UTILIZING CLASSIFICATION TECHNIQUES TO FORECAST MULTIPLE DISEASES

Project submitted to the

SRM University – AP, Andhra Pradesh

for the partial fulfillment of the requirements to award the degree of

Bachelor of Technology/Master of Technology

In

Computer Science and Engineering

School of Engineering and Sciences

Submitted by

G. Harshavardhan Reddy

AP20110010588

Under the Guidance of

Dr. Saleti Sumalatha

SRM University–AP
Neerukonda, Mangalagiri, Guntur
Andhra Pradesh – 522 240
[December, 2022]

Certificate

Date: 9-Feb-23

This is to certify that the work present in this Project entitled "**UTILIZING CLASSIFICATION TECHNIQUES TO FORECAST MULTIPLE DISEASES**" has been carried out by **G.Harshavardhan Reddy** under my/our supervision. The work is genuine, original, and suitable for submission to the SRM University – AP for the award of Bachelor of Technology/Master of Technology in **School of Engineering and Sciences**.

Supervisor

(Signature)

Prof. / Dr. [Name]

Designation,

Affiliation.

Co-supervisor

(Signature)

Prof. / Dr. [Name]

Designation,

Affiliation.

Acknowledgments

We would like to express our gratitude to Dr. Saleti Sumalatha for assisting us with the research. Without the advice and assistance of our adviser, this study effort would not be feasible. We would also like to recognize Dr. Neeraj Kumar Sharma for providing the initial step for this project by providing a dataset and outlining the route to begin our research project. The suggestions, feedbacks, and encouragement from our advisors were extremely helpful in assisting us in finishing this research project.

Table of Contents

Abstract

A predictive work for disease prediction is proposed in this work in the healthcare industry using machine learning techniques. Machine learning is now widely employed in a variety of industries, ranging from finance to healthcare. This paper post will look at the role of machine learning in illness prediction. The predictions of diseases based on symptoms where the dataset included 41 diseases were done with different algorithms and the dataset was pre-processed by removing the unwanted duplicate values where the most of rows were dropped and the feature selection was done based on highly correlated features and trained and tested data was applied to algorithms named RF, DT, GB and NB, and the results were acquired on the predictions are 86.66%, 81.52%, 74.46%, 83.69%.

Abbreviations

RF	Random Forest
DT	Decision Tree
GB	Gradient Boosting
NB	Naïve Bayes
LGR	Logistic Regression
AB	Ada boost
SVM	Support vector machine
$\mathrm{SVM_L}$	Support vector machine(Linear)
$\mathrm{SVM_P}$	Support vector machine(Polynomial)
$\mathrm{SVM_{IRB}}$	Support vector machine(Improved Radial bias)
CNN	Convolutional Neural Network
KNN	K-Nearest Neighbour
FT	Fine Tree
MT	Median Tree
CT	Course Tree
FKNN	Fine K-Nearest Neighbour
MKNN	Medium K-Nearest Neighbour
CKNN	Course K-Nearest Neighbour
WKNN	Weighted K-Nearest Neighbour
GNB	Gaussian Naive Bayes
KNB	Kernel Naive Bayes
SKNN	Subspace K-Nearest Neighbour
RUSBT	Random Under Sampling Boosted Tree
TP	Ture positive values
FP	False positive values
TN	True negative values
FN	False negative values

List of Tables

List of Figures

1. Introduction

Humans nowadays confront a variety of illnesses as a result of existing environmental conditions and lifestyle choices. It is critical to detect and forecast such diseases in their early stages in order to prevent them from progressing to their terminal phases. In today's environment, there is a lot of focus on the need of early illness detection. We can now diagnose illnesses sooner than ever before because to advances in technology. Early illness detection is critical because it allows therapy to begin sooner. Early identification also allows for better disease treatment and can prevent the disease from advancing to a more severe state. There are several methods for detecting illnesses early on. Screenings, testing, and physical examinations are examples of these. Early detection is critical for illness treatment effectiveness. It has the potential to save lives and enhance the quality of life for individuals suffering from the condition.

Early illness discovery can lead to early treatment and better outcomes. For example, if a woman diagnoses with breast cancer early on, she has a better chance of survival than if she waits until cancer has developed. Early detection may also result in less intrusive and less costly therapies. When certain diseases are diagnosed early, they are easier to cure. Cancers that are diagnosed early, for example, are frequently far easier to treat successfully than those that are not detected until they have advanced. Early detection can make or break the result of your therapy. Many illnesses, including cancer, can be treated if detected early. If the condition progresses, it becomes considerably more difficult to cure and may possibly be deadly. When other disorders, such as heart disease and stroke, are diagnosed early, they can be treated more successfully. Many ailments can be cured if they are detected early. Unfortunately, many people are unaware of the need of early illness identification and wait until it is too late. When they eventually see a doctor, the condition has advanced too far and is considerably more difficult to treat.

The development of analytical models is automated using the data analysis method known as machine learning. It is a branch of artificial intelligence that is based on the idea that computers can see patterns in data, learn from it, and make decisions with little or no human involvement. Machine learning is now widely employed in a variety of industries, ranging from finance to healthcare. Machine learning is used in the medical profession to improve the accuracy of diagnosis, identify which therapies work best for certain situations, and potentially uncover novel cures for diseases. This blog post will look at the role of machine learning in illness prediction. Machine learning is a strong technique for illness prediction. We can use machine learning to create models that can take data about a patient and forecast the chance of acquiring a disease. This is a critical tool for identifying at-risk individuals and offering tailored treatment. Machine learning has several advantages in illness prediction. For starters, machine learning can process vast volumes of data more efficiently than traditional approaches. Second, machine learning can detect patterns that people may find difficult to recognize. Third, when new data becomes available, machine learning models may be adjusted, making them more accurate over time. Finally, machine learning gives a more objective method of prediction.

The datasets used in this paper to develop the predictive models are freely available and may be downloaded from Kaggle machine learning datasets. A general framework for disease prediction is proposed in this work in the healthcare industry using machine learning techniques. In this dataset the symptoms are listed as different attributes with each disease that was labeled in it, this dataset contains 132 predictive factors, and 42 illnesses may be classified using these predictive variables. The predictive variables are binary variables, where 1 denotes a positive symptom and 0 denotes a symptom that is absent or negative. The data is cleaned before being imported in CSV format. and the data pre-processing and feature selection played a major role while predicting the accuracies. machine learning algorithms such as Decision Trees, Random Forest, Naïve Bayes, and Gradient Boosting are used for disease prediction, and their accuracy is compared to select the best model for that disease dataset. we also compared confusion matrix, accuracy, precision, and recall between different classification algorithms.

2. Review of literature

[1]They presented a framework for a decision support system that includes machine learning methods for illness prediction in this study. The decision support system for illness prediction was developed using an improved SVM-Radial bias kernel approach, and its efficiency was compared to that of existing machine learning techniques. The best method for predicting illnesses from patient data using symptoms is to use enhanced SVM radial bias. The datasets being gathered from UCI include those for chronic kidney disease, diabetes, and heart disease. During the pre-processing step of the data, missing values were calculated from previous records and filled up using mean values. They removed the unwanted characteristics from this prediction during feature selection and predictive model development. The subset will be created through feature extraction from the selected features. The feature extraction approach entails reducing the original features of the data to a more manageable feature set. The accuracy of the forecast made by the decision support system depends on how trustworthy the data and prediction algorithm are. The three illness datasets included in this suggested model are those for chronic kidney disease, diabetes, and heart disease. Several machines learning techniques, including SVM-Linear, Random Forest, Decision Tree, and SVM-Polynomial, are used to predict the accuracy of various illnesses. In line with predictions, the revised

model outperformed the other machine learning algorithms for chronic kidney disease, diabetes, and heart disease with accuracy rates of 98.3%, 98.7%, and 89.9%.

[2] The risk that patients will die from common diseases including breast cancer, diabetes, coronary artery disease, and malignancies can be limited and minimised by early detection. By utilising various classifiers and grouping algorithms, developments in machine learning and artificial intelligence make this possible. Three distinct illness databases—heart, breast cancer, and diabetes—all of which are accessible in the UCI repository for disease prediction—were subjected to several categorization methods in this work, each with its own advantages. They used the mean value of a continuous variable or the mode value of a categorical variable to fill in any missing values during the pre-processing of the data. In addition to using the backward selection approach, they employed feature selection to eliminate any duplicate characteristics that are associated to one another closely. They start with every attribute in the model and then filter them out depending on p-value. The model was reconstructed using the variables left behind after the characteristics with p-values higher than 0.05 were eliminated. To forecast illnesses like diabetes, breast cancer, and heart failure, they used a variety of algorithms. Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and Adaptive Boosting were the classification techniques employed. The suggested method of their model has a prediction accuracy of 87.1% for heart disease detection using LG Regression, 98.57% for breast cancer detection using the AB classifier, and 85.71% for diabetes prediction using the SVM(linear).

[3] Many machine learning prediction and other classification algorithms have benefited from the latest data mining techniques that have emerged in the new era. These algorithms are employed in the illness prediction process. They employed a variety of categorization and prediction methodologies on a wide range of symptoms, including back pain, chest discomfort, constipation, and many more. The model for this project begins with pre-processing the input data, which is presented as symptoms. Then, a series of algorithms, including Choice Tree, Random Forest, and Naive Bayes, are used to produce the decision. The algorithm was trained using the medical records of 4920 individuals who were at risk for 41 illnesses as a result of a confluence of different symptoms. To avoid overfitting, 95 out of 132 symptoms have been taken into account.

[4] Data mining is the process of employing a variety of techniques to uncover hidden patterns in vast volumes of data. Along with dealing with big data sets, heterogeneity, data privacy, and data accuracy, it is crucial. The most important area of research is medical data mining, and major efforts have been made in this area recently since inaccurate medical data systems can result in medical treatments that are significantly false. Medical data sets have been examined using a wide range of mining techniques. The prediction of illnesses has been used with data mining approaches. They used classification algorithms such as Naive Bayes, J48, REF Tree, Sequential Minimal Optimization, and Multi-Layer Perceptron on various data sets,

including as they have varying values for numerous metrics such as successfully categorized instances, precision, recall, and F-Measure, as well as time. They also presented general disease predictions based on patient symptoms. Numerous data mining algorithms may accurately forecast a wide range of illnesses, including diabetes, heart, liver, and kidney problems. SVM and Naive Bayes are the most popular and extensively utilized algorithms for illness prediction, according to the literature that is currently accessible. In terms of accuracy, both algorithms perform better than other algorithms.

[5] Many datasets, including those related to heart disease, diabetes, and breast cancer, have been subjected to data mining techniques. In the healthcare industry, data mining is essential for spotting hidden patterns and forecasting disease. In order to have clear data on which to run various data mining algorithms, predicting illness entails gathering several tests from a patient. This shorter test is important in terms of both execution and execution time. With an emphasis on predictions for datasets related to heart disease, diabetes, and breast cancer, this study investigates mining techniques used to anticipate various sorts of illnesses. The accuracy of three data mining approaches is examined in this essay. High precision, recall, and accuracy metrics are the desired results. . The confusion matrix can be used to produce these measurements, which can then be readily transformed into true positive and false positive measures.

[6] By identifying patterns and connections in the data, the author of this study hopes to use machine learning algorithms to forecast different illnesses. A single method will not suffice because the amount of medical-related data is so large, and the author has to locate all of those items. SVM, Naïve-Bayes classification, and decision trees are the methods employed. By specifically choosing these methods for mentioning their own aims, the author got right to the point. In contrast to several existing models that can only forecast one illness, the author seeks to create a single model that can predict numerous diseases. For instance, there are several models for forecasting diabetes but none for predicting heart disease linked to diabetes, or the opposite is true. They therefore initiated this action. Among the algorithms described, the nave Bayes method is used to uncover relationships between characteristics so that we can comprehend a collection of symptoms that will be connected to a select number of diseases. It analyses the conditions provided to determine the likelihood that the symptoms would manifest. We can rely on this powerful categorization method. One of the effective techniques for locating patterns in the provided large data is the widely used Support Vector Machine algorithm. Any form of data may be scaled and generalised by it. It employs the Radial Kernel Function (RDF), which among the kernel functions used by SVM is the best. For the purpose of locating the missing value in datasets, the decision tree approach is utilised. These are furthermore employed in choosing between two or more datasets. The aforementioned methods or techniques were applied to and pre-processed on a Cleveland dataset by the author. They discovered that the decision tree algorithm, which provides the best accuracy of any algorithm, is the most effective algorithm.

Finally, a model that can forecast the heart-related diabetes condition they are looking for was created.

[7] There are several models that forecast specific diseases, but there is no model for multi-disease prediction, according to the author of this study. The goal of the author's data analysis is to identify acquired diseases such as diabetes, diabetic retinopathy, heart disease, and breast cancer. TensorFlow and Flask API are the algorithms employed. These are the methods or algorithms applied in this prediction model. The primary objective of the author is to reduce mortality, which is mostly brought on by a few models that are unable to correctly diagnose the condition. The author said that some people may have many diseases, but because there aren't any models that can forecast multiple diseases, the other diseases can't be anticipated, which is a loss. The author intends to employ the Python pickling approach, which aids in serializing the data, to get around these kinds of problems. By doing this, he is able to leverage the flask API approach, which is used to determine how the model responds when used with pickled illness data. Given that it can aid in the prognosis of several diseases, it will be an effective tool. The symptoms of numerous patients from various nations make up the dataset that was used. For other illnesses, the author has used a variety of other machine learning techniques, such as naive Bayes, decision trees, and random forests. SVM achieved the greatest accuracy of 96% for cancer identification among these methods.

[8] Based on predictive modelling, a disease prediction system determines the user's condition from the symptoms they submit as input to the system. The technology evaluates the user's symptoms as input and returns a likelihood of the disease in the form of an Asian output. Naive Bayes Classifier implementation is used for disease prediction. The Naive Bayes Classifier determines the likelihood of the illness. Accurate analysis of medical data helps with early illness identification and patient care as big data usage increases in the biomedical and healthcare sectors. We are forecasting illnesses like Diabetes, Malaria, Jaundice, Dengue, and Tuberculosis using linear regression and decision trees.

[9] Due to the environment and their lifestyles, humans are susceptible to a wide range of illnesses. Predicting sickness early on hence becomes a crucial task. But for medical professionals, accurate prediction based on indications and symptoms becomes too challenging. Predicting illnesses accurately is the hardest task. They suggested using the patient's symptoms to forecast the illness. They used the K-Nearest Neighbour and Convolutional neural network techniques to forecast diseases. On three illness datasets, including those for diabetes, cerebral infarction,

and heart disease, they conducted disease prediction. They foretell whether a patient would suffer from a cerebral infarction with a high risk of occurrence or one with a low risk. They first acquired a disease dataset from UCI machine learning, which is composed of a list of diseases and their symptoms. After that, the dataset is pre-processed to get rid of commas, punctuation, and white spaces. The training dataset is then created using this. The feature was then retrieved and selected after that. The data is then categorized using methods for classification like KNN and CNN. Using machine learning, we can predict diseases with high accuracy. They applied a CNN-based multimodal illness risk prediction to test data. CNN-based multimodal disease risk prediction algorithms and CNN-based unimodal sickness risk prediction algorithms were compared. The ability to forecast diseases is 94.8% accurate.

[10] In this case study, the author aims to use machine learning techniques to speed up the medical diagnosis system in order to provide accurate and timely findings. Therefore, he employed several approaches from both data mining and machine learning in order to accomplish his aim. This suggested system employs KNN, Naive Bayes Classification, and Decision Tree algorithms. The author notes that while a doctor's results may occasionally have errors, the model he is going to suggest shouldn't. Therefore, he employed three different types of algorithms and the different types of those algorithms to accomplish his purpose. He can use this to discover an algorithm that is accurate and effective. The data set was subsequently analyzed using a variety of machine learning (ML) models, including fine, medium, and coarse decision trees, gaussian and kernel naive Bayes, fine, medium, and coarse KNN, weighted and subspace naive Bayes, and RUS boosted trees. The ensuing illness and the precision of the employed procedure are the two sorts of outputs provided by this model. Out of all of them, the weighted KNN produces the best results. The weakest performance came from RUS Boosted. Of all the KNN methods, the Fine KNN is the best. Of the 11 models used, the Weighted KNN had the greatest accuracy (93.5). The condition and its specific medical information will be forecasted after determining which algorithm is the best or most effective. The author has taken a chance by using 11 different types of approaches for the prediction because the data is in raw form. Additionally, because the dataset is so large, 230 disorders are included.

[11] This project's primary objective is to advance computer-aided techniques. Using machine learning, they are able to anticipate a number of ailments, including diabetes and heart disease. Additionally, they are attempting to reduce human

participation by using AI. All of the machine learning methods, including supervised learning, unsupervised learning, semi-supervised learning, naive bayes classification, SVM, decision trees, and random forests, are discussed. To determine the sort of sickness the patient could contract, they have employed many algorithms from various types of learning. The author has chosen to use as many of the approaches in this model as possible to treat the diseases that are now in the public eye and widespread. The author employed machine learning techniques such the SVM, decision tree, random forest, and naive bayes algorithm. Each method is used to determine the prevalence of the condition in question. The naive bayes algorithm is used to categorize high dimensional data using supervised learning. For group modelling, which may be done on categorized data, the random forest is employed. It is necessary to take this action. Making judgments using data that has been reorganized and sorted is made easier with the aid of the decision tree. Regarding various sorts of data, several decision trees are created. The random forest incorporates these decision trees, which clarifies a number of issues. SVM is one of the most effective supervised learning algorithms for classification and regression problems. These methods have been applied by the author to treat conditions including diabetes, heart disease, liver disease, and dengue fever. When the accuracy of algorithms and illnesses are compared, the decision tree method produces the best results for dengue sickness. Other techniques have a better degree of precision. The author's primary objective of developing a model for illness prediction utilizing a variety of ML approaches has been accomplished.

[12] Medical databases are appealing to researchers from all around the world. Decision support systems for sickness prediction have been built using a range of medical datasets and data mining approaches. In this paper, we propose a novel knowledge-based method for sickness prediction using clustering, noise reduction, and prediction algorithms. Our knowledge-based approach uses Classification and Regression Trees to generate fuzzy rules (CART). On a number of available medical datasets, we assess our recommended approach. Results employing datasets from the Pima Indians for diabetes, mesothelioma, WDBC, Stat Log, Cleveland, and Parkinson's telemonitoring show a significant increase in the proposed method's prognostic capability. The results showed that clustering algorithms can be successful in predicting illnesses when combined with fuzzy rule-based, CART with noise reduction, and real-world medical datasets. The knowledge-based system can aid medical practitioners in the practice of healthcare as a clinical analytical method.

	Methodology	Datasets	Results
[1]	Using improved SVM-radial Bias and comparing with existing algorithms.	Chronic Kidney Disease dataset Diabetes dataset, and heart disease dataset form UCI.	RF – 97.8,79.9,82.0 DT - 66.3,97.4,73.0 SVM$_L$ - 96.7,77.6,84.3 SVM$_P$ - 96.7,77.6,86.5 SVM$_{IRB}$- 98.3,98.7,89.9
[2]	Predicting using the different classification algorithms like AB, RF, DT etc…	Wisconsin Breast Cancer dataset, Pima Indians Diabetes dataset, Heart disease dataset	AB – 98.57, 80.52, 83.87 DT – 94.29, 74.03, 70.97 LGR – 95.71, 84.42, 87.10 RF –97.14, 81.82, 77.42 SVM$_{RB}$ -95.71%,66.25%,54.84% SVM$_L$ - 96.7,77.6,84.3
[3]	Predicting the diseases using DT, RF, NB	Multi disease dataset.	DT -93% RF – 93% NB – 93%
[4]	Comparing the results from different prediction models from different papers based of diseases	Heart disease dataset, Diabetes dataset, Liver dataset, Kidney dataset From UCI	-
[5]	Predicting different diseases with ML algorithms	Heart disease dataset, Diabetes dataset, Breast cancer dataset	DT – 77%, 100%, 75.52% NB – 79%, 77.6%, 82.5%
[6]	Classification algorithms are used to predict the diseases.	Heart disease dataset, from UCI	NB – 77% SVM – 86.5% DT – 90%
[7]	Multiple Diseases Forecasting Model Using Flask API and ML Algorithms	diabetes dataset diabetes retinopathy datasets, heart disease and breast cancer datasets	LGR-92%(diabetes) RF – 95%(heart) SVM – 96%(cancer) And GUI interface is built with this models
[8]	Based on predictive modelling, a disease prediction system forecasts the user's condition	Real life hospital dataset	Prediction utilizing the real-world hospital dataset using LGR, NB, and KNN was suggested.

	based on their symptoms.		
[9]	Using KNN and CNN's various classification algorithms, the prediction was done	Patient disease dataset, from UCI	CNN – 96% KNN – 92%
[10]	Using machine learning, diagnose diseases based on various symptoms	Symptoms dataset, from UCI	RUSBT – 0.5% SKNN – 73.5% KNB – 16.8% GNB – 1685 WKNN – 93.5 CKNN – 5.3% MKNN – 61.8% FT -21.8% MT – 12.3% CT – 6.4%
[11]	Diseases Prediction from Symptoms Using Machine Learning Techniques	Heart Disease dataset, Diabetes Disease dataset, Liver Disease dataset, Dengue Disease dataset From open sources ,UCI	-
[12]	A Machine Learning-Based Analytical Method for Disease Prediction	Wisconsin Diagnostic Breast Cancer dataset, StatLog Heart Disease dataset, Cleveland Heart Disease dataset, Mesothelioma dataset Pime Indian Diabetes dataset	PCA-KNN – 81.2%, 75.3%, 79.2%, 82.3%, 81.8 % PCA-SVM – 89.8%, 88.2%, 84.9%, 86.1%, 85.6% EM-PCA-Fuzzy Rule-Based – 93.2%, 91.4%, 92.8%, 93.6%, 92.9%

Table 1. Summarized report for the papers we studied

3. Methodology

With each disease that was labelled in this multi-disease data set, the symptoms are listed as a different attribute. Data pre-processing and feature selection played a major role in predicting the accuracy of the predictions. The algorithms used were Random Forest, Decision tree, Gradient boosting, and Naive Bayes.

3.1 Data collection

In this multi-disease data set the symptoms are listed as different attributes with each disease that was labeled in it, and the data pre-processing and feature selection played a major role while predicting the accuracies. This dataset contains 132 predictive factors, and 42 illnesses may be classified using these predictive variables. The predictive variables are binary variables, where 1 denotes a positive symptom and 0 denotes a symptom that is absent or negative.

3.2 Data pre-processing

The goal of the case study is to develop an accurate and precise machine-learning model to recognize these disorders. We can observe that 131 of the 133 variables in the training set contain binary categorical (0 or 1) values. We removed the column "Unnamed: 133" for further investigation because it appeared to be null. Due to the dataset's large number of duplicate tuples, we also need to remove the unnamed column. After pre-processing the data, we get 4920 data entries (in a row), 132 variables (in columns), and 1 target variable in a string format that has been reduced to 304 tuples with 132 symptoms. The target variable has remained unchanged. Now that there are no null values in the data, we can go on to create the ML model.

3.3 Feature selection

In feature engineering, features will be chosen based on their relative value to the selected model. By better representing the data with a subset of the original set of features, this enables the machine learning method to train more quickly while also lowering the computational complexity and cost of the model. When the appropriate subset is chosen, the model may also become simpler to understand for humans, more accurate in some circumstances, and easier to interpret. Consequently, we must examine and deal with linked features. We also created a heatmap to identify connected characteristics.

3.4 Model description

The two datasets show how the training set and test set are divided. They do not have the same attribute values even though they have the same characteristics. The training set is used to construct and train the classification models. The test set is used to forecast the classifications of fresh, unbiased data that wasn't used to train the model before evaluating the model's performance using the performance metrics of accuracy, precision, recall, and F1-score of those classifications.

3.5 Decision Tree

A decision tree may be used in decision assessment to show decisions and decision-making quickly and aesthetically. It employs decision-making models with tops that resemble trees, as suggested by the name. The decision tree algorithm is a member of the supervised learning algorithm family. Its function is regression. It has a root node at the beginning, splits in the dominant input feature, and then splits again. When all inputs have been deposited, these processes are repeated, and the input is then categorized using the weights stored in the very last node, which contains the weights.

3.6 Naïve Bayes

A supervised learning method that handles categorization issues is the Naive Bayes algorithm. The Bayes theorem forms the basis of it. One of the simplest and most powerful classification algorithms, the Nave Bayes Classifier, assists in the creation of rapid ML models that may produce prompt predictions. The most common use for it is text classification, which necessitates a bidimensional training data set. The Bernoulli model is perfect for use in Naive Bayes classification since the dataset is binary and we know that the probability generation will be highest when the values are between 0 and 1.

3.7 Random Forest

Random Forest is a well-known machine learning method that is a component of the supervised learning system. It may be used to address problems with regression and differentiation in ML. It focuses on the idea of group modeling, a technique for combining many classifiers to address a specific problem and increase the model's effectiveness. It boosts the supplied dataset's projected accuracy on various subsets using a variety of decision trees and averages. The RandomForest predicts the results using the prediction from each decision tree and is based on the maximum voting of forecasts instead of relying on a single decision tree.

3.8 Gradient Boosting

A prediction method comprised of numerous wobbly estimating strategies, the most popular of which is the decision tree, is produced using the machine learning technique known as gradient boosting for regression and classification. By fusing many smaller models, it produces a robust, huge model with outstanding prediction performance. Because they can efficiently categorise datasets, these models are commonly employed. Gradient-boosting classifier models are often constructed using decision trees. Reduce the loss function, teach a weak learner to generate accurate predictions, and adapt weak learners to an optimization approach are the three main traits of gradient boosting.

3.9 Analysis of Prediction :

In machine learning, where illness prediction is used, there are numerous analysis prediction techniques for bringing the prediction findings to a conclusion.

3.9.1 Confusion Matrix

Four performance assessment measures are utilized to assess the proposed illness prediction model. The true positives (TP) are accurate predictions of diseases, the true negatives (TN), are accurate predictions of people without diseases, the false positives (FP), are inaccurate predictions of healthy people as diseased people, and the false negatives (FN), which are inaccurate predictions of the target as healthy people, make up the confusion matrix. The whole positive tuples are mentioned as (P) where P = TP+FN and the whole negative tuples are named as (N) where N = TN+FP, and P′ = FP+TP AND N′ = FN+TN.

3.9.2 Accuracy

The classification accuracy is expressed mathematically as follows, where the following is the ratio of the accurately predicted values to the total predicted values.

Accuracy $= {}^{\mathrm{TP}\,+\,\mathrm{FP}}/_{P\,+\,N}$

3.9.3 Precision

The precision, also known as the is mathematically represented as follows. It is defined as the proportion of right forecasts to all correct values, including both true and erroneous predictions

Precision = ${}^{\mathrm{TP}}/_{\mathrm{P}'}$

3.9.4 Recall

The precision, also known as the Ture positive rate, is mathematically represented as follows. It is defined as the proportion of right forecasts to all correct values, including both true and erroneous predictions

Recall = ${}^{\mathrm{TP}}/_{\mathrm{P}}$

3.9.5 F-score

The weighted average of the values obtained from the computation of the accuracy and recall parameters is referred to as the F-measure (F). When there is an uneven distribution of students in a class, the F_1 score is far more significant than the accuracy score. And anytime false positives and false negatives have different values, the F_1 Score value is quite appropriate. The mathematical representation of the F_1 score is as follows

$$F_1 = {}^{2 \times \text{precision} \times \text{recall}}/_{\text{precision} + \text{recall}}$$

Algorithms	Recall	Precision	F1 Score
Random Forest	89	89	87
Decision Tree	76	89	79
Naïve Bayes	83	81	81
Gradient Boosting	81	95	85

Table 2.results of recall, precision, F1Score

4. Discussion

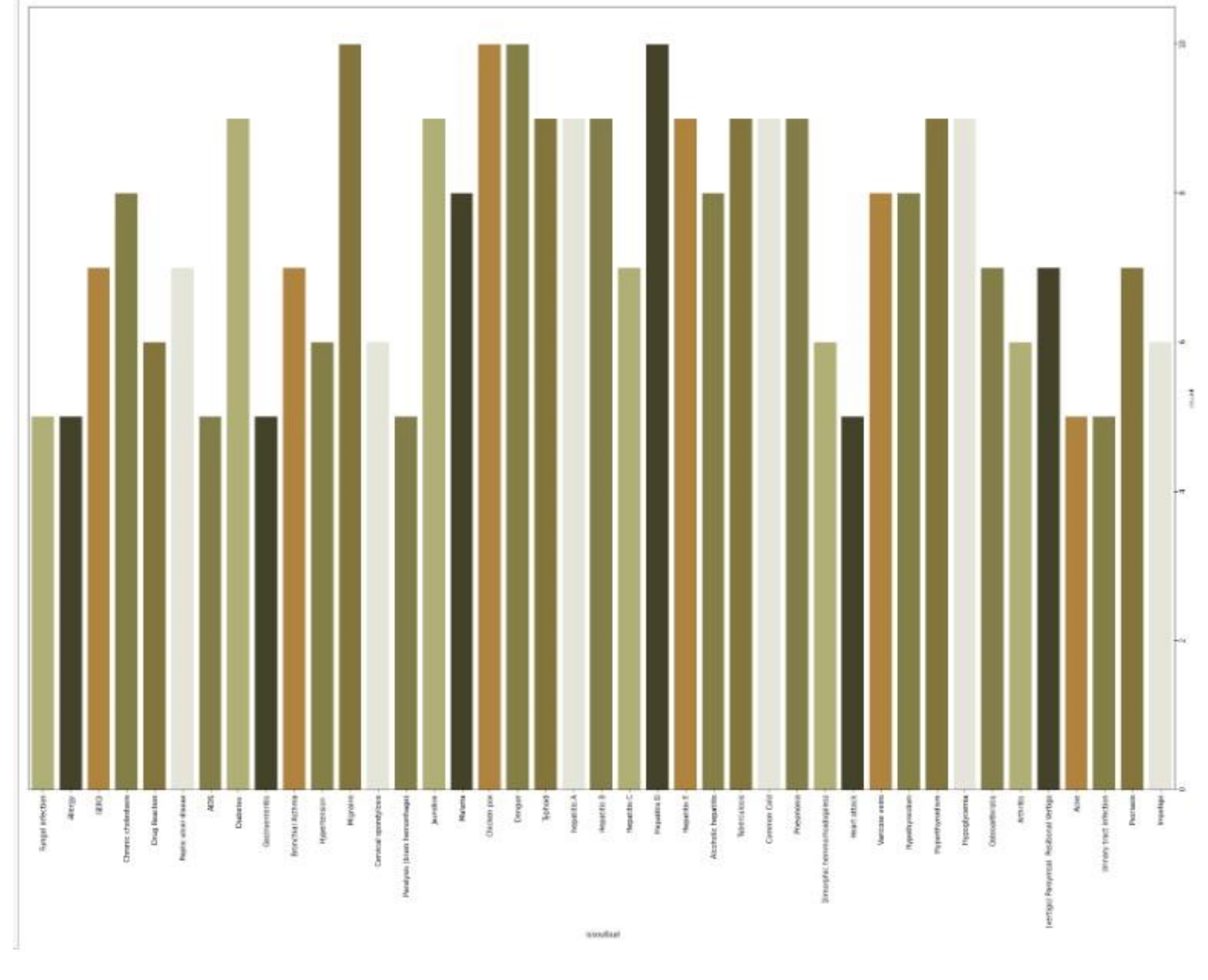


Figure 1. Count of symptoms

A visual depiction of the number of symptoms that are produced when all the duplicate tuples from the dataset are removed during data preprocessing.

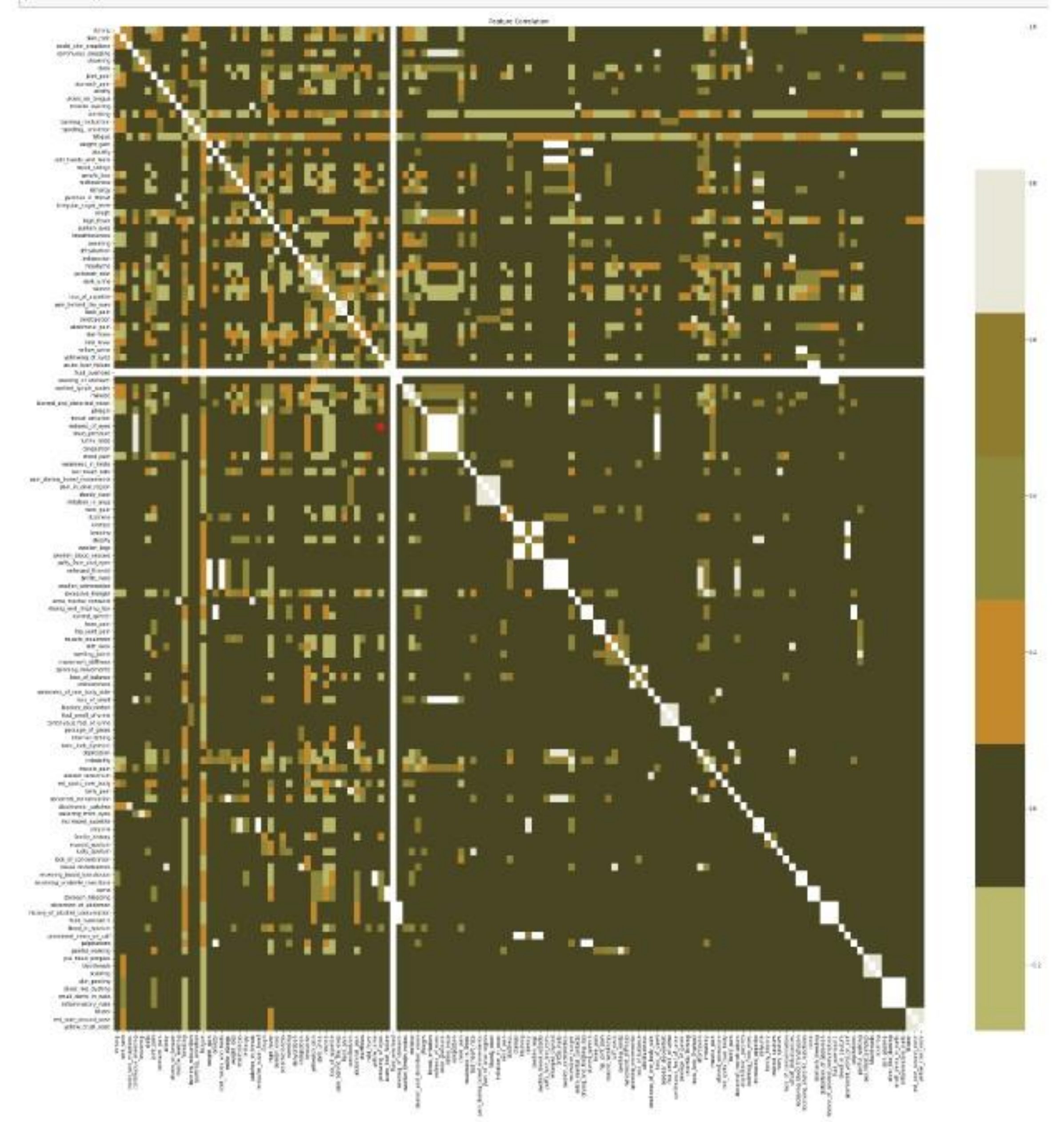


Figure 2. Correlation heatmap

This correlation heatmap shows the relationship between the numerical symptoms factors and the symptoms themselves where this takes place in feature selection part to remove the highly corelated symptoms.

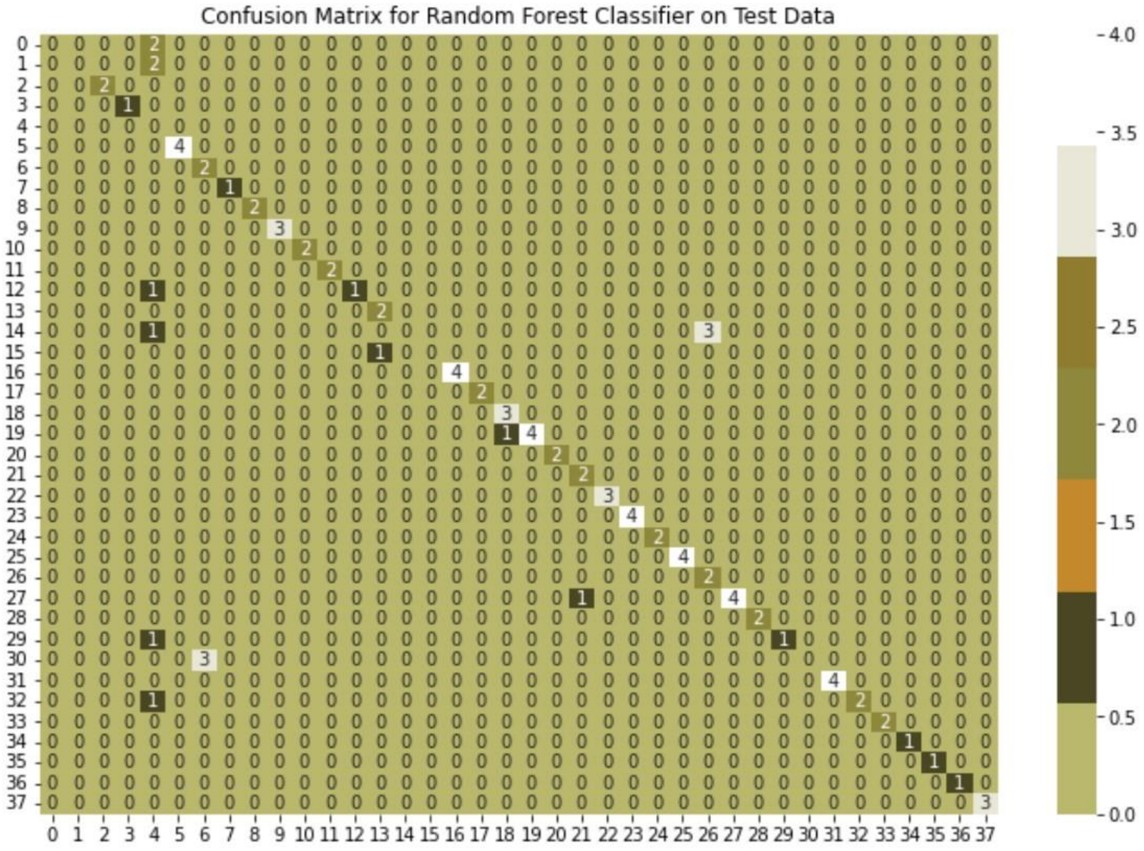


Figure 3. Confusion matrix of Random Forest

The values of the true-false predictions in this confusion matrix are obtained by fitting the train and test data to the RF model. The confusion matrix is produced from the TP,FP,TN, and FN.

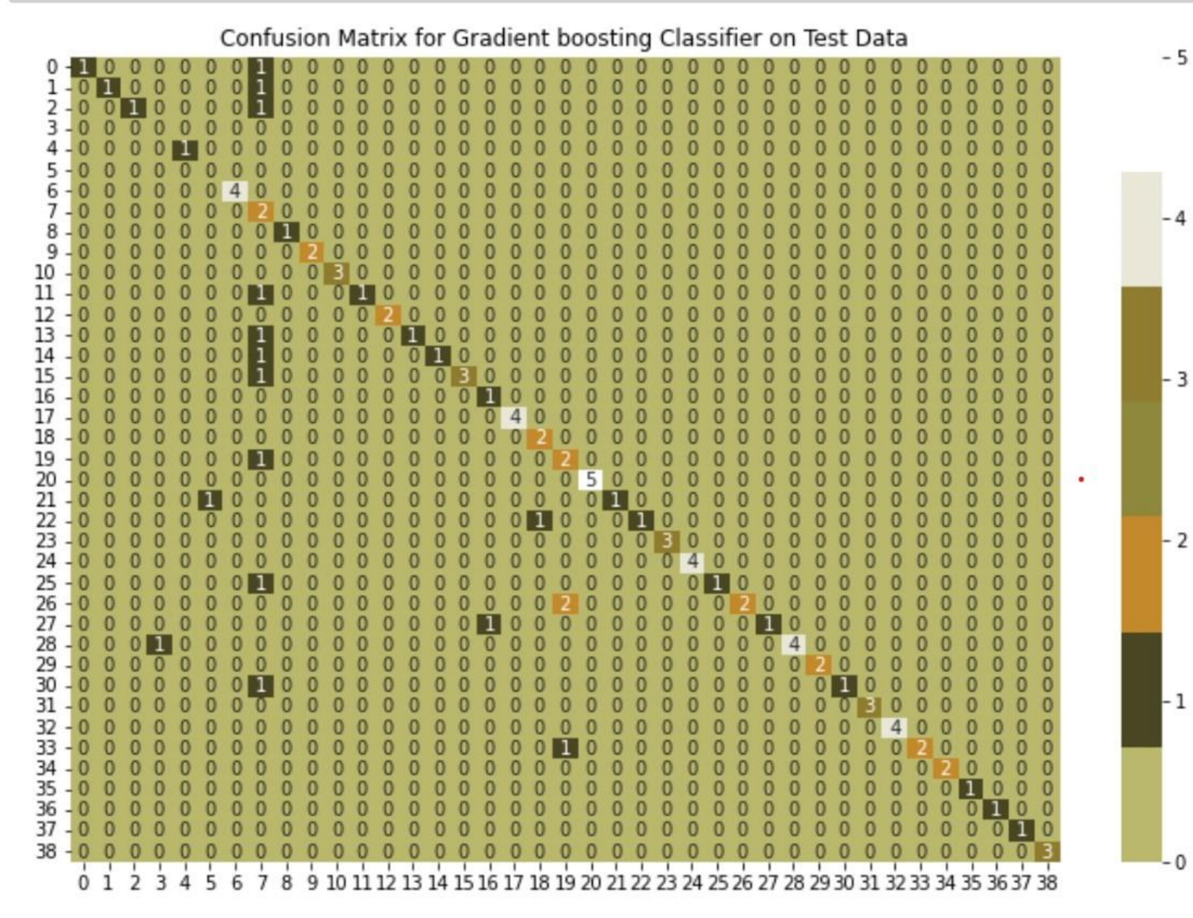


Figure 4. Confusion matrix for the Gradient Boosting.

This confusion matrix is formed using the TP,FP,TN, and FN, and the values of true-false predictions in this matrix are generated by fitting the train and test data in the GB model.

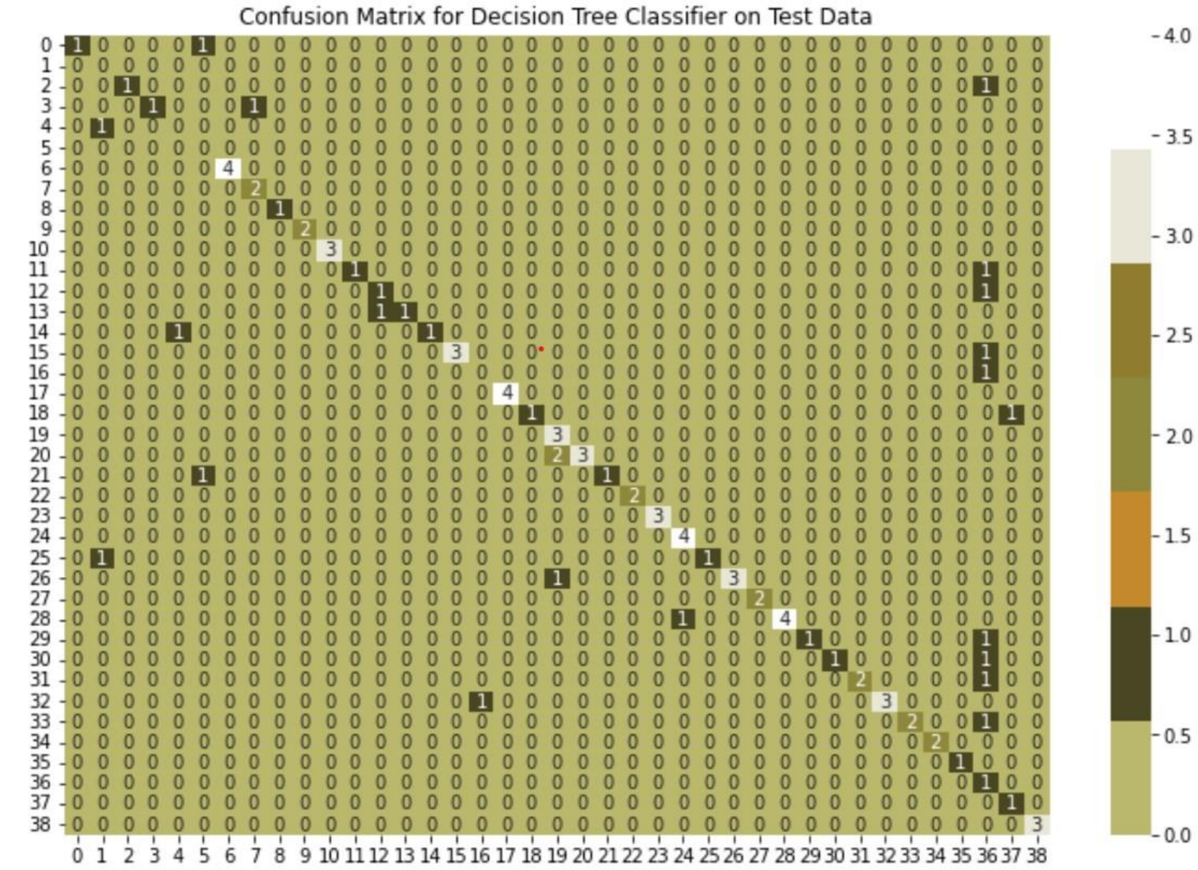
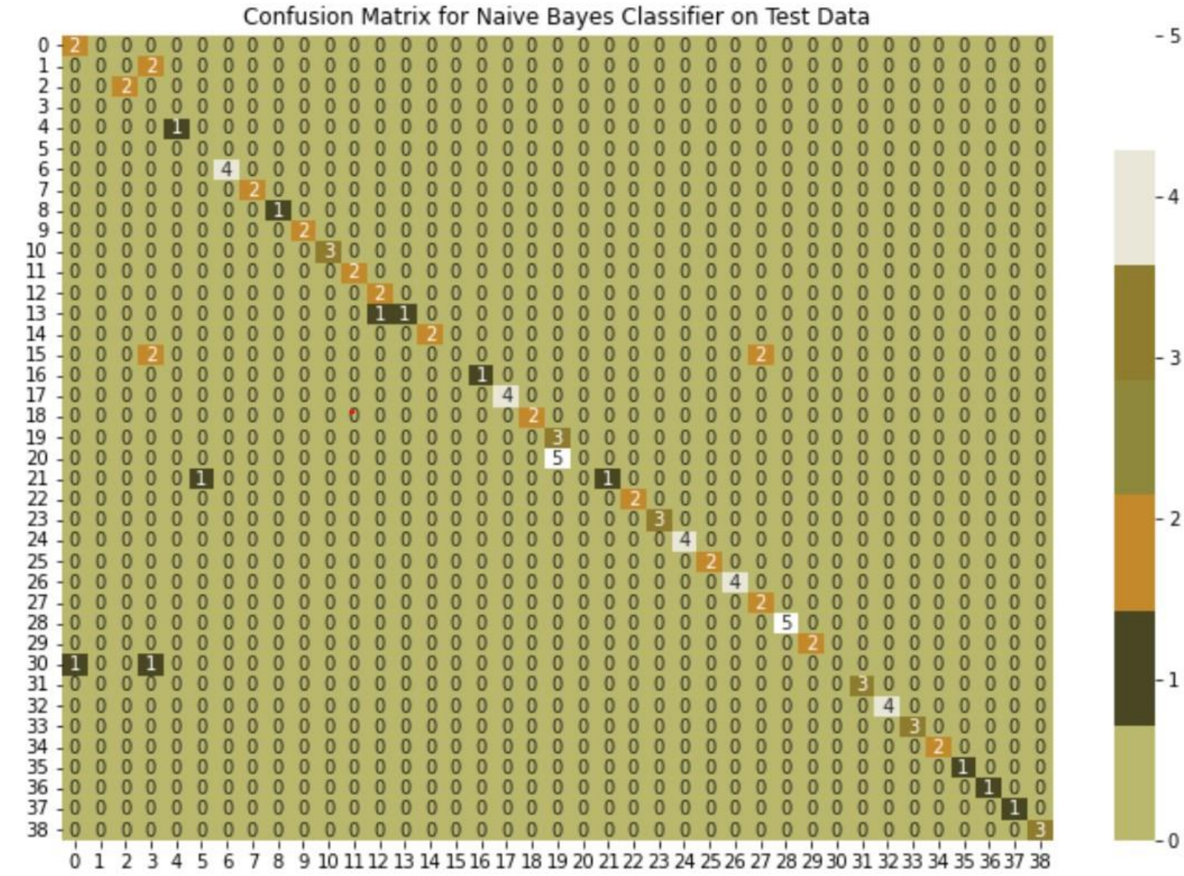


Figure 5. Confusion matrix for the Decision Tree.

The values of the true-false predictions in this confusion matrix are obtained by fitting the train and test data to the DT model. The confusion matrix is produced from the TP,FP,TN, and FN. and the 38 symptoms are the reason that the matrix is 38 x 38.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
30	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0
35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
37	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
38	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3

Figure 6. Confusion matrix for the Decision Tree.

The values of the true-false predictions in this confusion matrix are obtained by fitting the train and test data to the NB model. The confusion matrix is produced from the TP,FP,TN, and FN. and the 38 symptoms are the reason that the matrix is 38 x 38.

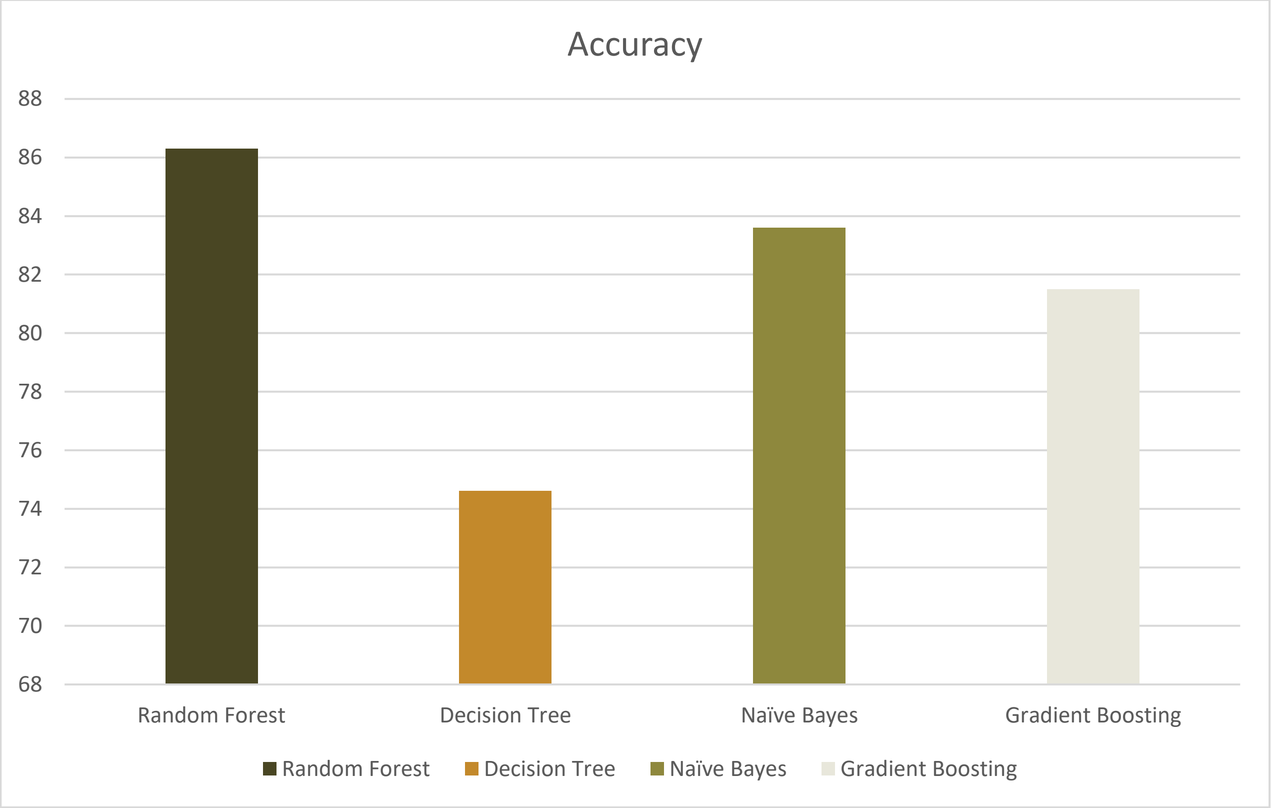


Figure 7. Results of the predicted models.

A graphical comparison of the specified algorithms and accuracy results were being displayed. This graph shows the variances in the four algorithms, relative prediction accuracy values, which are 86%, 81%, 74%, and 83% for Random Forest, Decision Tree, Nave Bayes, and Gradient Boosting. This demonstrates that when compared to other machine learning methods, the proposed approach obtains the greatest accuracy of 86% from Random Forest.

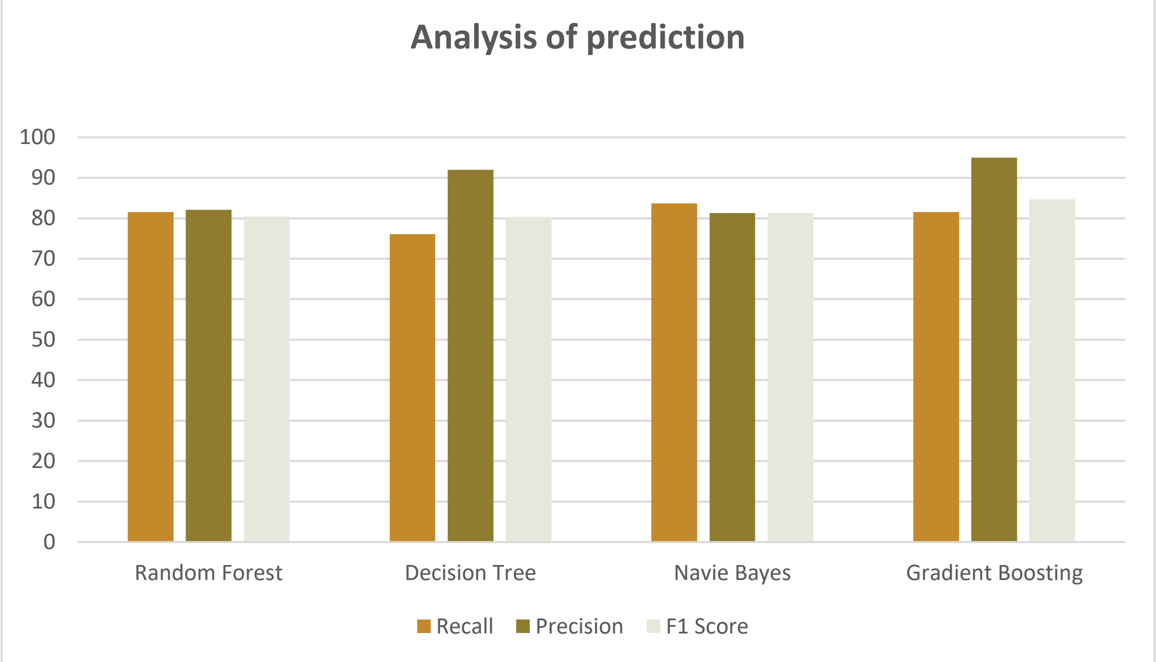


Figure 8. Predictive analysis results for the implemented algorithms

A graphical depiction for the results of implemented algorithms' comparison precision, recall, and F1-score values. This graph shows the variations in the precision, recall, and F1-score performance evaluation parameters for the four algorithms Random Forest, Decision Tree, Nave Bayes, and Gradient Boosting as they are, respectively, 82.08%, 91.99%, 81.29%, and 94.96% for precision, 81.52%, 76.08%, 83.69%, and 81.52% for recall. The Gradient Boosting is superior to the others, with values of 93%, 99%, and 97% for precision, recall, and F1-score, respectively.

5. Concluding Remarks

This paper deals with the identification of disease based on the symptoms. The analysis was done on the data set which was collected from the hosptals. Various datasets can be generated from different hospitals with different diseases and those predictions are used by doctors to recognize the percentage of the predicted disease. But in this project the data set was purely based on the symptoms that can be directly handled by the patients as per there knowledge of symptoms that they are facing. By doing data-preprocessing and feature selection, tuples are reduced by up to 70% and the most useful symptoms are evaluated, and the performance of algorithms such as Random Forest, Decision Tree, Nave Bayes, and Gradient Boosting is better than predicted. The findings demonstrate that the Random Forest method outperforms the other three algorithms with an accuracy of 86%.

6. Future Work

In accordance with this phase, the prediction component was being completed. The research paper that resulted from these models' conclusions will be published, and in subsequent days, the image data will also be added to this text data using deep-learning models like CNN, TensorFlow, etc., and the predictions will be made with this image and text data. Finally, the backend of a GUI interface that enables the user to forecast the desired disease using the symptoms as input will be connected to the stored model. The percentages of diseases that are associated with the user's symptoms will be the result that is displayed.

References

1. [Hamsagayathri, P., & Vigneshwaran], 2021. Symptoms Based Disease Prediction Using Machine Learning Techniques. In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 5, Publisher Site.
2. [Keniya, R., Khakharia, A., Shah, V., Gada, V., Manjalkar, R., Thaker, T., ... & Mehendale], 2020. Disease prediction from various symptoms using machine learning, pp. 7, Publisher site.
3. [Pingale, K., Surwase, S., Kulkarni, V., Sarage, S., & Karve,] 2019. Disease prediction using machine learning. International Research Journal of Engineering and Technology (IRJET), 6, 831-833, Publisher site.
4. [Grampurohit, S., & Sagarnal], 2020. Disease prediction using machine learning algorithms. In 2020 International Conference for Emerging Technology (INCET), pp. 1-7, Publisher site.
5. [Yaganteeswarudu, A.], 2020, Multi disease prediction model by using machine learning and Flask API. In 2020 5th International Conference on Communication and Electronics Systems (ICCES) pp. 1242-1246., Publisher site.
6. [Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T.], 2021. Multiple disease prediction using Machine learning algorithms., pp.1-4, Publisher site.
7. [Nilashi, M., bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L.], 2017. An analytical method for diseases prediction using machine learning techniques. Computers & Chemical Engineering, 106, 212-223, Publisher site.
8. [Gomathi, K., & Priyaa, D. D.], 2016. Multi disease prediction using data mining techniques. International journal of system and software engineering, 4(2), 12-14, Publisher site.
9. [Harimoorthy, K., & Thangavelu, M.], 2021. Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. Journal of Ambient Intelligence and Humanized Computing, 12(3), 3715-3723, Publisher site.
10. [Nabeel, M., Majeed, S., Awan, M. J., Muslih-ud-Din, H., Wasique, M., & Nasir, R.], 2021. Review on Effective Disease Prediction through Data Mining Techniques. International Journal on Electrical Engineering & Informatics, 13(3), Publisher site.
11. [Dahiwade, D., Patle, G., & Meshram, E.], 2019. Designing disease prediction model using machine learning approach. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1211-1215), Publisher site.
12. [Kohli, P. S., & Arora], 2018. Application of machine learning in disease prediction. In 2018 4th International conference on computing communication and automation (ICCCA) (pp. 1-4), Publisher site.

13. [G. Battineni, G. G. Sagaro, N. Chinatalapudi, and F Amenta], 2020. Applications of machine learning predictive models in the chronic disease diagnosis, Journal of Personalized Medicine, vol. 10, no. 2, p. 21, Publisher Site.
14. [B. Manjulatha and P. Suresh], 2021. An ensemble model for predicting chronic diseases using machine learning algorithms, in *Smart Computing Techniques and Applications*, pp. 337–345, Publisher Site
15. [Alanazi, R.], 2022. Identification and prediction of chronic diseases using machine learning approach. *Journal of Healthcare Engineering, pp. 1-12, Publisher Site*.