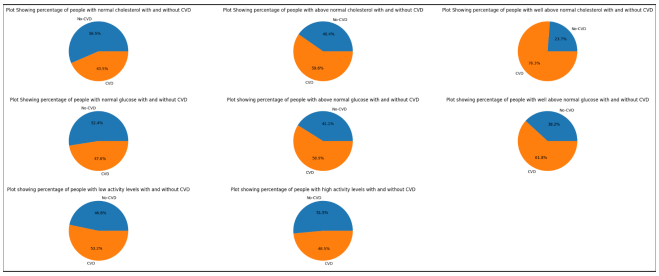
- **Objective Feature :** Factual Information
- **Examination Feature :** Results of Medical Examination.
- **Subjective Feature :** Information given by the patient
- **Derived Features :** Features derived from already existing features

Derived Features The two derived features in our dataset include the Mean Arterial Pressure (MAP) and the Pulse Pressure(PP), which are defined as:-

- PP = Systolic Blood Pressure(SBP)-Diastolic Blood Pressure(DBP)



(a) Figure 1



(b) Figure 2

- MAP = Diastolic Blood Pressure(DBP)+$\frac{1}{3}$Pulse Pressure (PP)

4.3. Exploratory Data Analysis

4.3.1 Univariate Analysis

Univariate analysis was performed to analyze the effectiveness of each feature in predicting the occurrence of CVD. The results obtained from each of the plots have been summarized below:-

Box-Plots

The box plots shown in **Figure-4** indicate that many values lie beyond the $\pm$ 1.5 IQR mark and are outliers. Thus, performing outlier detection is a must before training the models.
It can also be observed that among the BP indices, the highest number of outliers are shown by the Pulse Pressure. In contrast, the Mean Arterial Pressure (MAP) and the Systolic Blood Pressure (SBP) demonstrate the lowest number of outliers.
It can also be observed that the people having CVD have a higher median value of BP indices and weight as compared to those not having CVD, indicating the possibility of a positive correlation between these metrics and cardiovascular diseases.

Histograms

Histograms were used to closely analyze the distribution of the participants, with or without CVD, in specific ranges of the numerical features. From the histograms shown in the **Figure-1**, it was concluded that, as the age of the person increases, the number of people with CVD is higher, compared to the number of people without CVD, especially for people beyond the age of 22500 days (61 years). Similarly, as the person's weight increases, the proportion of people having CVD is significantly higher compared to those not having CVD in the higher weight ranges, especially in the 75 kg and beyond range. A similar trend was also observed in the Blood Pressure indices, wherein the participants in the higher blood pressure ranges (above the standard value) had a higher number of people with CVD than those without CVD.

Pie-Charts

We used pie charts to analyze the distribution of participants, with or without CVD, in various categories of categorical features.

From the pie charts shown in **Figure-2**, it can be concluded that the proportion of people not having CVD in the normal cholesterol level category is higher as compared to the people having CVD in the above-normal and well-above normal cholesterol categories the proportion of people, having CVD, is higher, as compared to those not having CVD. A similar trend can be seen in the glucose level as well. It can also be seen that, out of the people with an active lifestyle, the proportion of people not having CVD is higher than those with CVD. Similarly, the people with a sedentary lifestyle have a higher proportion of people who have CVD than the ones not having CVD.

4.4. Multivariate Analysis

To determine the correlation between the features and how effective a combination of elements was in predicting the CVD, multi-variate analysis was done using pair plots and co-relation heatmaps. The conclusions obtained from each of these plots have been summarized below

.

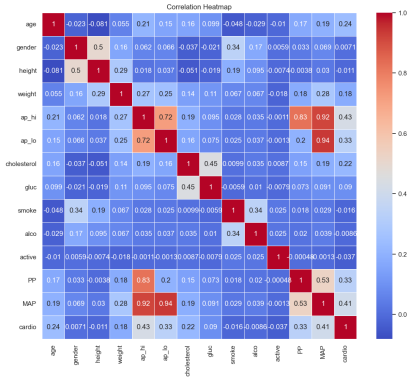
Pair Plots and Co-relation Heat Map

The correlation heatmap shown in **Figure-3** the correlation between the different features of data, including the target attribute. The values and color of each cell indicate the degree of correlation. Gender and height have a moderate correlation(around 0.5), with males having greater height than females

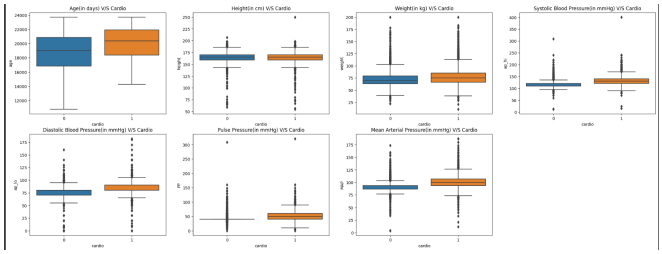
- ap_lo and ap_hi have a strong correlation(close to 1) as greater ap_lo means a greater ap_hi
- PP and MAP both have a strong correlation with ap_hi and ap_lo
- Age and cardio have a somewhat moderate(around 0.3) correlation
- Ap_hi and map have a moderate correlation (around 0.5) with CVD as higher blood pressure generally means a greater risk of CVD
- PP ap_lo and cholesterol have somewhat moderate(around 0.3) correlation with CVD
- PP and MAP have a moderate correlation with each other

4.5. Pre-processing

From the box plots, it can be seen that some outliers are present in the data. Moreover, it can also be seen that specific columns, such as ap_hi and ap_lo, have few negative values that are not possible and few values that are much beyond the possible blood pressure limit. These might be present because of data entry errors. Hence, detecting and removing these values is imperative for finding the best-fit model.



(a) Figure 3



(b) Figure 4

4.5.1 Outlier detection

The first step in our outlier detection involved removing all the negative values and the values in the columns of ap_hi, ap_lo, PP, and MAP that were beyond 500 mmHg. Once this step was done, the number of records in the dataset reduced from 70,000 to 68,727.

After performing this initial detection of out-of-bounds values, the box plots still showed the presence of outliers, and hence, Z-score and Local Outlier Factor methods were used to remove them. Thus, two copies of the datasets were created. On one of the datasets, **the Z-Score method** was used to detect the outliers, with a bound of ± 2.75. In each column, the value whose Z-Score lay beyond this limit was considered an outlier and was removed from the dataset. The final number of records after using the Z-Score method reduced from 68,727 to 65,048

The Local Outlier Factor method was applied in the second copy with a 20 percent contamination rate and 20 neighbors. After using LOF, the final number of records were reduced from 68,727 to 56,000.

4.5.2 Data Standardization

The Standard Scaler library of scikit learn was used to scale the numerical value columns of the filtered data (subtracting each value of the column from the mean of the column and then dividing by the standard deviation for that feature). This ensured that the standardized numerical features had a mean of 0 and a standard deviation of 1 to nullify the effect of different scales of measurements of various physical quantities.

4.5.3 Data Encoding

Initially, label-based encoding was used for the categorical features, followed by one-hot encoding, and the models were trained. With label-based encoding the dataset had 14 features and with one-hot encoding the number of features increased to 16.

5. Models and Methodologies

Since the problem being addressed in this paper is a binary classification problem, classification algorithms viz: Naive Bayes, Logistic Regression, and Support Vector Machines (SVM) were used.

For applying each of these models, the scikit-learn Python library was used. For using

5.1. General Flow

Two copies of the dataset were made. Z-Score and Local Outlier Factor (LOF) methods were used for outlier detection on the first and second copies of the dataset, respectively, as mentioned above.

Data standardization was performed on each dataset copy after doing a **70:30** train-test split. PCA was applied to the one-hot encoded dataset to improve the performance further. To determine the optimal number of components to which the data should be reduced for training each model, K-Fold Cross Validation was used, with the number of folds set to 10.

The results obtained for each round of re-training on the two copies of the dataset have been summarized in the result and analysis sections. The specific details regarding the parameters set for training each model have been discussed in the section below:-

5.2. Models Used

Gaussian Naive Bayes

Since the data involves real values that are nearly normally distributed, as shown from the histograms in **Figure-1**, Gaussian Naive Bayes (GNB) was used as an initial starting point for finding the most optimal model. Gaussian Naive Bayes was applied with the default parameters of the scikit-learn mainly to determine the baseline estimates of the expected accuracy and F1-Score. Default parameters(i.e smoothing $=1^{-9}$) were used for training the model

Logistic Regression

After Gaussian Naive Bayes, logistic regression was applied to the dataset. The threshold value for the cut-off probability was determined using the ROC curve. Although the optimal threshold value obtained using the ROC curve was 0.7 for the highest recall, the best F1 score was obtained at a threshold value of 0.5. Hence, the threshold probability used was 0.5.
L2 regularisation was used with default values of parameters for regularisation (which is 1 in scikit learn library), and the learning rate (which is 0.1) was used for training the model.

Support Vector Machines

The Gaussian kernel was used for support vector machines, as it gave the best performance among all other choices. The default values of the margin (gamma), which is $\frac{1}{(numberof features * varianceof data)}$ and regularisation parameters were used to train the model.

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
	Standard Data	0.7209	0.7191	0.7570	0.7592	0.6159	0.6148	0.6792	0.6794
Gaussian Naive Bayes	One Hot Encoded	0.7169	0.7139	0.7520	0.7529	0.6113	0.6088	0.6744	0.6732
	One Hot Encoded+ PCA(components=3)	0.7126	0.7105	0.7611	0.7633	0.5841	0.5825	0.6610	0.6608
	Standard Data	0.7244	0.7206	0.6480	0.6433	0.7444	0.7448	0.6928	0.6903
Logistic Regression	One Hot Encoded	0.7251	0.7219	0.6470	0.6438	0.7461	0.7468	0.6930	0.6915
	One Hot Encoded+ PCA(components=13)	0.7251	0.7219	0.6470	0.6438	0.7461	0.7468	0.6930	0.6915
	Standard Data	0.7312	0.7281	0.6391	0.6338	0.7620	0.7642	0.6952	0.6929
Support Vector Machines	One Hot Encoded Data	0.7325	0.7274	0.6450	0.6393	0.7608	0.7595	0.6981	0.6942
	One Hot Encoded+ PCA(components=15)	0.7330	0.7276	0.6456	0.6395	0.7614	0.7597	0.6988	0.6944

Table 2. Metrics on the dataset cleaned using Z-Score

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
	Standard Data	0.7275	0.7313	0.7763	0.7889	0.6403	0.6493	0.7018	0.7123
Gaussian Naive Bayes	One Hot Encoded	0.7236	0.7290	0.7712	0.7857	0.6371	0.6476	0.6978	0.7100
	One Hot Encoded+ PCA(components=3)	0.7217	0.7253	0.7694	0.7823	0.6343	0.6426	0.6954	0.7056
	Standard Data	0.7311	0.7350	0.6810	0.6902	0.7575	0.7688	0.7172	0.7274
Logistic Regression	One Hot Encoded	0.7316	0.7371	0.6808	0.6914	0.7586	0.7716	0.7176	0.7293
	One Hot Encoded+ PCA(components=13)	0.7316	0.7370	0.6808	0.6913	0.7586	0.7715	0.7176	0.7294
	Standard Data	0.7370	0.7388	0.6821	0.6892	0.7669	0.7760	0.7220	0.7300
Support Vector Machines	One Hot Encoded	0.7378	0.7392	0.6834	0.6900	0.7675	0.7761	0.7230	0.7305
	One Hot Encoded+ PCA(components=12)	0.7377	0.7388	0.6838	0.6899	0.7672	0.7754	0.7231	0.7302

Table 3. Metrics on the dataset cleaned using Local Outlier Factor (LOF)

6. Results and Analysis

6.1. Results

The results (evaluation metrics) after applying the three models to each of the two copies of the dataset using the specified methodology are summarized in **Table-2** and **Table-3**.

6.2. Analysis of the results obtained

Based on the methodology followed and the evaluation metrics used to evaluate the models, the following analysis was obtained:-

- If either accuracy or F1 score is used as a metric to evaluate the models, then a clear-cut trade-off can be observed in the precision and recall of the models simply because the increase in accuracy comes either at the expense of a lower recall or a lower precision.
- In our case scenario, we are more concerned by false negatives than by false positives. Thus, unequal weight is given to misclassification. Thus, a model having higher recall is more desirable. Out of all the models trained, the best **Recall** was observed in the case of Support Vector Machines (0.7754) on the one hot encoded data, with PCA cleaned using LOF outlier detection method, which was closely followed by logistic regression with a recall of 0.7715 on the same dataset and following the same methodology.
- Since we are giving unequal weights to misclassification, the F1 Score would be an overall better metric to gauge the model's performance as compared to accuracy. It can be observed that the F1 scores of all the models trained on the data with LOF as a method of outlier removal were higher than the models trained on data with Z-Score as the outlier removal method.

- Among the values for the F1 score achieved on the models trained on the dataset with LOF as the outlier removal method, the best performance was achieved by Support Vector Machines (SVM) with an F1 score of 0.7305 on the testing data after one hot encoding was applied on the dataset. This was closely followed by Logistic Regression, which was able to achieve an F1 score of 0.7293 on the testing data.

7. Conclusion

- From the above analysis, it can be concluded that out of Naive Bayes, Logistic Regression and Support Vector Machines, the best performance (considering F1 score as the overall evaluation metric) was achieved by Support Vector Machines (0.7305) followed by Logistic Regression (0.7294) and Naive Bayes (0.7274). This makes SVM the best model if only model performance is the concern.

- If training times are also taken into consideration, then that makes logistic regression is the best model since it offers an almost comparable F1 score as SVM but has much lower training times.

- Further improvement of the F1 score can be achieved by using other classification models like Decision Trees and Random Forests. The existing models can also further be improved by hyper-parameter tuning via K-Fold Cross Validation, which will be taken up after the midsem evaluation.

References

[1] Karolina Drożdż, Katarzyna Nabrdalik, Hanna Kwiendacz, Mirela Hendel, Anna Olejarz, Andrzej Tomasik, Wojciech Bartman, Jakub Nalepa, Janusz Gumprecht, and Gregory YH Lip. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach. *Cardiovascular Diabetology*, 21(1):240, 2022.

[2] Andre-Pascal Kengne, Sébastien Czernichow, Rachel Huxley, Diederick Grobbee, Mark Woodward, Bruce Neal, Sophia Zoungas, Mark Cooper, Paul Glasziou, Pavel Hamet, et al. Blood pressure variables and cardiovascular risk: new findings from advance. *Hypertension*, 54(2):399–404, 2009.

[3] Svetlana Ulianova. Cardiovascular disease dataset, Jan 2019.