A HEART DISEASE CLASSIFICATION MODEL ANALYSIS MADE BY DIFFERENT CLASSIFICATION ALGORITHMS AND DATASETS.

Akhunova Robiya[1]
Master of Technology in Computer Science and Engineering
Sharda University Uzbekistan
e-mail:abdukhalilovarobiya@gmail.com

Puladjonov Otabek[2]
Master of Technology in Computer Science and Engineering
Sharda University Uzbekistan
e-mail:otabekpolatjonov@gmail.com

Dr. Pooja[3]
Faculty of Engineering and Technology Sharda University Uzbekistan,
e-mail: pooja@shardauniversity.uz
Andijan, Uzbekistan

Abstract: Heart disease is a leading cause of death worldwide, necessitating the development of accurate and efficient diagnostic methods. Data mining techniques including classification have shown promise in analyzing large datasets and identifying patterns that can assist in the early detection of heart disease. In this report, a Heart diseases dataset is analyzed using different classification algorithms such as Random Forest, K-Nearest Neighbors, Support vector machines, Naïve Bayes and Logistic regression and more enhancement is performed where relevant. The dataset for experiment used here is Cleveland Heart Disease dataset available on UCI machine learning repository.

Key words: Datamining, Classification, Heart disease, random forests, K-Nearest neighbors (KNN), Support Vector Machines (SVM), Naïve Bayes and Logistic Regression, Cleveland Heart Disease dataset

I. INTRODUCTION:

Heart disease is a broad term for a number of disorders affecting the circulatory system that can be difficult to correctly diagnose. The World Health Organization (WHO) lists coronary heart disease (CHD) among the world's most serious illnesses. Approximately 17.9 million individuals lose their lives to CHD every year, according to the WHO. Angina pectoris, myocardial infarction, and hyperlipidemia are all part of CHD. Medical professionals typically use the results of blood tests, angiography, electrocardiography, and sonography to make diagnosis. Although it is difficult to diagnose CHD in the early stages of the disease, early detection is crucial for successful treatment. On the other hand, diagnoses are based on the personal experiences and knowledge of medical professionals regarding the illness, which raises the possibility of mistakes, delays the proper course of therapy, lengthens the duration of treatment, and significantly raises costs. Numerous studies on clinical decision support systems have been carried out in an effort to address these issues. Since many issues are hard to answer analytically in a practical amount of time, researchers are attempting to find suitable solutions in a reasonable amount of time through search strategies. It is discovered that data mining, which aids in the extraction of patterns throughout the

information discovery process in databases where intelligent approaches are employed, is crucial to employ in order to address the issue. Classification, regression, clustering, rule creation, finding association rules, summarization, dependency modeling, and sequence analysis are common tasks in modern data mining practice. In this study, we want to examine a Cleveland dataset to determine its validity as a source for a model that accurately reports a person's presence of CHD. The outputs of the machine learning models can also be made easier to understand by using graph-based visualization approaches. To get the desired outcomes, model should immediately be trained ,analyzed the association between the given features' indices, and see how each impacts the others.

II. CLASSIFICATION OF HEART DISEASE DATASET AND PREPROCESSING:

This experiment takes into account medical data pertaining to heart disorders. This dataset, which includes approximately 4000 entries of patient information, was taken from the Cleveland Database, an open-source dataset available to the public. In a study utilizing the Cleveland dataset for cardiac disorders, which has 303 cases and was conducted using 10-fold Cross Validation while taking into account 14 variables and implementing 5 different algorithms, it was shown that Random Forest and Gaussian Naïve Bayes provided the highest accuracy of 91.2 percent. The Cleveland dataset focuses on categorizing individuals with cardiac disorders into normal and abnormal groups.

For the purpose of training and assessing neural network models, a comprehensive dataset comprising a range of heart disease-related features is required, including age, gender, type of chest pain (typical angina 1, 2, 3, 4, atypical angina, non-anginal pain, asymptotic), resting blood pressure (in mm Hg), serum cholesterol levels in mg/dl, fasting blood sugar (120 mg/dl), electrocardiogram (ECG) measurements, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise rate, slope of the peak exercise, number of major vessels, results of nuclear stress test, and target variable representing diagnosis of heart disease (angiographic disease status) in any major vessel are crucial. To guarantee validity and correctness, the data should be gathered from a variety of sources and appropriately preprocessed. The dataset has 303 occurrences in , which is based on the UCI Heart Disease Data Set. UCI states that although this database has 76 properties, only a subset of 14 of them are used in the published studies. It is reasoned that having too many features would create too much noise, therefore 76 features to 14 features were reduced by feature extraction.

A clean dataset is simply obtained from Kaggle because the original dataset

1	age	Age in years
2	Sex	Sex (1=male; 0=female)
3	Cp	Chest pain type Value 1: typical angina 1,2,3,4 Value 2: atypical angina Value 3: non-anginal pain Value 4: asymptotic
4	Trestbps	Resting blood sugar on admission to the hospital (in mm Hg)
5	Chol	Serum cholesterol in mg/dl
6	Fbs	Fasting blood sugar is greater than 120 mg/dl or not (1=True; 0=False)
7	Restecg	Resting electrocardiographic Value 0: normal Value 1: having ST-T wave abnormality (T wave inversions and/or ST, Elevation or depression of >0.05mV) Value 2: showing probable or definite left ventricular Hypertrophy by Ete`s criteria
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina (1=yes; 0=no)
10	Oldpeak	ST depression included by exercise relative to rest
11	Slope	The slope of the peak exercise ST segment Value 1: up sloping Value 2: flat Value 3: down sloping
12	Ca	Number of major vessels (0-3) colored by fluroscopy
13	Thal	The heart status (normal/fixed/defect/reversible defect)
14	num	Diagnosis of heart disease

contained missing values. The dataset has been divided into two categories: 20% (61 instances) for testing and 80% (242 instances) for training. This dataset was normalized in order to prevent overfitting. In order to make value 1 indicate the existence of heart disease and value 0 represent the absence of heart illness, some changes were made to the certain dataset by changing the 1s in the target column to 0s and vice versa. A variety of fascinating predicative tasks with such datasets can be performed. These traits, for instance, can be used to forecast the type of chest pain. However, the most crucial issue is that, given the patient's 13 characteristics, the aim is to determine whether or not the patient has heart disease because maintaining one's health is very important to people.

III. METHODS USED:

Numerous techniques, such as random forests, K-Nearest neighbors (KNN), support vector machines (SVM), Naïve Bayes, and logistic regression, have been used for tests throughout this project. A number of parameters, including dataset size, interpretability, feature complexity, and computational efficiency, influence the choice of algorithm.

A.Logistic Regression

A supervised learning technique called logistic regression calculates the probabilities for classification problems with two possible outcomes. It can be expanded to forecast many classes as well. The sigmoid function is used in the Logistic Regression model, which is

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

Any number can be successfully mapped by this function to a value between 0 and 1, which is used to represent the likelihood of predicting classes. As an illustration, there are two classes: those with and without cardiac disease. It will forecast that the man has heart disease if the threshold is put at 0.5. Applying the sigmoid function yields a value of 0.7, indicating that the man has a 70% chance of having heart disease.

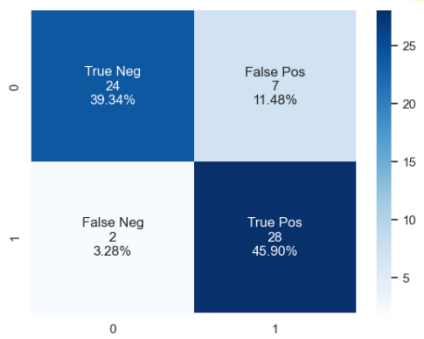


Figure 1: Confusion matrix(all feature used)

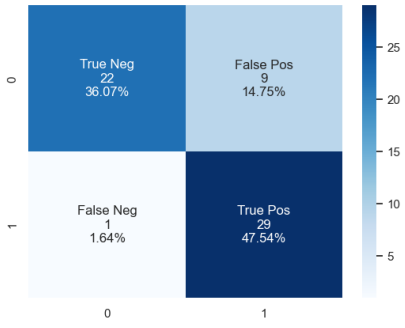


Figure 2: Confusion matrix(after dropping)

B.Support Vector Machine

SVM looks for a hyperplane that classifies the dataset in multiple dimensions, or multiple features.

The hyperplane form's equation is

$w^{\mathrm{T}}x + b = 0$

where w is a weight vector, x is input vector and b is a bias. The margin is the distance of closest points from the hyperplane and is calculated as

$$\frac{w}{||w||} * (x_+ + x_-) = \frac{w^T(x_+ + x_-)}{||w||} = \frac{2}{||w||}$$

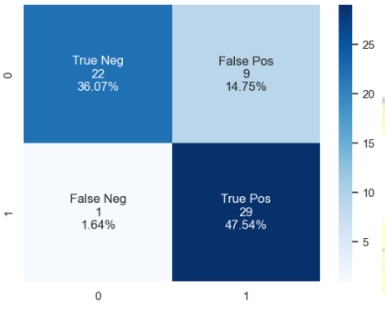
where $w.x' + b = +1$ and $w.x' + b = -1$. Our object is to maximize the margin or equivalently to minimize $||w||$. After adding loss function, the learning problem is to find a weight vector w that minimizes the cost function of

$$\|w\|^2 + c\sum_{i}^{N} \max(0,1 - y_i f(x_i))$$

And Gradient descent algorithm is able to minimize the cost function by iteratively updating the equation of

$$w_{t+1} \leftarrow w_t - \eta_t \nabla_w c(w_t)$$

where η is the learning rate.



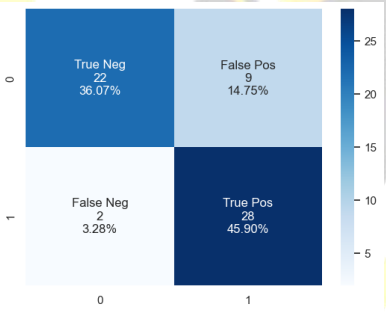


Figure 3: Confusion matrix(all feature used)

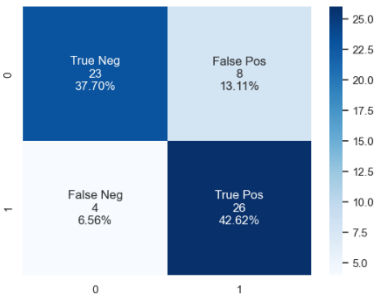
Figure 4: Confusion matrix(after dropping)

C. Naïve Bayes

Naïve Bayes assumes the independence between the features of the dataset and the Bayes Rule is

$$P(y|x) = \frac{P(x|y)}{P(x)}$$

where P(y|x) represents the likelihood that y will be classified given x data. By utilizing Bayes theory, a Naïve Bayes model can be constructed that computes probabilities from training data and then predicts based on test data attributes.



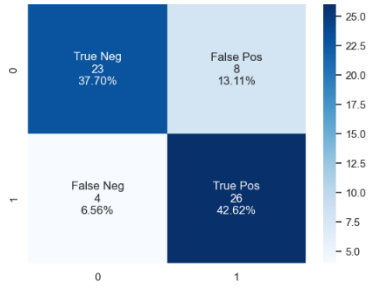


Figure 5: Confusion matrix(all feature used) *Figure 6: Confusion matrix(after dropping)*

D. Random Forest

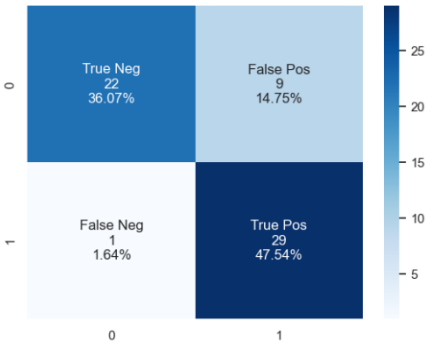
By building several decision trees during training and producing the classification or prediction (regression), Random Forest is an ensemble learning technique for classification and regression. Combining weak leaning models into a strong and reliable learning model is the aim of Random Forest. It is discovered via an internet tutorial that there are four steps that make up the Random Forest algorithm:

Step 1:Randomly draw M bootstrap samples from the training set with replacement.

Step 2: Grow a decision tree from the bootstrap samples. At each node: Randomly select K features without replacement and split the node by finding the best cut among the selected features that maximizes the information gain.

Step 3:Repeat the steps 1 and 2 T times to get T trees;

Step 4:Aggregate the predictions made by different trees via the majority vote.



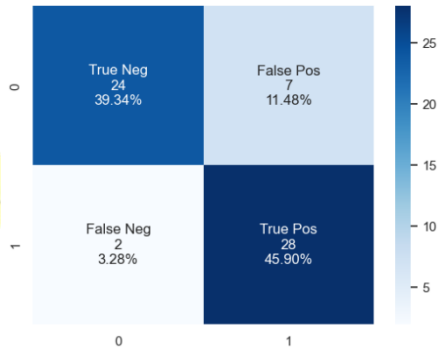


Figure 7: Confusion matrix(all feature used) *Figure 8: Confusion matrix(after dropping)*

E. K-Nearest Neighbors

One of the most fundamental yet crucial machine learning classification techniques is K-Nearest Neighbors. It is heavily used in pattern recognition, data mining, and intrusion detection and is a member of the supervised learning

domain.Since it is non-parametric—that is, it does not make any underlying assumptions about the distribution of data—it is extensively applicable in real-life circumstances (in contrast to other algorithms like GMM, which assume a Gaussian distribution of the given data). An attribute-based prior data set (also known as training data) is provided to us, allowing us to classify coordinates into groups. As is well known, the KNN algorithm aids in locating the groups or closest points to a query point. However, a measure is required here in order to identify the closest points or groups for a given query point. The following distance measurements for this purpose is employed:

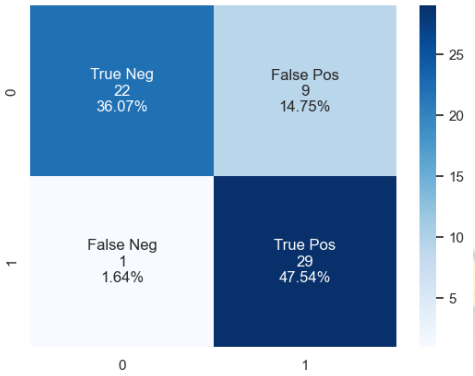


Figure 9: Confusion matrix(all features used)

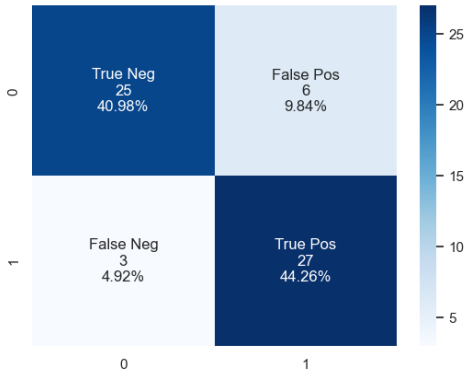


Figure 10: Confusion matrix(after dropping)

Euclidean Distance

The cartesian distance between the two locations in the plane or hyperplane is all that this represents. Another way to represent euclidean distance is as the length of the straight line connecting the two points under investigation. This metric aids in the computation of the net displacement that an object undergoes between its two states.The cartesian distance between the two locations in the plane or hyperplane is all that this represents. Another way to represent euclidean distance is as the length of the straight line connecting the two points under investigation. This metric aids in the computation of the net displacement that an object undergoes between its two states.

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

Manhattan Distance

When there is more interest in the object's total distance traveled than its displacement, the Manhattan Distance Metric is typically employed. The absolute difference between the point coordinates in n dimensions is added together to determine this measure.

$$distance = \sum_1^n |p_i - q_i|$$

Minkowski Distance

We can say that the Euclidean, as well as the Manhattan distance, are special cases of the Minkowski distance.

$$(\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

It can infered from the preceding formula that, for p = 2, it is equivalent to the Euclidean distance formula, and that, for p = 1, it yields the Manhattan distance formula.

Other distance metrics, such as Hamming Distance, are useful when handling problems that call for overlapping comparisons between two vectors whose contents can include both string and boolean values. The metrics mentioned above are the most frequently used ones when handling machine learning problems.

IV.MODEL TRAINING AND EVALUATION:

The dataset on heart disease has been preprocessed to separate the data into training and testing sets, manage missing values, and normalize features. This guarantees an impartial assessment of the categorization methods. Python and other widely used programming languages provide suitable libraries or packages that are used to implement each algorithm. The algorithm is optimized by using certain hyperparameter tuning approaches like grid search or cross-validation. To assess each algorithm's efficiency, performance metrics like accuracy, precision, recall, and F1-score are computed. Furthermore, the operational feature of the receiver.

Since the particular research involves classification, we assess the models using F1 score, accuracy, precision, and recall.

Lets first clarify what TP, FP, TN, and FN signify is. A positive outcome that the model predicts accurately is called a true positive (TP), whereas a good event that the model predicts incorrectly is known as a false positive (FP). A negative event that the model predicts accurately is called a true negative (TN), whereas a negative outcome that the model predicts incorrectly is known as a false negative (FN). The Cleveland dataset is not very sufficient, so cross-validation was not employed. The dataset is split into 80% for training and 20% for test. Here is the table of results of different methods and each evaluation of methods is explained in detail.

Methods	Train accuracy	Precision	Recall	F1 Score
Logistic Regression	0.84	0.86	0.84	0.83

Naiive Bayes	0.80	0.81	0.80	0.80
SVM	0.85	0.87	0.85	0.85
Random Forest	0.84	0.88	0.84	0.83
KNN	0.72	0.81	0.80	0.80

V. RESULTS AND DISCUSSION:

The outcomes of using various categorization methods were compared and assesed. The assessment metrics shed light on how well each algorithm performs and how well it can categorize the risks of heart disease. The distinct algorithms' advantages and disadvantages as well as their applicability to the prediction of heart disease are covered in this section. This examination takes into account various factors, including robustness,interpretability, and computational efficiency. First and foremost, each of the five algorithms are tested using a subset of the data. After that, the model, which resulted in the separation of the process into testing and training is trained.

The next stage was using Sklearn's StandartScaler modules to scale the data. In order to prevent overfitting and get the desired outcome, the data taken is successfully and completely shuffled throughout the procedure. Not a single feature has a very strong association with our target value, even after all these procedures. Additionally, a few features have a negative association and a few have a positive correlation with the goal value.

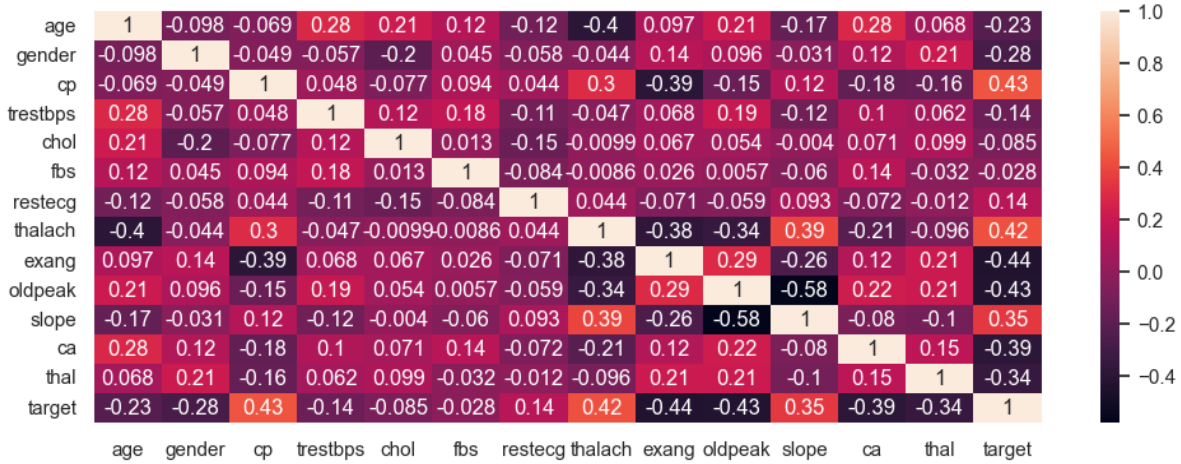


Figure 11: Correlation diagram (all features used)

In addition, we should not overlook the crucial point that the accuracy level of the module drops more during testing than it does throughout training. After seeing these results, it is decided that dropping two features that has no single correlation with our target value and affecting our model negatively, helps to check the model`s working

rate. Now two features which are #chol and #fbs, has been dropped. Before training is done, the data should be scaled and shuffled once more.

Now it shows the perfect result.

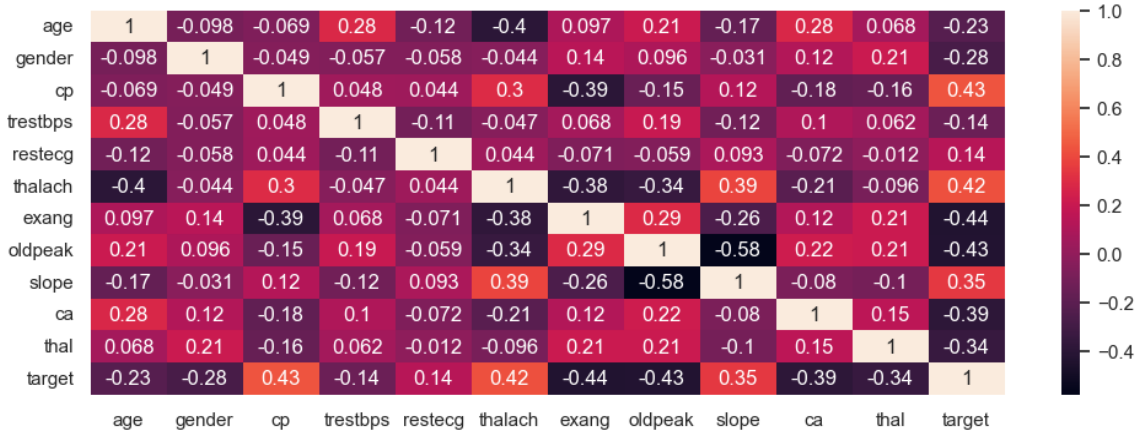


Figure 12: Correlation diagram (after dropping)

However the model now has became biased only as it is working on this dataset only. If another unfamiliar or new data of an even single person is added to the data, the model will not show the accurate information to a patient. Because those features that are dropped were the most important factors on detecting any type of the heart diseases on human organism.

Serum cholesterol and Heart disease

An international investigation showed that elevated serum cholesterol levels were linked to the highest attributable risk for the development of CVD, particularly ischemic heart disease, among all modifiable risk factors of CV illness. Many people believe that high serum total cholesterol (TC) is the primary cause of coronary atherosclerosis, and it is widely known that raised TC is linked to a higher risk of cardiovascular disease (CVD). The main job of a lipoprotein, a biochemical assembly, is to carry fat molecules in water, such as blood plasma or other extracellular fluids. There are three types of lipoproteins: very low-density lipoprotein (VLDL), low-density lipoprotein (LDL-C), and high-density lipoprotein (HDL-C).

Numerous epidemiological and interventional investigations have determined that low-density lipoprotein (LDL-C) is the primary risk factor for cardiovascular disease (CVD) due to its prominent role in the pathogenesis of atherosclerosis. In order to predict cardiovascular risk, LDL-C has largely supplanted TC as the primary lipid measurement in recent times. On the other hand, a large amount of data indicates that HDL-C, which is regarded as an anti-atherosclerotic lipoprotein, has an inverse relationship with the risk of vascular problems. The significance of HDL-C, which is independent of LDL-C levels in predicting the risk of cardiovascular disease, is becoming more widely recognized.

To our knowledge, no thorough study has examined the relationship between serum cholesterol levels and cardiovascular mortality, despite the fact that numerous studies have documented the association between serum cholesterol and risks of cardiovascular disease and coronary heart disease. Furthermore, it's unclear what risk each cholesterol level carries. Setting treatment objectives for cholesterol levels requires an understanding of the we. Therefore, in order to ascertain if total cholesterol, LDL-C, and HDL-C are risk factors for CVD mortality, a dose-response meta-analysis and systematic review are carried out.

It is conducted a systematic search of the MEDLINE and EMBASE databases in January 2021 to find cohort studies including human participants that evaluated the relationship between cholesterol and cardiovascular mortality risk between 2000 and 2020. The two topics of Medical Subject Headings phrases and associated exploding versions were the focus of the computer-based searches. The Medical Subject Headings (MeSH) cholesterol, HDL cholesterol, and LDL cholesterol were integrated into an exploding version for the first theme, cholesterol. Heart arrest, cardiovascular death, cardiac death, or cardiovascular mortality was the second theme. The Boolean operator "and" was used to integrate the two themes.

The importance of checking fasting blood sugar in detecting heart diseases

The likelihood that the glucose deaths tracked here are related to known risk factors is minimal because subjects with glucose intolerance—that is, all patients with diabetes mellitus—were excluded from the study and adjusted for various CV risk factors, including age, sex, BMI, systolic blood pressure, total cholesterol, smoking, and the use of antihypertensive medications. However, in a non-experimental investigation, it is impossible to totally rule out this option. It is still unknown what mechanism(s) glucose uses to increase the risk of death in people with CVD. Because glucose is able to non-enzymatically glycosylate low-density lipoprotein (LDL) cholesterol, other apolipoproteins, and blood clotting factors, it can directly damage the endothelium or atherosclerotic plaque of blood arteries.

VI. CONCLUSION:

Based on the results and discussion, an algorithm is recommended that demonstrates superior performance in terms of accuracy, precision, recall, F1-score analysis. The report summarizes the findings and highlights the potential clinical implications of using such algorithms to aid in the early detection and management of heart disease.

In this conducted research, the most effective and optimal algorithm that could provide sufficient model for the given dataset[1] was Random Forest algorithm. The final accuracy score was 83.6% meaning 2 out 10 predicted results does not match for

expected results which might lead to inconveniences when tested with huge amount of data.

To improve the performance of the model a dataset which has higher correlation rate of target with features especially cholesterol and fasting blood sugar in comparison to the present dataset[1] as the relativeness of features are lower than expected to target values.

VII. FUTURE DIRECTIONS:

The analysis presented in this report may have certain limitations, such as reliance on a specific dataset or generalizability issues. Future research could involve exploring other classification algorithms, incorporating additional features, or using larger and diverse datasets to further improve heart disease prediction.

In conclusion, the analysis of heart disease through different classification algorithms provides valuable insights into the potential application of machine learning techniques in healthcare. The findings can aid healthcare professionals in implementing effective strategies for early detection and management of heart disease, ultimately improving patient outcomes and reducing mortality rates.

REFERENCES:

[1]. Heart Disease UCI dataset: https://archive.ics.uci.edu/ml/datasets/heart+disease

[2]. Kaggle Heart Disease dataset: https://www.kaggle.com/datasets/priyanka841/heart-disease-aaa-uci

[3]. Cleveland Heart Disease dataset: https://archive.ics.uci.edu/ml/datasets/heart+disease

[4]. Framingham Heart Study dataset: https://www.framinghamheartstudy.org/

[5]. MIMIC-III dataset for cardiovascular diseases: https://mimic.physionet.org/

[6]. Nabeel Al Zaza, Mohammed Cherkaoui, and Ridha Soua. "A review of heart disease detection system based on machine learning with different classifier techniques." International Journal of Information Technology and Electrical Engineering, vol. 9, no. 2, 2020.

[7]. Hamzah Mahfooz, Anam Khan, and Mazin Al-Sheddy. "Detection of heart disease using machine learning techniques: A review." Journal of Computer Science and Technology, vol. 1, no. 1, 2019.

[8]. Zahra Saleem, Hafiz Tayyab Rauf, and Muhammad Umer. "A comprehensive review of heart disease detection methods using machine learning techniques." Journal of Healthcare Engineering, vol. 2020, 2020.

[9]. Ashish Gupta, Monika Sharma, and Harshul Vashist. "A review of heart disease prediction system using machine learning techniques." International Journal of Computer Applications, vol. 181, no. 31, 2018.

[10]. Shubhada Aras and Renu Balyan. "A survey on heart disease prediction using machine learning algorithms." International Journal of Advanced Research in Computer Science, vol. 9, no. 1, 2018.

[11]. Ayan Mukherjee, Aniruddha Chandra, and Samrat Roy. "A review on heart disease prediction using machine learning techniques." International Journal of Advance Research, Ideas and Innovations in Technology, vol. 5, no. 4, 2019.

[12]. Ahmed G. Radwan, Mohamed A. AbuGabal, and Mohammed S. Elbashir. "Recent trends and evaluation methods for heart disease detection systems using machine learning algorithms." International Journal of Advanced Computer Science and Applications, vol. 10, no. 4, 2019.

[13]. Sushant Agarwal and Apoorva Sapre. "A survey on heart disease prediction using machine learning techniques." International Journal of Emerging Trends in Engineering Research, vol. 7, no. 4, 2019.

[14]. Devi and Chaitanya. "Heart disease prediction system using machine learning." International Journal of Pure and Applied Mathematics, vol. 120, no. 6, 2018.

[15]. Rajdeep and Akash. "A review on machine learning algorithms for heart disease prediction." International Journal of Computer Science and Engineering, vol. 7, no. 2, 2019.