Cardiovascular Disease Prediction using Machine Learning Models

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

1. Abstract

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, various classification models 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 and the optimal hyperparameters for each model were determined using the K-Fold Cross Validation method.

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. [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. 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 Kaggle 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. 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. 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. The dataset was split into 80:20 for training and 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 analyse 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 obtained 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%

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

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 (including the target column).

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

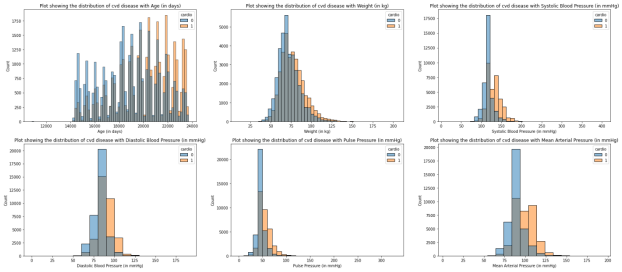
4.3. Exploratory Data Analysis

4.3.1 Univariate Analysis

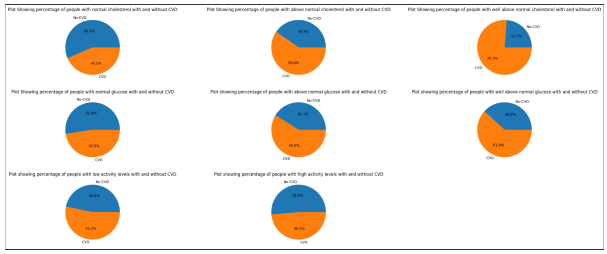
Univariate analysis was performed to analyze the effectiveness of each feature in predicting the occurrence of CVD. The results obtained from each of the plots have been summarised below:-

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,



(a) Figure 1



(b) Figure 2

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 analyse 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 summarised 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 a 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.

4.5.1 Outlier detection

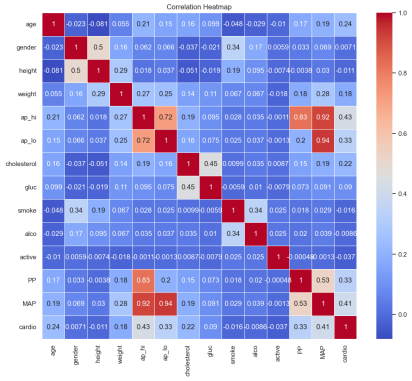
The first step in our outlier detection involved removing all the negative values and the values in the columns of ap_hi, ap_lo, PP, and MAP that were beyond 500 mmHg. Once this step was done, the number of records in the dataset reduced from 70,000 to 68,727.

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

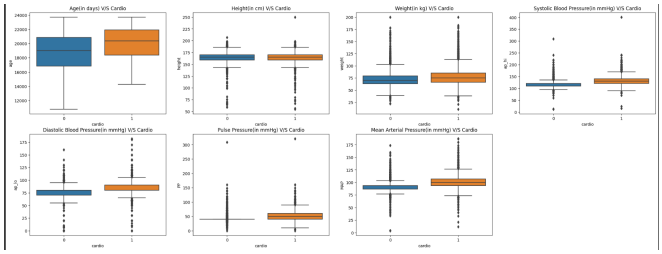
The Local Outlier Factor method was applied in the second copy with a 20 percent contamination rate and 20 neighbors. After using LOF, the final number of records were reduced from 68,727 to 56,000.

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



(a) Figure 3



(b) Figure 4

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 13 features, and with one-hot encoding, the number of features increased to 15.

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 Machine (SVM), Decision Tree, Random Forest (and Xg-Boost) along with Multi-Layer Perceptron were used.

For applying each of these models, the scikit-learn Python library was used. For applying MLP, however, the TensorFlow library was used in order to facilitate faster convergence of the model.

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 standardisation 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 5. However, for MLP and Random Forest, PCA was not applied; the reasons for the same are discussed in the subsequent sub-sections for the particular models.

K-Fold Cross Validation was applied to determine the optimal hyper-parameters for each model on the data set cleaned using LOF after applying PCA except for Random Forest and MLP, for which PCA was not applied. For these models, the optimal hyper-parameters were directly determined on the One-Hot encoded dataset.

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

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.

Logistic Regression

After Gaussian Naive Bayes, logistic regression was applied to the data set. The threshold value for the cut-off probability was set at 0.5. L2 regularisation was used for training all the models, and K-Fold Cross Validation was applied to the dataset cleaned using the LOF method for outlier detection to determine the optimal value of the regularisation constant and the best solver. The optimal parameters obtained are as follows:-

Solver: newton-cg, Regularization Strength: 10

Support Vector Machines

Soft Margin Support Vector Machines were trained with an 'rbf' kernel and a regularisation strength of 1 as the default parameters. This was followed by the application of PCA one hot encoded data cleaned using LOF for outlier detection. K-Fold cross-validation was applied to the training data, and the value of the optimal parameters obtained are as follows:-

Regularisation Strength: 1, Kernel: rbf

Decision Trees

Decision Trees were used for classification, and in order to reduce over-fitting on the training data, pre-pruning was used. K-Fold Cross Validation was used on the training data after applying PCA on the dataset cleaned using the LOF outlier detection method. The hyperparameters obtained were as follows:-

Max Depth: 7, Min Samples Split: 5, Min Samples Leaf: 1

Random Forests and Xg Boost

Random forests were used for binary classification and after using k-fold validation on one-hot encoded data. PCA was not used because the best components were found to be 1, but there are trade-offs as the model would not capture the complexity of the dataset. The best hyper-parameters were:

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Naïve bayes	Standard Data	0.72097	0.71919	0.75703	0.75921	0.61599	0.61486	0.67927	0.67945
	One Hot Encoded	0.71690	0.71396	0.75209	0.75294	0.61132	0.60882	0.67444	0.67326
	One Hot Encoded + PCA(components=3)	0.71260	0.71053	0.76114	0.76335	0.58417	0.58257	0.66102	0.66082
Logistic	Standard Data	0.72444	0.72067	0.64800	0.64334	0.74442	0.74482	0.69287	0.69037
	One Hot Encoded	0.7251	0.72195	0.64704	0.64387	0.74614	0.74680	0.69306	0.69152
	One Hot Encoded + PCA(components=13)	0.7251	0.72195	0.64704	0.64387	0.74614	0.74680	0.69306	0.69152
SVM	Standard Data	0.73120	0.72810	0.63916	0.63381	0.76209	0.76423	0.69523	0.6929
	One Hot Encoded	0.73252	0.7274	0.64511	0.63931	0.76088	0.75952	0.69823	0.69425
	One Hot Encoded + PCA(components=14)	0.73307	0.72769	0.64580	0.63963	0.76148	0.75980	0.6988	0.69456
Xgboost	Standard Data	0.72575	0.72282	0.76239	0.76486	0.62217	0.61708	0.68518	0.68307
	One Hot Encoded	0.72611	0.72144	0.76234	0.76222	0.62332	0.61698	0.68586	0.68195
	One Hot Encoded + PCA(components=8)	0.72448	0.72047	0.76403	0.76483	0.61581	0.6101	0.68196	0.67875
Decision tree	Standard Data	0.72554	0.72262	0.70317	0.69659	0.71860	0.72093	0.71080	0.70855
	One Hot Encoded	0.72551	0.72272	0.70326	0.69690	0.71852	0.72095	0.71081	0.70872
	One Hot Encoded + PCA(components=12)	0.72336	0.71734	0.69845	0.69076	0.71737	0.71546	0.70779	0.70289
Random forest	Standard	0.99806	0.70458	0.99693	0.67171	0.99903	0.70429	0.99798	0.68761
	One Hot Encoded	0.99780	0.70602	0.99665	0.6720	0.99876	0.70635	0.99770	0.68876
MLP	Standard Data	0.73507	0.72482	0.69964	0.68992	0.73522	0.72750	0.71699	0.70821
	One Hot Encoded	0.74256	0.72564	0.67057	0.65488	0.76388	0.74710	0.71419	0.69795

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

max depth=15, max features='log2', min samples leaf=4,n estimators=300, Bootstrap=True

XgBoost was used for further boosting the random forest, and optimal hyper-parameters were calculated on one-hot encoded training data after applying PCA on the dataset cleaned using the LOF method. The optimal parameters are as follows:

learning rate: 0.2, max depth: 3, n estimators: 100

Multi Layer Perceptron

Multi-Layer Perceptron was the only Deep Learning Model applied, with 2 hidden layers. Since the task at hand was a binary classification task, the Sigmoid Activation function was used at the output layer, whereas for the hidden layers, the ReLU activation function was used. The default layer sizes for each of the hidden layers were chosen to be 200, and the number of epochs was fixed to 20. PCA was applied on the input one-hot encoded dataset in the case of MLP so that it is able to effectively capture the non-linear patterns in the data and preserve the local and global information. The hyperparameter tuning using K-Fold Cross Validation, in this case, was directly done on the one hot encoded dataset cleaned using the LOF method, and the parameters obtained were as follows:-

Neurons in hidden layer-1 : 300, Neurons in hidden layer-2: 250

6. Results and Analysis

6.1. Results

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

Note that the metrics reported in **Table-3** for the column for One Hot Encoded + PCA have been reported after parameter tuning of the models. For the models which do not have this metric reported (which include MLP and Random Forests), retraining the model with optimal hyper parameters on the One Hot Encoded Data itself.

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 misclassifi-

Model	Method	Accuracy		Precision		Recall		F1 Score	
		Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data	Training Data	Testing Data
Naïve Bayes	Standard Data	0.72788	0.73089	0.77897	0.77872	0.64221	0.65262	0.70401	0.71012
	One Hot Encoded	0.72385	0.72555	0.77291	0.77273	0.64004	0.64686	0.70023	0.70421
	One Hot Encoded+ PCA(components=3)	0.72141	0.73022	0.77009	0.77765	0.63747	0.65238	0.69753	0.70953
Logistic Regression	Standard Data	0.73221	0.73495	0.68475	0.69379	0.76007	0.76043	0.72045	0.72558
	One Hot Encoded	0.73229	0.73610	0.68408	0.69367	0.76059	0.76240	0.72031	0.72642
	One Hot Encoded+ PCA(components=13)	0.73229	0.73622	0.68403	0.69391	0.76062	0.76246	0.72029	0.72658
SVM	Standard Data	0.73666	0.73883	0.68243	0.69199	0.76898	0.76795	0.72313	0.72800
	One Hot Encoded	0.73822	0.73919	0.68588	0.69451	0.76957	0.76707	0.72532	0.72899
	One Hot Encoded+ PCA(components=13)	0.73837	0.73889	0.68691	0.69499	0.76921	0.76628	0.72574	0.72890
Xgboost	Standard Data	0.73305	0.73659	0.76171	0.76237	0.68434	0.69511	0.72095	0.72719
	One Hot Encoded	0.73380	0.73701	0.76413	0.76553	0.68238	0.69091	0.72094	0.72631
	One Hot Encoded+ PCA(components=12)	0.74697	0.73853	0.76718	0.75645	0.71481	0.71132	0.74007	0.73319
Decision Tree	Standard Data	0.73479	0.73877	0.72631	0.73545	0.74195	0.74429	0.73405	0.73984
	One Hot Encoded	0.73479	0.73877	0.72631	0.73545	0.74195	0.74429	0.73405	0.73984
	One Hot Encoded+ PCA(components=12)	0.73747	0.72298	0.67360	0.65964	0.75303	0.73985	0.71110	0.69745
Random forest	Standard Data	0.99787	0.71525	0.99675	0.71708	0.99902	0.71855	0.99788	0.71781
	One Hot Encoded	0.79112	0.74132	0.75472	0.71420	0.81684	0.75932	0.78455	0.73607
MLP	Standard Data	0.74419	0.73743	0.71996	0.72116	0.75980	0.74950	0.73935	0.73505
	One Hot Encoded	0.74629	0.73786	0.73172	0.73040	0.75677	0.74543	0.74403	0.73784

Table 3. Metrics on the dataset cleaned using LOF

cation. 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 Machine (0.76921)

- 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.
- PCA was not applied to MLP since it is capable of handling complex relationships between the features and the data points. Applying PCA preserves the global structure, whereas the local structure is distorted. However, the initial layers of the MLP are capable of recognising the local structure within the data; hence, distorting it might render these initial layers useless, resulting in lower performance metrics.

7. Conclusion

From the above analysis, it can be concluded that out of all the models applied, the best performance (considering the F1 score as the overall evaluation metric) was achieved by the Decision Tree (0.73984) followed by Multi-Layer Perceptron (0.73784).

If we observe the accuracies of the model, we can observe that the best accuracy was observed in the case of Random Forest, with an accuracy of 0.74132, which was closely followed by 0.73919.

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.