JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

E-CHIKITSALEYA

[1]Anchal Gupta , [2]Ankit Shah , [3]Farhan Yezdani , [4]Chakshu Chawla

[1]Student , [2]Student , [3]Student , [4]Student

[1]B.Tech AI & ML,

Dr. Akhilesh Das Gupta Institute of Professional Studies , New Delhi , India

Abstract: This study has been undertaken for a well-targeted and timely analysis for all health-related issues. However, for serious diseases such as Diabetes and Heart disease, traditional methods may not be adequate. Therefore, we have developed a disease prediction model based on Machine Learning algorithms for Diseases like Diabetes and Heart Disease. The dataset we have developed is a variable dataset which includes factors such as Age, Gender, Height, Weight etc. for the analysis of a specific disease to get a diagnosed output. For Diabetes , the Random Forest algorithm gives the best results with 84% accuracy. For Heart Disease , the SVM algorithm provides the best results with 86% accuracy. Our disease prediction model can be a virtual doctor in the future for timely prediction so that the patient can receive timely treatment and live a longer life.

I. INTRODUCTION

Diabetes and heart disease are two grave diseases that are both related in a way that a person having diabetes has a risk of getting a heart disease. Using Machine learning [ML] [14], we are using variable datasets to predict diabetes and also to predict heart disease. This dataset includes patient level data. There are many ways to predict diabetes and heart disease, but the recommended and used approach in this project is to use MLand compare the ability of the four MLalgorithms that are Random Forest [RF], SVM, KNN, Decision Tree [DT] to predict the chances of developing diabetes and heart disease. In this prediction we have used a training data of 300 patients under different circumstances and situations. The prediction will help in the prevention of these diseases. Diabetes is an increasingly prevalent medical condition that currently has no known cure. A person with elevated blood sugar levels is known as having diabetes mellitus. This may be brought on by the pancreas not producing enough insulin, a hormone that controls blood sugar levels, or by the body being resistant to insulin. Over time, high blood sugar levels can cause harm to several body systems including damaging the blood vessels which in turn can lead to the development of a heart disease. Heart disease refers to a condition where plaque has built up within the walls of the coronary arteries. It is common yet a serious condition that can lead to further complications. It is very common for people with high blood sugar levels to later develop a heart disease. The basis and connection between diabetes and heart disease will be very relevant to this project.

Numerous research have been carried out utilizing ML algorithms to forecast diseases based on symptoms displayed by an individual using a statistical model. Sun and Zhang [1], for instance, have covered a few deep learning and classification techniques, including SVM, DT, artificial neural networks, and RF. A logistic regression classification strategy has been created by Qawqzeh [3] for the purpose of classifying diabetic data. There training data consists of 459 patients and the testing data contains 128 patients data. By employing logistic regression, the authors were able to reach 92% classification accuracy. The model's main drawback was that it could not be validated because it wasn't compared with other diabetic prediction models. Tafa [2] divided half of the dataset into training and the other half in testing . The system was proposed with a combination of naïve Bayes and SVM calculations for diabetes expectation. The collected dataset was from three diverse areas, the system was approved on the dataset. It comprised data of 402 patients, among 80 patients 2 were diabetic. Tafa's system accomplished the exactness of 97.6%, Naïve Bayes accomplishing an precision of 94.52 and SVM accomplishing 95.52

Fig.1 E-Chikitsaleya v/s Doctor

Figure 1 depicts that this disease prediction model can be a virtual doctor in the future for timely prediction so that the patient can receive timely treatment . Sreevalli and co. utilized a irregular woodland MLcalculation to anticipate infection based on indications. This framework brought about in moo time-consuming and moo taken a toll for anticipating infections with 84.2 percent exactness. Different MLcalculations were optimized by Chen to successfully foresee a unremitting malady flare-up. The preparing information collected was found to be fragmented, provoking the utilize of a idle figure demonstrate. A novel convolutional neural network-based multimodal infection hazard expectation (CNN-MDRP) was created, accomplishing an exactness of around 94.8%. Chae utilized 4 particular profound learning models – deep neural systems (DNN), long short-term memory (LSTM), standard slightest squares (OLS), and autoregressive coordinates moving normal (ARIMA) – to screen 80 irresistible infections over 6 bunches. Among these models, DNN and LSTM illustrated predominant execution. The DNN model excelled in overall performance, while the LSTM model provided accurate predictions for larger occurrences. Haq employed a database that contained patient information related to heart disease. They utilized three different feature selection methods, namely relief, minimum redundancy, and maximum relevance (MRMR), in addition to the least absolute shrinkage and selection operator. The validation of these algorithms was done using the K-fold method. To analyze the extracted features, six different MLalgorithms were utilized, and based on whether the heart disease is present or not and the data was classified. Mohan achieved a successful development of an efficient heart disease prediction system, which achieved an accuracy level of 88.4% through the hybrid RF with a linear model (HRFLM). Similarly, Maniruzzaman utilized ML algorithms to classify diabetes disease, employing logistic regression (LR) to identify the associated risk factors. The ML-based system demonstrated an overall accuracy of 90.62%. Sanakal and Jayakumari's system , concatenated and predicted text data using SVM and fuzzy C-means clustering [FCM] . Their study concluded that FCM can effectively predict diabetes having accuracy of 94.3%.

Table1 . Related Work

Ref .No.	Algorithm Name	Applied on Disease	Objective	Accuracy
[2]	Naïve Baye's Theorem and SVM	Diabetes	a method for detecting early diabetes	94.52% 95.52%
[3]	Logistic Regression	Diabetes	a method for detecting early diabetes	92%
[4]	Decision Tree	Diabetes	a method for detecting early diabetes	73.82%
[5]	Logistic Regression	Diabetes	a method for detecting early diabetes	75.32%
[6]	Quantum Neural Network	Cardiovascular Disease	a method for detecting cardiovascular disease	98.57%
[7]	KNN	Cardiovascular Disease (CVD)	a method for detecting cardiovascular disease	90.8%
[8]	Decision Tree	Predicting Cardiovascular disease	a method for detecting cardiovascular disease	93.19%
[9]	KNN	cardiovascular disease (CVD)	a method for detecting cardiovascular disease	0.87

[Table1] displayed the related works done on diabetes and heart disease prediction models in the past. Various investigate thinks about have appeared that the larger part of ML models created for healthcare examination are centered on person infections. For

illustration, there are particular models custom fitted for analyzing liver issues, cancer, and lung issues. As a result, people trying to find exact forecasts over a wide extend of ailments must allude to different online assets. Endeavoring to foresee numerous illnesses through a single examination needs a clearly characterized prepare, and wrong comes about from certain models can posture a chance to quiet wellbeing.

The current framework emphasizes anticipating particular maladies like diabetes and heart malady, utilizing different MLcalculations such as KNN, SVM, DT, and RF. This framework accomplishes an amazing exactness rate of up to 84% for diabetes and 86% for heart infection. Rest of the paper is sorted out as takes after : The Research Methodoly for malady forecast is portrayed within the first area, it describes the strategy and algorithms for the illness forecast and their usage. The second segment pins the results of E-Chikitsaleya and the project's dialog and points the project's conclusion and its progressing improvement, whereas the third area is committed to Acknowledgement and in the last references.

1. RESEARCH METHADOLOGY

This methodology section outlines the plan and method that how the study for E-Chikitsaleya is conducted. This includes Universe of the study, Data and Sources of Data, study's variables and analytical framework. The details are as follows;

1.1 Population and Sample

E-Chikitsaleya has the data of 300 patients under different circumstances. The dataset acquired in this model is a variable dataset used for the implementation of the model , it differs from the real time data. The data used for the implementation is unique as there is no repetition of the data values to reduce the data redundancy.

1.2 Data and Sources of Data

For this study secondary data has been collected from the website of Kaggle for Diabetes as well as for the Heart Disease. The data that is used is a variable dataset for the implementation and comparative analysis of the algorithms to find the optimal one for a particular dataset.

1.3 Theoretical Framework

ML : It is the foremost widely utilized innovation within the world nowadays.

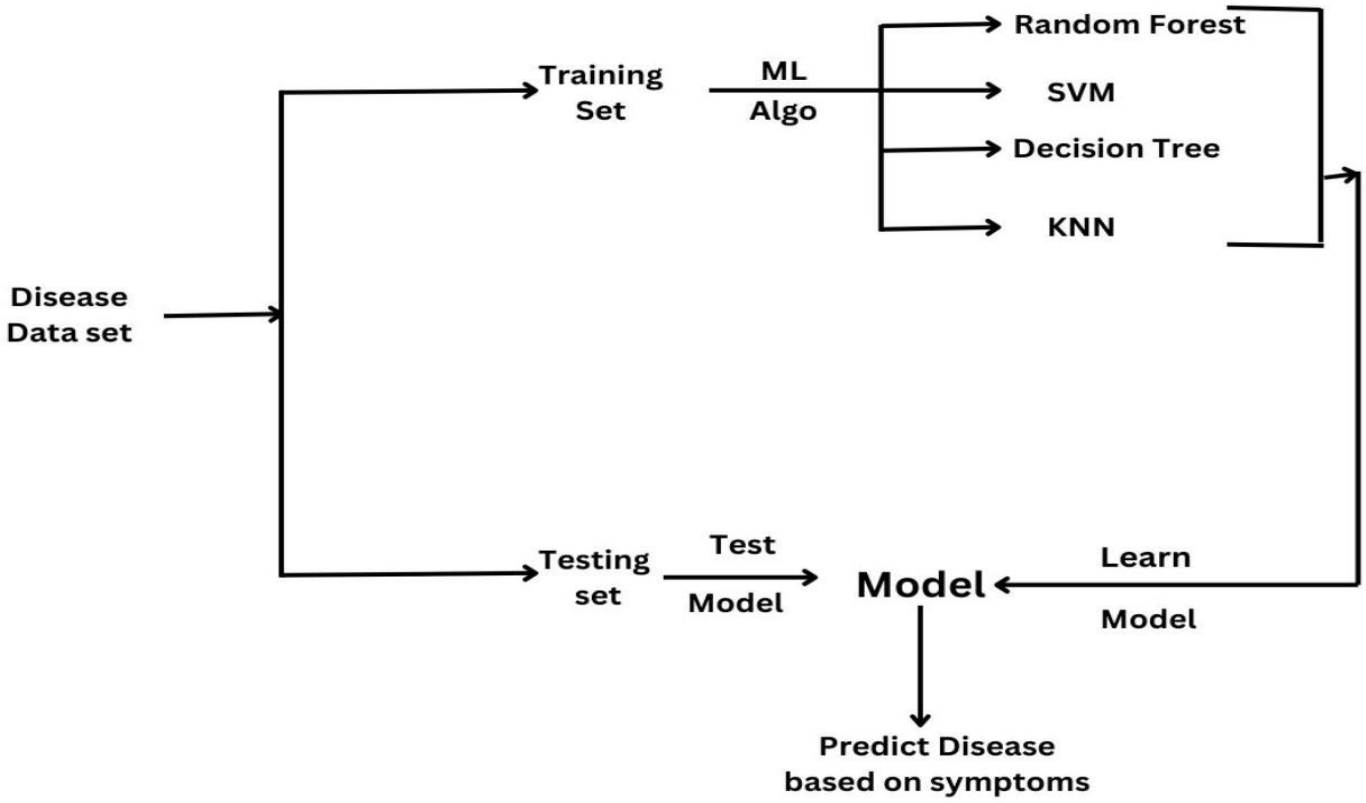

Fig.1.3 Implementation of E-Chikitsaleya

Figure 1.3 illustrates the implementation of the E-Chikitsaleya. ML is an approach to preparing the computer to memorize from past encounters or illustrations. ML can be isolated into three primary sorts: Supervised Learning [SL] , Unsupervised Learning [USL] , and Reinforcement Learning [RL] . In administered learning, the computer learns from information that has names and labels. With named information, it is simple to anticipate the modern

information. SL calculations are comparable to understudy learning under the supervision of their instructors. In USL, the machine learns from information that does not have any names or labels. USL classifies or bunches the data based on the similarties or connections between other information. Reinforcement Learning could be a sort of calculation where the machine is trained with algorithms to learn with the environment by performing certain activities and examining the information.

1.4 Statistical Tools and Econometric Tools

The section elaborates the statistical tools that is Confusion Matrix and the econometric tools that are the Algorithms used for the implementation of the model. It also outlines the softwares used for implementing the model.

1.4.1 Statistical Tools or Confusion Matrix

A tabular representation of a classification model's performance used in ML is called a confusion matrix. It displays an overview of the classifier's predictions in comparison to the dataset's actual labels. The anticipated classes are represented by the rows of the matrix, and the actual classes are represented by the columns. The elements of a confusion matrix for a binary classification task are explained below:

The quantity of cases that were anticipated to be positive but turn out to be positive is known as True Positives (TP).
The quantity of cases that, although projected to be negative, turn out to be negative is known as True Negatives (TN).
False Positives (FP): Also referred to as Type I errors, these are situations in which a good outcome was anticipated but a negative outcome occurred.

Moreover, False Negatives (FN).

Several metrics are used in ML to assess how well categorization models perform. Here's a quick rundown of each, along with formulas:

1.4.1.1 Accuracy

Accuracy is the percentage of right predictions among all the model's predictions.

Equation: Accuracy = TP + TN / (TP+TN+FP+FN)

1.4.1.2 Precision

The percentage of true positive predictions among all of the model's positive predictions is known as precision.

Equation: Precision = TP / (TP+FP)

1.4.1.3 Recall or Sensitivity

Sensitivity quantifies the percentage of actual positives in the dataset that correspond to true positive forecasts.

The formula is recall = TP / (TP+FN)

1.4.1.4 F1 Score

The percentage of accurate negative predictions among the dataset's actual negatives is measured by specificity.

Formula: Specificity = TN / (TN+FP)

1.4.2 Econometric Tools or Algorithms

Algorithms utilized in E-CHIKITSALEYA are as follows:

a. KNN:

K Nearest Neighbor (KNN) may be horrendously simple, straightforward to get a handle on, flexible and one among the highest MLalgorithms. Within the Healthcare Framework, the client will predict the

disease. In this system, the client can anticipate whether the infection will detect or not. Within the proposed framework, classifying illness in different classes that appears which illness will happen on the premise of indications.

b. Decision TREE:

A DT may be a structure that can be utilized to isolate up a huge collection of records into effectively littler sets of records by applying a arrangement of basic choice tree. With each progressive division, the individuals of the resulting sets gotten to be increasingly comparable to each other. A DT model comprises of a set of rules for isolating a huge heterogeneous population into smaller, more homogeneous (commonly exclusive) bunches with regard to a specific target.

c. Random Forest:

RF classifier could be a broadly utilized administered ml procedure that finds its essential application in classification tasks, in spite of the fact that it can moreover handle regression issues. It falls beneath the category of ensemble learning strategies. The usage and utilization of RF are straightforward, making it an perfect choice when time is constrained for model improvement.

d. SVM:

Support Vector Machine or SVM is among the foremost prevalent SL algorithm, which solves Classification and Regression problems. Fundamentally , it is utilized for Classification problems in ML. The main objective of the SVM algorithm is to make the leading choice boundary to segregate n-dimensional space into classes so it is ready to effortlessly put the given point within the correct segment or category in the near future. This choice boundary is known as the hyperplane.

1.4.3 Software used in the Model Implementation

The implementation of the model is as follows:
An Excel spreadsheet was generated from a publicly available dataset, containing a comprehensive list of symptoms associated with various diseases. Subsequently, age and gender information were included for each disease within the dataset.

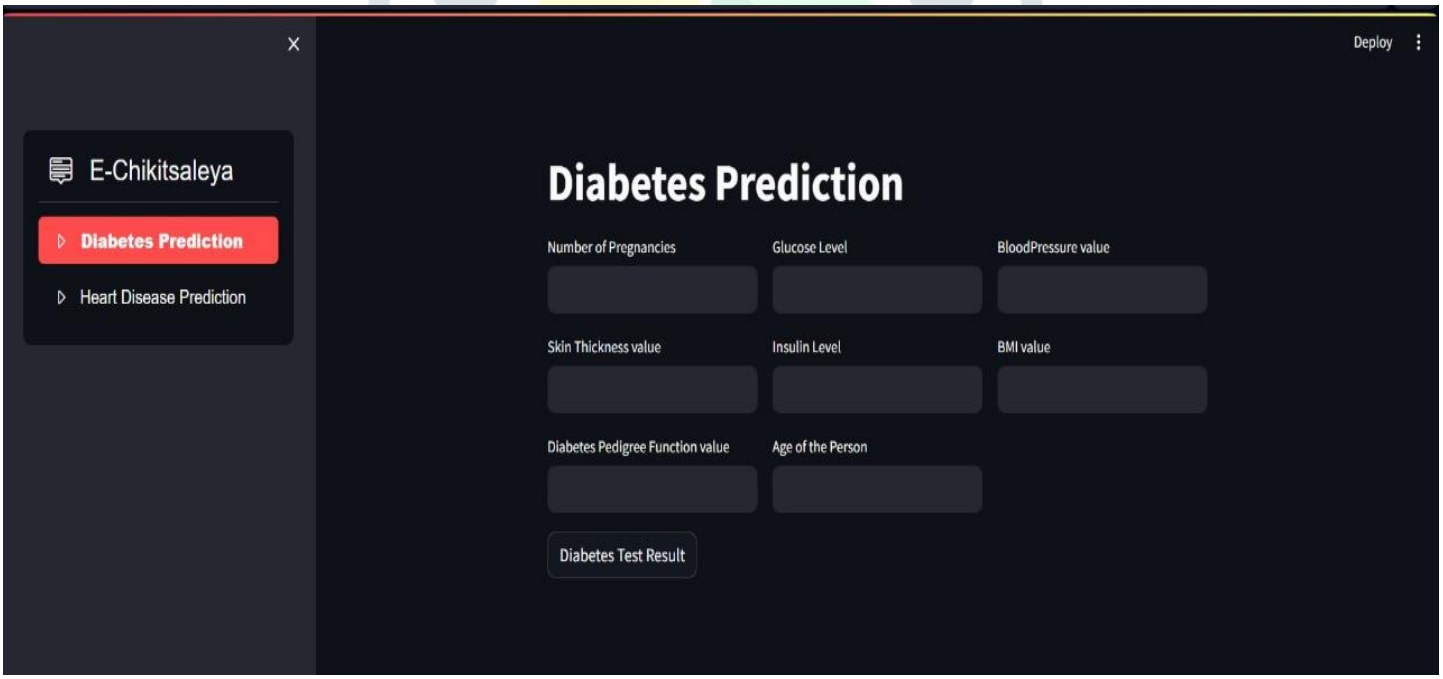

Fig.1.4.3.1 Diabetes Prediction Interface

Figure 1.4.3.1 depicts the symptoms , along with the age, bp level, glucose level and other data of individuals, were utilized as input for multiple ML algorithms in case of diabetes.

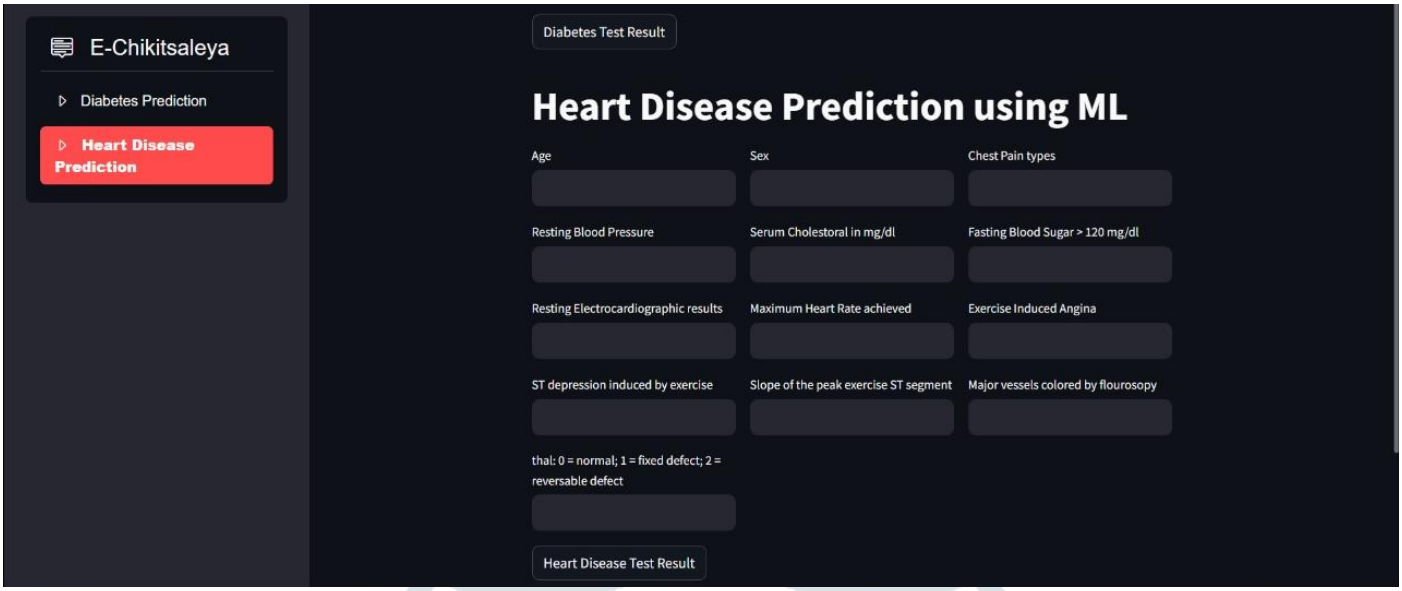

Fig.1.4.3.2 Heart Disease Prediction Interface

Figure 1.4.3.2 depicts the symptoms such as age, gender, bp level and other data of individuals were used as the input for the algorithms in case of heart disease.
For our model, Jupyter Notebook, Google Collab, and Spyder were the used software. For the aim of training and testing the model with 300 patients under various conditions, Jupyter Notebook and Google Collaboration are utilized.

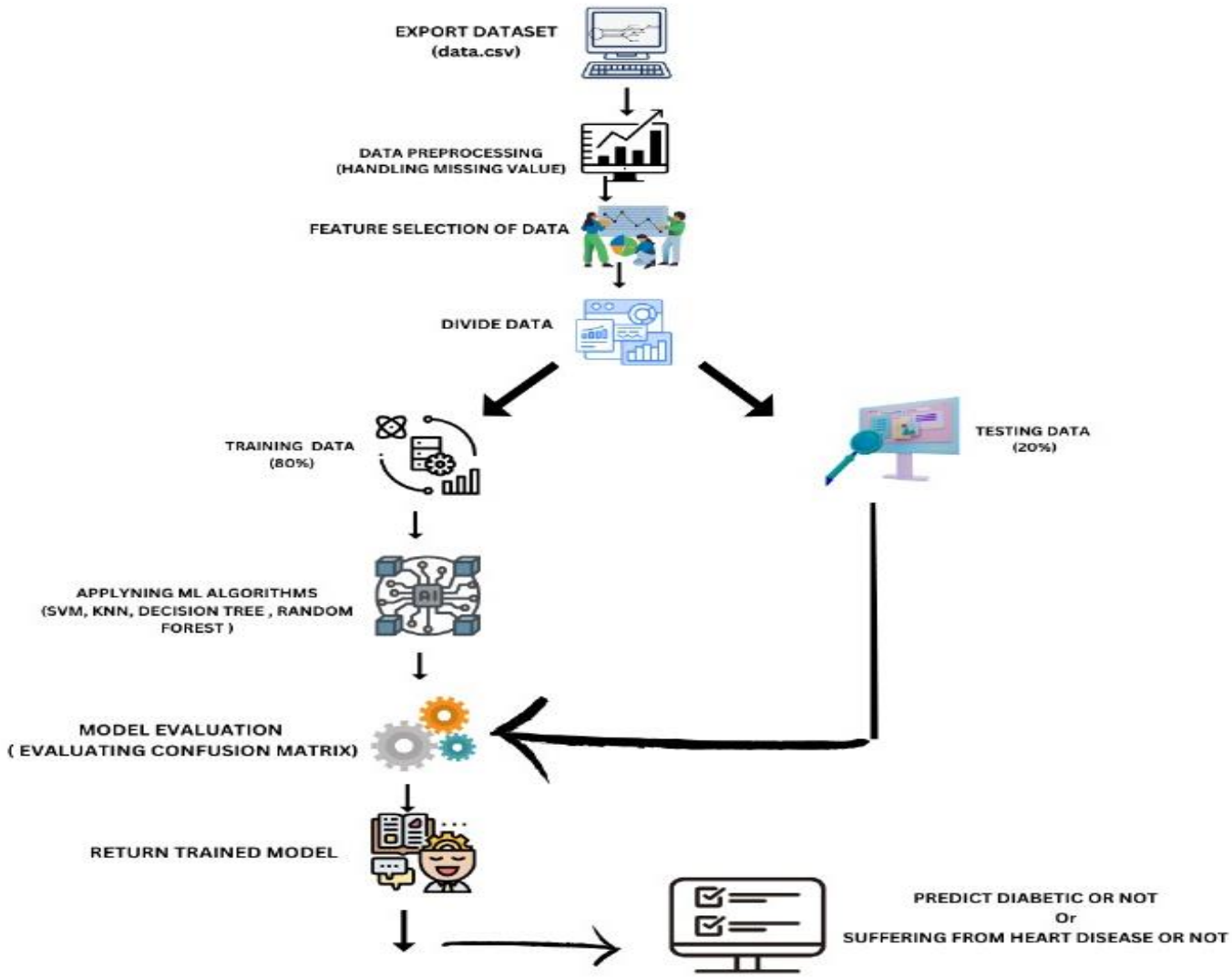

Fig.1.4.3.3 Framework Of E-chikitsaleya

Figure 1.4.3.3 illustrates that how the trained model for the web interface is deployed using Spyder.
Python Language and its many libraries are used for this specific ML model. For the Disease prediction model; NumPy, Pandas, Seaborn, SkLearn, MatPlotLib, and StreamLit are the libraries that were utilized. The pandas library[15] is an assortment of packages and modules with many different features. Developers can complete a variety of tasks without starting from scratch while using these libraries. SKLEARN is a free ML library in python . Many applications, including grade boosting, K-means, naive Bayes, model selection, regression, clustering, and preprocessing, can be successfully carried out using sklearn. With streamlit, you may quickly, rather than over the course of weeks, convert Python scripts into interactive web applications. Make report generators, chat apps, or dashboards. After creating an app, you can manage, distribute, and deploy it using our Community Cloud platform.seaborn: One of the trustworthy resources for statistical model visualization, such as heat maps, is Seaborn. This Python library works closely with Pandas data structures and is based on Matplotlib. To find out how to install this package, go to the installation page. matplotlib: While all the libraries we've seen can perform a wide range of numerical functions, Matplotlib shines in dimensional charting. For publishing good-quality figures in hard copy formats and interactive settings across platforms, sklearn is frequently used. In a few lines of code, we can create a variety of graphs and charts, including pie charts, scatterplots, histograms, error charts, and more. NumPy: Another python package that supports scientific computing, NumPy supports massive N-dimensional arrays and matrices and offers a set of quite good level for mathematical functions to quickly perform these operations. NumPy uses BLAS and LAPACK to perform computations in linear algebra efficiently. NumPy can also be applied as a productive multi-dimensional generic data container.

1.4.4 Comparison of the algorithms

After the application of the confusion matrix, algorithms, and compilation in the software , the next step of the study is to perform the comparative analaysis between these algorithms to find the optimal one that is more supported by the data. The comparison is done by the values of accuracy obtained from the confusion matrix for each algorithm in Diabetes as well as in Heart Disease.

2. RESULTS AND DISCUSSION

2.1 Results of Algorithms of Model

Table 2.1 Table for the results of Diabetes Prediction

Name of Algorithm	Accuracy	Precision	Recall
KNN	0.71	0.76	0.80
DECISION TREE	0.75	0.79	0.85
SVM	0.79	0.89	0.81
RANDOM FOREST	0.84	0.89	0.87

Table 2.2 Table for the results of Heart Disease Prediction

Name of Algorithm	Accuracy	Precision	Recall
KNN	0.61	0.65	0.48
DECISION TREE	0.79	0.75	0.80
SVM	0.86	0.75	0.91
RANDOM FOREST	0.85	0.85	0.81

Table 2.1 represents the results of different algorithms for diabetes prediction model and Table 2.2 represents the results of different algorithms for heart disease prediction model.
A assortment of ML algorithms were tried to predict disease for the input dataset. Four diverse ML algorithms were utilized for the prediction: RF, DT, SVM, and KNN. The RF algorithm was found to have the most elevated accuracy[10] of 84% for diabetes, which changed concurring to the dataset estimate (small and expansive for the training set). As a result, it was found to be the most

accurate algorithm among the other ML algorithms. The DT showed an accuracy of 75%, the SVM was accurate at 79%, and the KNN was accurate at 71%.

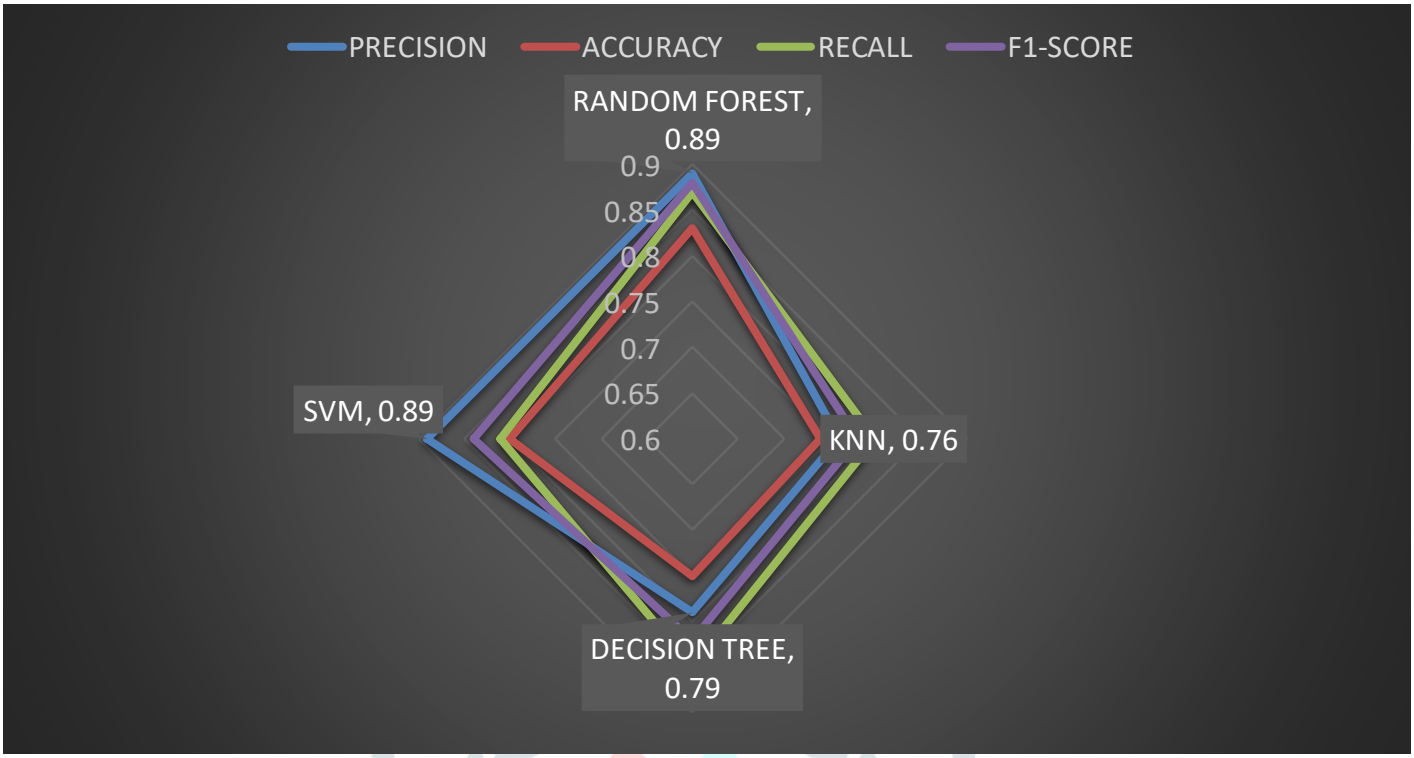

Fig.2.1 Confusion Matrix Values

Figure 2.1 depicts the values from the confusion matrix that contains the accuracy of the algorithm , recall value of the algorithm , precision of the algorithm and the f1 score of the algorithm.

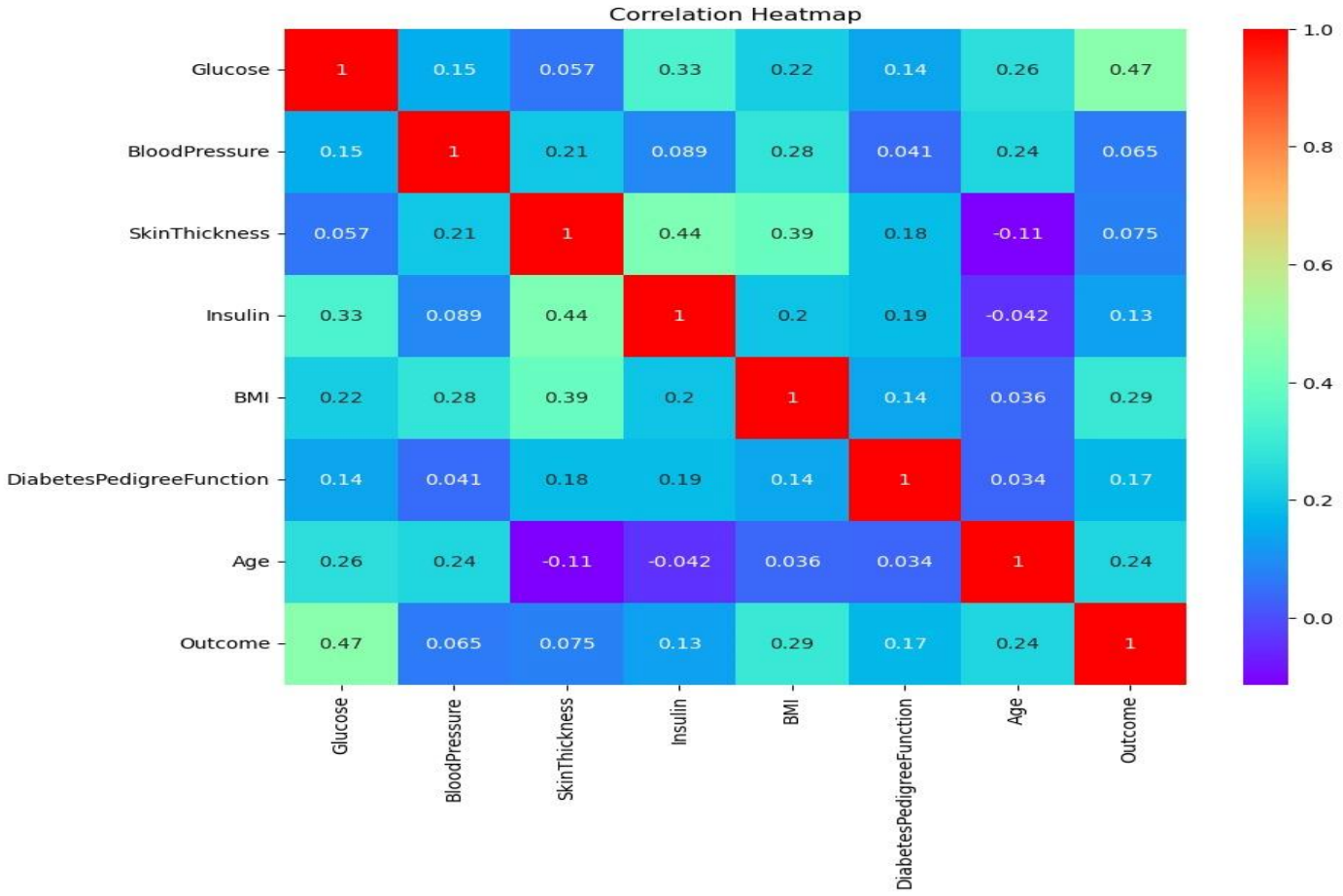

Fig.2.2 Heat Map For Diabetes

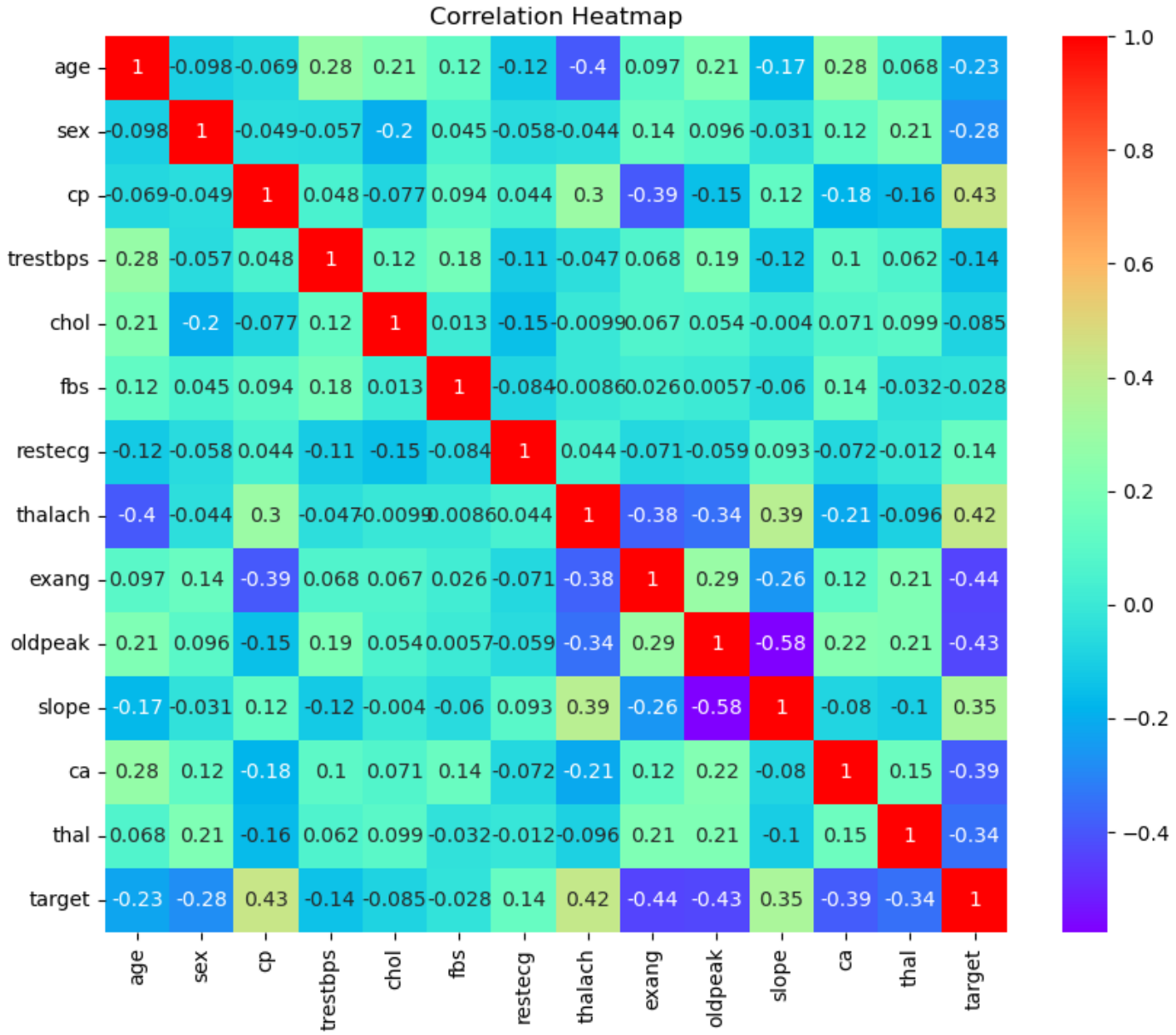

Fig.2.3 Heat Map For Heart Disease

Figure 2.2 depicts the heat map for diabetes and Figure 2.3 depicts the heat map for the Heart disease. The heat maps depicts the data visualization for Diabetes and Heart disease and represents the magnitude of different parameters within these dataset with different colours. The variation in colour is also known as hue or intensity.

The SVM algorithm was found to have the highest accuracy of 86% for Heart disease, which changes according to the dataset size. As a result, it was found to be the most accurate algorithm among the other ML algorithms. The DT showed an accuracy of 79%, the RF was accurate at 85%, and the KNN was accurate at 61%.In the realm of medical research, we have conducted a thorough analysis of various methodologies documented in the literature. Among these, the SVM and KNN models have been commonly employed for disease prediction. In our study, we have also utilized these models along with two different machine-learning algorithms to predict diabetes and heart disease. Remarkably, we achieved an impressive accuracy rate of 84% in diabetes and 86% in heart disease, surpassing the majority of previously reported methodologies. This remarkable precision can be credited to the execution of the RF Algorithm and SVM Algorithm separately.

Fig.2.4 Distribution Of Class Variable And Features For Diabetes

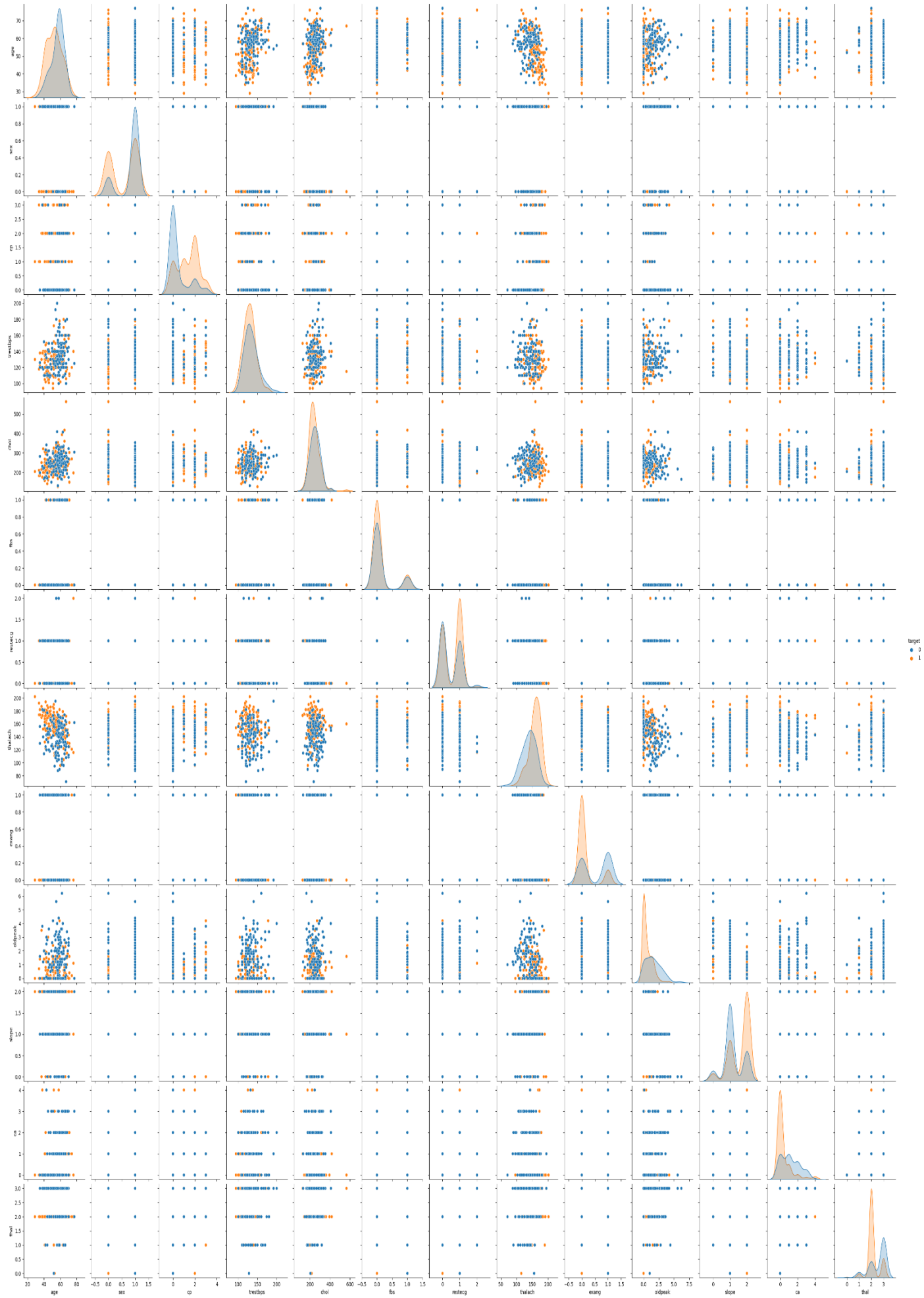

Fig.2.5 Distribution Of Class Variable And Features For Heart Disease

Figure 2.4 it represents the graph which illustrates the distribution between the class variables and features for diabetes and in Figure 2.5 it represents the graph which illustrates the distribution between the class variables and different features for heart disease.
The E-Chikitsaleya may be a portrayed method for anticipating disease based on patient's side effects, age, and gender. The RF Algorithm accomplished the most elevated accuracy, 84% for diabetes, and the SVM Algorithm accomplished the most elevated accuracy, 86% for heart disease, in predicting diseases using these variables. Most ML models tested provided good accuracy. However, some parameter-dependent models struggled, yielding lower accuracy. Accurately predicting disease could optimize medicine allocation and improve treatment outcomes while lowering costs. Overall, the RF Model and SVM Model show promise for leveraging patient data to predict disease and inform care management.
There is a possibility to enhance the existing study by delving into supplementary factors and components like medical prescription for the particular disease , the dosage will be planned according to the age , gender, sugar level ,bp level and etc of the patient. In the future, we'll collaborate with real-time hospital-validated information. In any case, due to time confinements, the scope of this research is limited, and further examination is required to dive deeper into these domains. Our plans involve implementing extra categorization strategies, diverse discretization approaches, and different vote-by-classifier strategies. We'll incorporate other illnesses into the current framework for future extension. Furthermore, we aim to create a chatbot that can address regularly inquired questions and guarantee the framework is as user-friendly as conceivable.

3. ACKNOWLEDGMENT

We extend our sincere gratitude to the individuals and organizations whose contributions have been vital to the completion of this research endeavour.
Anchal Gupta and Ankit Sah deserve special acknowledgment for their invaluable assistance in refining this research paper. Their meticulous attention to detail and insightful inputs have greatly enhanced the clarity and coherence of our findings.
We are deeply thankful to Farhan Yezdani and Chakshu Chawla for their expertise in model building, deployment, and the creation of visual aids. Their technical proficiency and creativity have significantly enriched the execution and presentation of our research.
Dr. Suman Bhatia's guidance, mentorship, and unwavering support have been indispensable throughout every stage of this project. Her expertise has played a crucial role in shaping the direction and outcomes of our study. The successful completion of this research project stands as a testament to the collaborative efforts and of all those mentioned above, for which we are truly grateful.

REFERENCES

[1] Sun Y. L., Zhang D. L. MLTechniques for Screening and Diagnosis of Diabetes: A Survey. Technical Gazette . 2019.

[2] Tafa Z., Pervetica N., Karahoda B. An Intelligent System for Diabetes Prediction. IEEE Explore ; Proceedings of the 4th Mediterranean Conference on Embedded Computing (MECO)

[3] Qawqzeh Y. K., Bajahzar A. S., Jemmali M., Otoom M. M., Thaljaoui A. Classification of diabetes using photoplethysmogram (PPG) waveform analysis: logistic regression modeling.

[4] Kandhasamy J. P., Balamurali S. Performance analysis of classifier models to predict diabetes mellitus. Procedia Computer Science .

[5] Tigga N. P., Garg S. Prediction of type 2 diabetes using MLclassification methods.

[6] Narin, A.; Isler, Y.; Ozer, M. Early prediction of Paroxysmal Atrial Fibrillation using frequency domain measures of heart rate variability. In Proceedings of the 2016 Medical Technologies National Congress (TIPTEKNO), Antalya, Turkey, 27–29 October 2016.

[7] Shah, D.; Patel, S.; Bharti, S.K. Heart Disease Prediction using MLTechniques. SN Comput. Sci. 2020, 1, 345.

[8] Alotaibi, F.S. Implementation of MLModel to Predict Heart Failure Disease. Int. J. Adv. Comput. Sci. Appl. 2019, 10, 261–268

[9] E. M. Dos Santos, R. Sabourin, and P. Maupin, Information Fusion, vol. 10, no. 2, pp. 150–162, 2009.

[10] What is a Confusion Matrix in MLby Jason Brownlee on November 18, 2016 in Code Algorithms from Scratch.

[11] Breiman, Leo, Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). Classification and regression trees. Monterey, CA: Wadsworth and Brooks/Cole Advanced Books and Software. (Google citation: 37373)

[12] Cortes, C., and Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273297.

[13] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability 1(14), 281-297.
[14] Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., ... and Zhou, Z. H. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), 1-37.s

[15] “Top 30 Python Libraries To Know" by Shveta Rajpal in My Great Learning (2024)