VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"JnanaSangama", Belgaum -590014, Karnataka.

PROJECT WORK-4 REPORT
on

"Analyzing classification algorithms to predict the disease"

Submitted by

Harshitha Mahadev (1BM19CS059)
Jahnavi Satish Shanbhag(1BM19CS065)
Harshitha R M (1BM19CS060)
Himani B M (1BM19CS062)

Under the Guidance of

Dr. Selva Kumar S

Assistant Professor, BMSCE

in partial fulfillment for the award of the degree of
BACHELOR OF ENGINEERING
in
COMPUTER SCIENCE AND ENGINEERING

B. M. S. COLLEGE OF ENGINEERING
(Autonomous Institution under VTU)
BENGALURU-560019

April-2022 to July-2022

B. M. S. College of Engineering,

Bull Temple Road, Bangalore 560019

(Affiliated To Visvesvaraya Technological University, Belgaum)

Department of Computer Science and Engineering

CERTIFICATE

This is to certify that the project work entitled "**Analyzing classification algorithms to predict a disease**" carried out by **Harshitha Mahadev (1BM19CS059) , Jahnavi Satish Shanbhag(1BM19CS065), Harshitha R M (1BM19CS060) and Himani B M (1BM19CS062)** who are bonafide students of **B. M. S. College of Engineering.** It is in partial fulfillment for the award of **Bachelor of Engineering in Computer Science and Engineering** of the Visveswaraiah Technological University, Belgaum during the year 2021. The project report has been approved as it satisfies the academic requirements in respect of **Project Work-4 (20CS6PWPW4)** work prescribed for the said degree.

Signature of the Guide Signature of the HOD Prof. Namratha M Dr. Jyothi S Nayak
Assistant Professor Professor & Head, Dept. of CSE BMSCE, Bengaluru BMSCE,
Bengaluru

External Viva Name of the Examiner Signature with date

1.______________________________ __________________________

2. ______________________________ __________________________

B. M. S. COLLEGE OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We, Harshitha Mahadev (1BM19CS059) , Jahnavi Satish Shanbhag(1BM19CS065), Harshitha R M (1BM19CS060) and Himani B M (1BM19CS062), students of 6th Semester, B.E, Department of Computer Science and Engineering, B. M. S. College of Engineering, Bangalore, hereby declare that, this Project Work-1entitled "Project Title" has been carried out by us under the guidance of Dr. Selva Kumar S, Assistant Professor, Department of CSE, B. M. S. College of Engineering, Bangalore during the academic semester Mar-2021-Jun-2021

We also declare that to the best of our knowledge and belief, the development reported here is not from part of any other report by any other students.

Signature

Harshitha Mahadev (1BM19CS059)

Jahnavi Satish Shanbhag(1BM19CS065)

Harshitha R M (1BM19CS060)

Himani B M (1BM19CS062)

1. Introduction

Healthcare data science has been growing rapidly for several years. The development and exploitation of several prominent Data mining techniques in numerous real-world application areas (e.g. Industry, Healthcare and Bio science) has led to the utilization of such techniques in machine learning environments, in order to extract useful pieces of information of the specified data in healthcare communities, biomedical fields etc. The techniques of machine learning have been successfully employed in assorted applications including Disease prediction. The aim of developing classifier system using machine learning algorithms is to immensely help to solve the health-related issues by assisting the physicians to predict and diagnose diseases at an early stage.

Three of the most prominent applications of data science in healthcare
are as follows:

- Image recognition for diagnostic purposes (e.g., malignancies and organ abnormalities)

- Developments towards precision medicine using genomics data (such as the 1000 Genomes Project from the National Institutes of Health), which tend to focus on chronic diseases like heart disease and diabetes

- Hospital operations analyses, which often look like standard business analyses (e.g., predicting when the ER will need more

1.1 Motivation

Disease prediction has the potential to benefit stakeholders such as the government and health insurance companies. It can identify patients at risk of disease or health conditions. Clinicians can then take appropriate measures to avoid or minimize the risk and in turn, improve quality of care and avoid potential hospital admissions. Due to the recent advancement of tools and techniques for data analytics, disease risk prediction can leverage large amounts of semantic information, such as demographics, clinical diagnosis and measurements, health behaviors , laboratory results, prescriptions and care utilization . In this regard, electronic health data can be a potential choice for developing disease prediction models. A significant number of such disease prediction models have been proposed in the literature over time utilizing large-scale electronic health databases, different methods, and healthcare variables.

1.2 Scope of the Project

The main objective of the project is predicting the diseases using classification algorithms . The algorithms used are KNN , decision tree , k means and random forest. So first we give the inputs . The inputs are the symptoms the patient is suffering from . For the give input we predict the diseases using the four algorithms and analyze the results. This analysis helps in predicting better results in the decision making.

1.3 Problem statement

Prediction of health diseases using classification algorithms in machine learning using symptoms as features. Comparing the results obtained and plotting the ROC curve.

2. Literature Survey

Name of the Research paper	Year Published	Algorithms Used	Results and Future Scope

Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms	2020	DT, KNN, NB, RFT, SVM	The accuracy of naïve Bayes was found to be the highest, although Random Forest was identified as the best model. Due to the fact that this problem produced imbalanced classes, the best-model selection was made on the basis of the f1 score, which is used for cases of imbalanced partitioning.
Mental Health Prediction using Data Mining: A Systematic Review	2020	Random Forest, Decision tree	The data is being subject to various machine learning techniques to obtain labels. These classified labels will then be used to build a model to predict the mental health of an individual. The model that is built will be integrated to a website so that it can predict the outcome as per the details provided by the user.
Big data Analysis on The Management Content of College Students' Mental Health Education	2021	C4.5 decision tree	The cause analysis based on actual data can provide a reliable basis for psychological educators, to improve the efficiency and effectiveness of school psychological consultation.
Digital Mental Health Challenges and the Horizon Ahead for Solutions	2021		Investments in digital mental health should ensure their safety and workability. End users should encourage the use of innovative methods to encourage developers to effectively evaluate their products and services and to render them a worthwhile

			investment. Technology-enabled services in a hybrid model of care are most likely to be effective (eg, specialists using these services among vulnerable, at-risk populations but not severe cases of mental ill health).

Name of the Research paper	Year	Algorithms Used	Results and Future Scope
Prediction of Mental Health in Medical Workers During COVID-19 Based on Machine Learning	2021	Stepwise logistic regression, binary bat algorithm, hybrid improved dragonfly algorithm	The results show that the prediction accuracy of the proposed model is 92.55%, which is better than the existing algorithms. This method can be used to predict the mental health of global medical workers. In addition, the method proposed in this paper can also play a role in the appropriate work plan for medical workers.
Sentiment Analysis on the News to Improve Mental Health	2021	Sequential, LSTM, BERT, and SVM models	Results have been successful – 1,300 users rate the app at 4.9 stars, and 85% report improved mental health by using it.

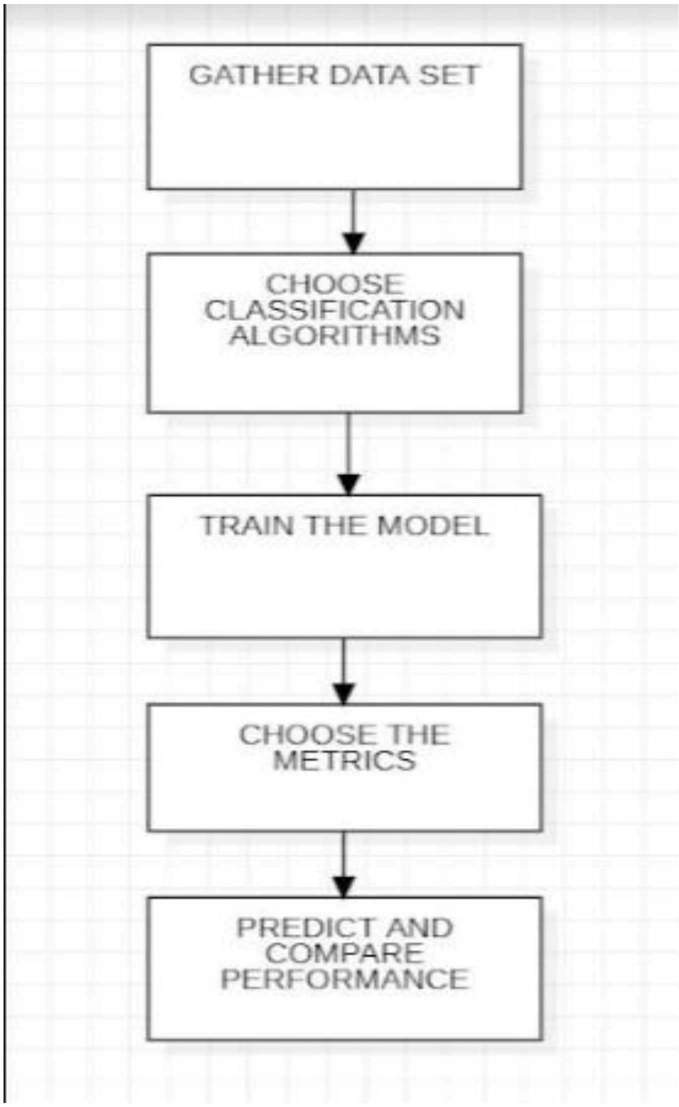
A machine learning algorithm to differentiate bipolar disorder from major depressive disorder using an online mental health questionnaire and blood biomarker data	2021	Extreme Gradient Boosting and nested cross-validation	Core predictors included elevated mood, grandiosity, talkativeness, recklessness and risky behaviour. Additional validation in participants with no previous mood disorder diagnosis showed AUROCs of 0.89 (0.86–0.91) and 0.90 (0.87–0.91) for separating newly diagnosed BD (N = 98) from MDD (N = 112) and subclinical low mood (N = 120), respectively.
Natural language processing applied to mental illness detection: a narrative review	2022		The review reveals that there is an upward trend in mental illness detection NLP research. Deep learning methods receive more attention and perform better than traditional machine learning methods.
Intelligent Parkinson Disease Prediction Using Machine Learning Algorithms	2013	Neural Networks, DMneural, Regression and Decision Trees	Analysis of voice data is important in the present decade to understand and diagnostic methods for human diseases. The present method provides the diagnosis of PD using voice dataset through machine learning algorithm
Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction	2011	Decision Tree , Bayesian classification, KNN, Neural Networks	The outcome of predictive data mining technique on the same dataset reveals that Decision Tree outperforms and some time Bayesian classification is having similar accuracy as of

			decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering are not performing well
Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques	2021	K-Nearest Neighbor Algorithm (KNN), Decision Trees (DT), Genetic algorithm (GA), and Naive Bayes (NB)	The proposed hybrid HRFLM approach is used combining the characteristics of Random Forest (RF) and Linear Method (LM). HRFLM proved to be quite accurate in the prediction of heart disease. The future course of this research can be performed with diverse mixtures of machine learning techniques to better prediction techniques.
Comparing different supervised machine learning algorithms for disease prediction	2019	Naïve Bayes algorithm ,Random Forest (RF) algorithm	We found that the Support Vector Machine (SVM) algorithm is applied most frequently (in 29 studies) followed by the Naïve Bayes algorithm (in 23 studies). However, the Random Forest (RF) algorithm showed superior accuracy comparatively. Of the 17 studies where it was applied, RF showed the highest accuracy in 9 of them, i.e., 53%. This was followed by SVM which topped in 41% of the studies it was considered.

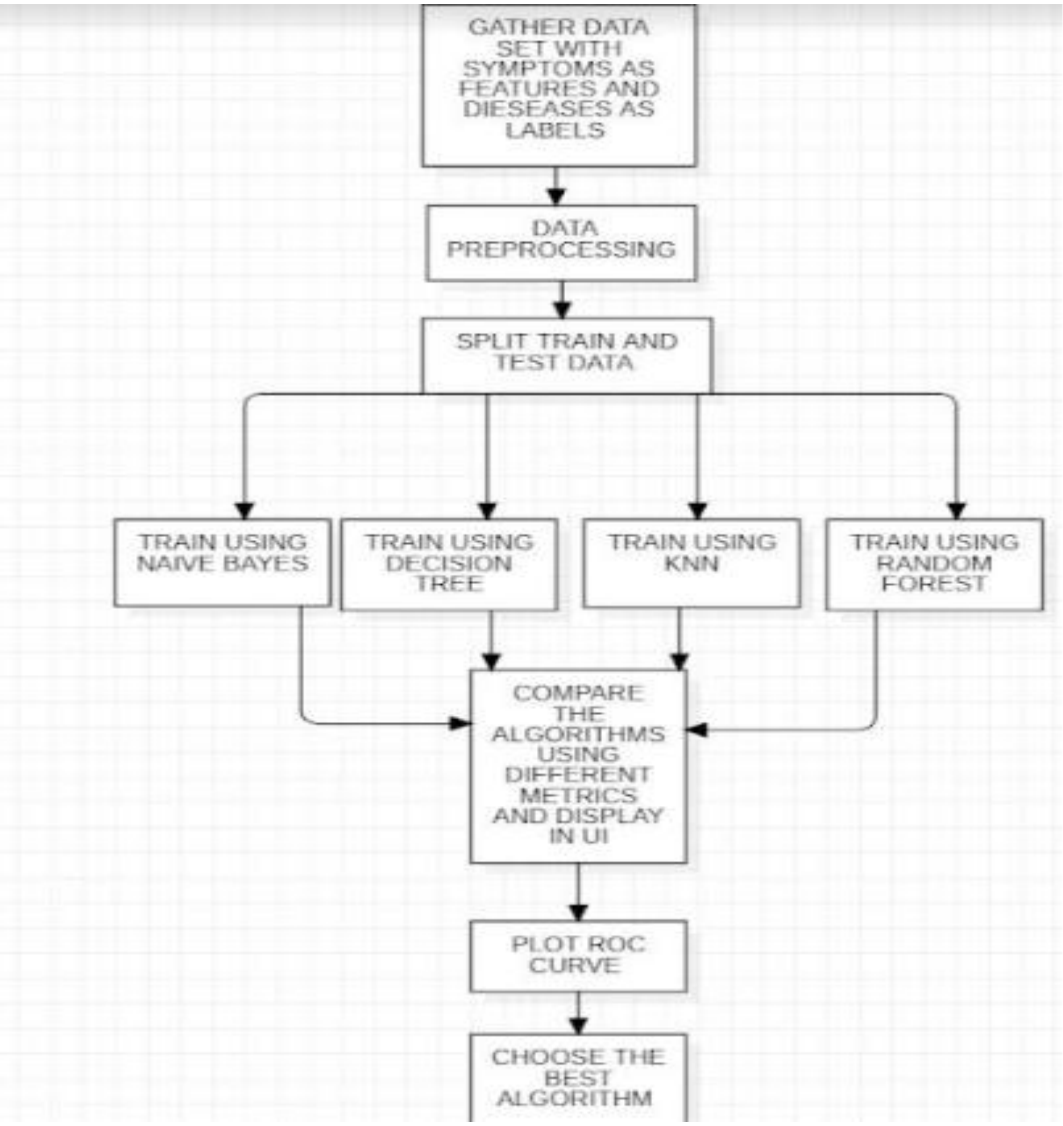
Clinical Data Analysis for Prediction of Cardiovascular Disease Using Machine Learning Techniques	2022	REP Tree, M5P Tree, Random Tree, Linear Regression, Naive Bayes, J48, and JRIP	When it came to the prediction of cardiovascular disease patients, the Random Tree model performed exceptionally well with the highest accuracy of 100%, the lowest MAE of 0.0011, the lowest RMSE of 0.0231, and the quickest prediction time of 0.01(secs). Future research could focus on enhancing the given CDPS model to achieve better performance in the classification of other types of medical data, resulting in a more cost-effective and time-saving option for both patients and doctors.

3. Design

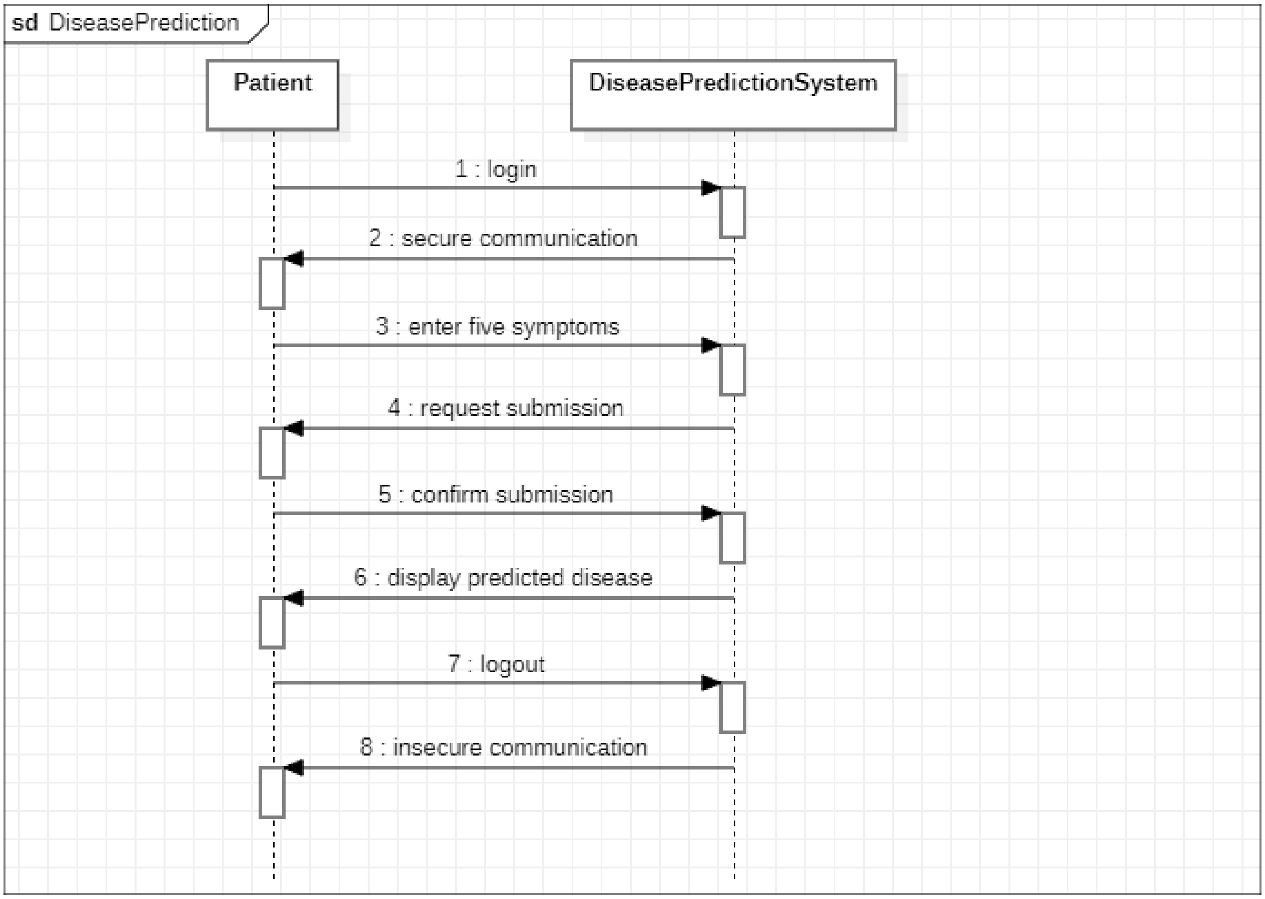
3.1 High Level Design


GATHER DATA SET
CHOOSE CLASSIFICATION ALGORITHMS
TRAIN THE MODEL
CHOOSE THE METRICS
PREDICT AND COMPARE PERFORMANCE

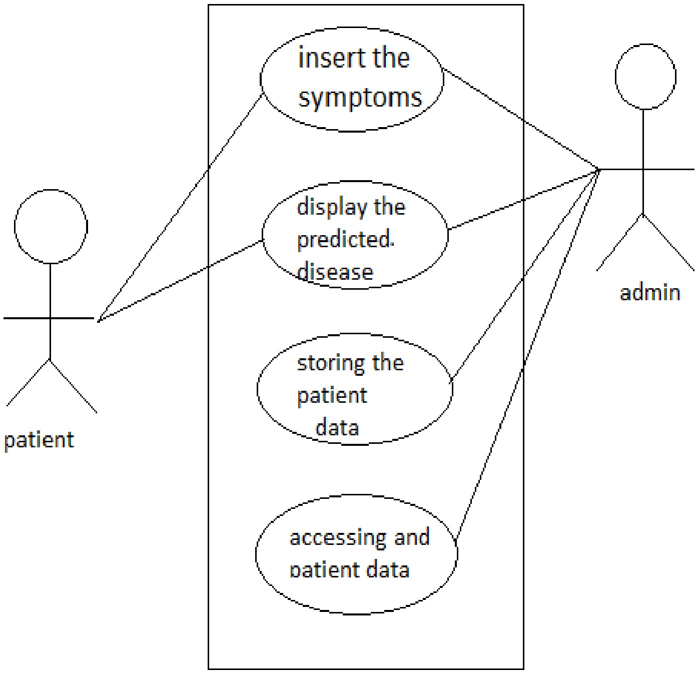
3.2 Detailed Design


GATHER DATA SET WITH SYMPTOMS AS FEATURES AND DIESEASES AS LABELS
DATA PREPROCESSING
SPLIT TRAIN AND TEST DATA
TRAIN USING NAIVE BAYES
TRAIN USING DECISION TREE
TRAIN USING KNN
TRAIN USING RANDOM FOREST
COMPARE THE ALGORITHMS USING DIFFERENT METRICS AND DISPLAY IN UI
PLOT ROC CURVE
CHOOSE THE BEST ALGORITHM

3.3 Sequence Diagram



3.4 Use Case Diagram



The use case diagram is as shown above.

Actions:

- Insert the symptoms - User is supposed to choose 5 symptoms among the options.
- Display the predicted disease - The disease prediction along with the analyzing of the different algorithms is shown on the screen.
- Storing the patient data - The symptoms chosen by the user is stored in the server in order to analyze the result using the algorithms.
- Accessing the patient data - The admin only will be able to access the data.

4. Implementation

The symptoms chosen by the user are stored and used to classify and predict the disease. Four different algorithms are used to analyze the results.

4.1 Proposed methodology

- Classification algorithms in machine learning are used to predict a disease, given the symptoms.
- K-nearest neighbor, Random Forest algorithm, Naive Bayes Classifier and Decision Tree algorithm are used to predict the disease.
- The accuracy of the prediction using each algorithm is compared with each other.
- After analyzing the results, the prediction is shown with the appropriate algorithm that gave the result.

4.2 Algorithm used for implementation

Random Forest algorithm

- Random Forest is a supervised machine learning algorithm that is used widely in classification and regression problems.
- The forest it builds is an ensemble of decision trees, usually trained with the bagging method.
- The general idea of the bagging method is that a combination of learning methods increases the overall result.

K-Nearest Neighbor

- K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- The K-NN algorithm assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories.

Naive Bayes Classifier

- The Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is mainly used in text classification that includes a high-dimensional training dataset.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

Decision tree algorithm

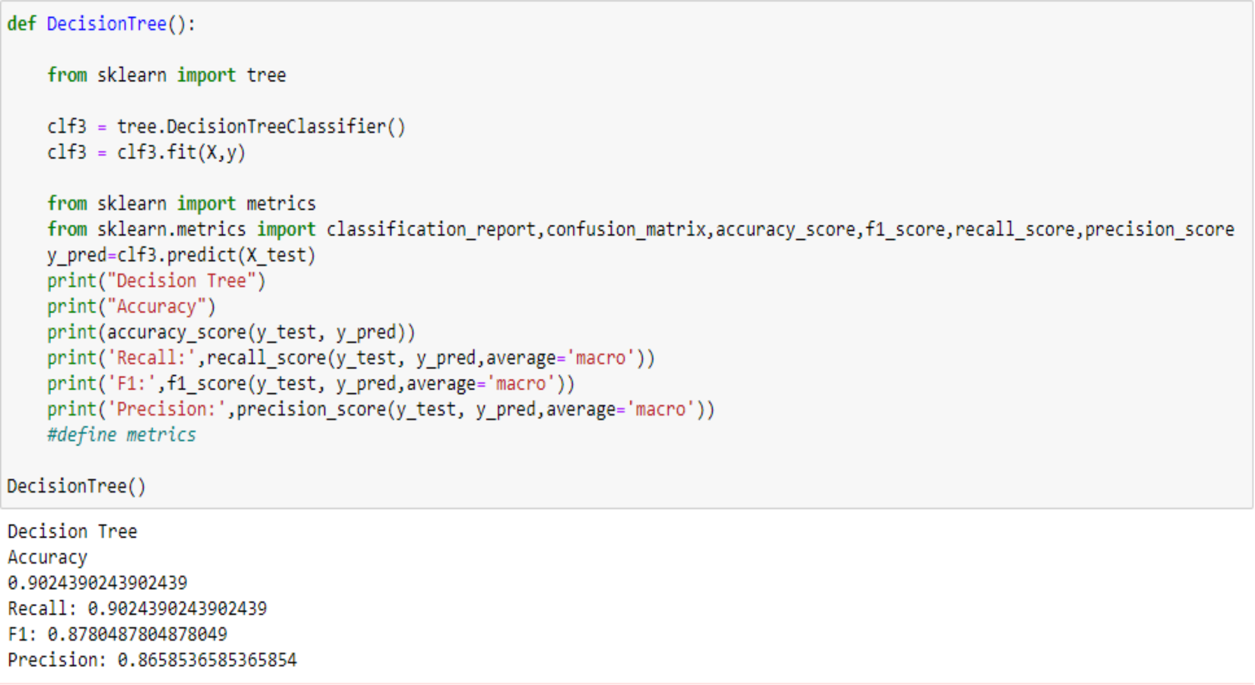
- Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.
- It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

4.3 Tools and technologies used

- Django for frontend
- Python machine learning libraries like numpy, pandas, matplotlib and scikit learn for developing the machine learning model

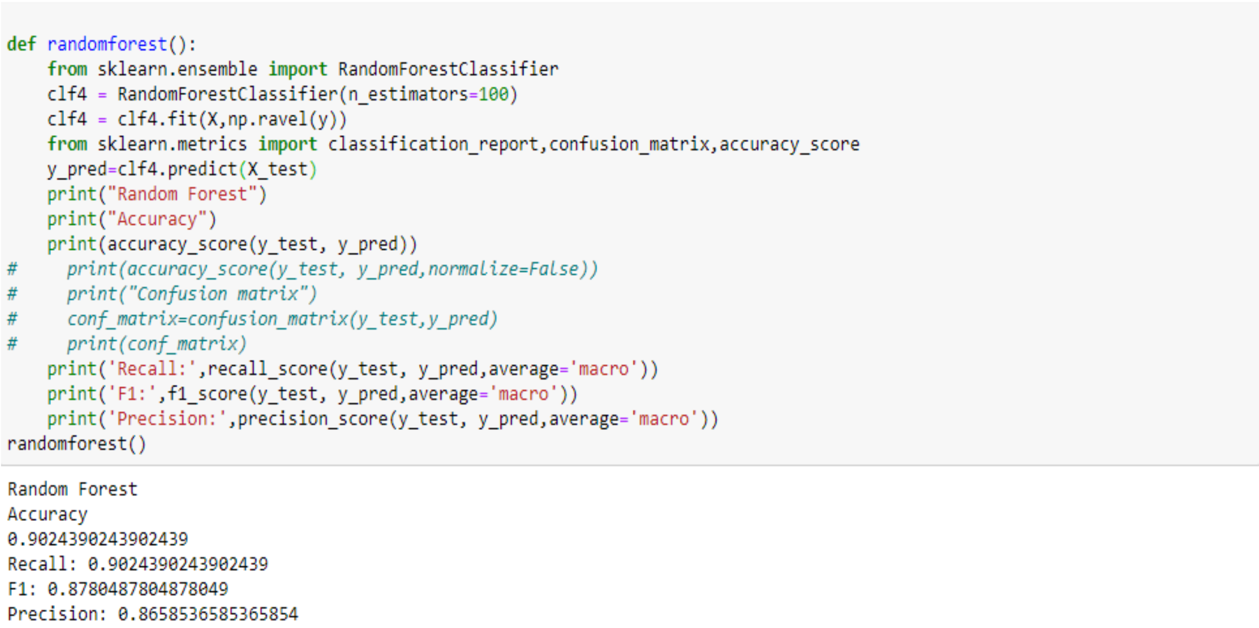
4.4 Testing

The interface looks like as shown in the picture below. The user selects the symptoms among the given symptoms.



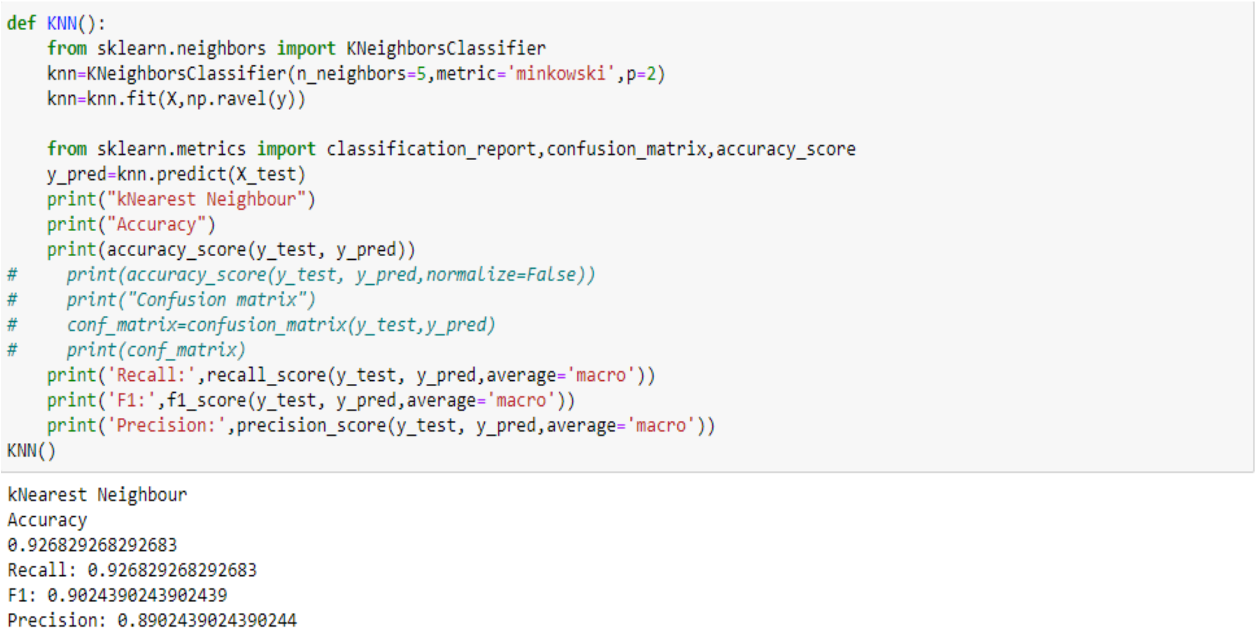
```
def DecisionTree():

    from sklearn import tree

    clf3 = tree.DecisionTreeClassifier()
    clf3 = clf3.fit(X,y)

    from sklearn import metrics
    from sklearn.metrics import classification_report,confusion_matrix,accuracy_score,f1_score,recall_score,precision_score
    y_pred=clf3.predict(X_test)
    print("Decision Tree")
    print("Accuracy")
    print(accuracy_score(y_test, y_pred))
    print('Recall:',recall_score(y_test, y_pred,average='macro'))
    print('F1:',f1_score(y_test, y_pred,average='macro'))
    print('Precision:',precision_score(y_test, y_pred,average='macro'))
    #define metrics

DecisionTree()
```

```
Decision Tree
Accuracy
0.9024390243902439
Recall: 0.9024390243902439
F1: 0.8780487804878049
Precision: 0.8658536585365854
```



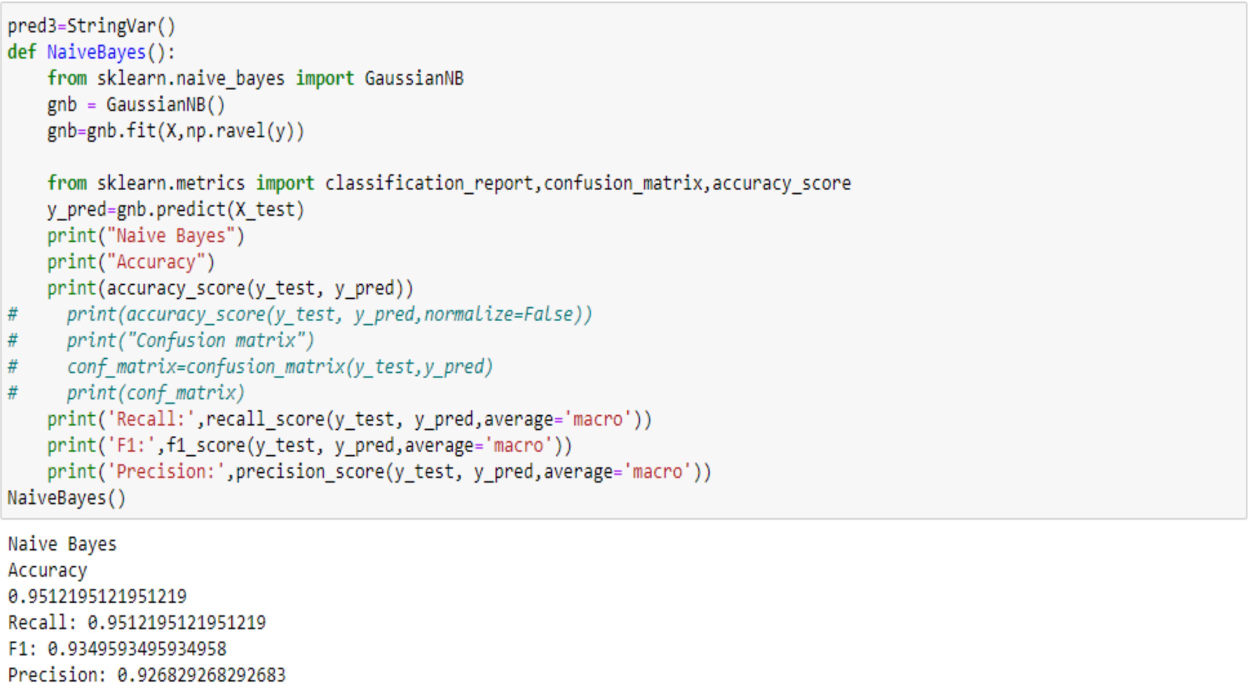
```
def randomforest():
    from sklearn.ensemble import RandomForestClassifier
    clf4 = RandomForestClassifier(n_estimators=100)
    clf4 = clf4.fit(X,np.ravel(y))
    from sklearn.metrics import classification_report,confusion_matrix,accuracy_score
    y_pred=clf4.predict(X_test)
    print("Random Forest")
    print("Accuracy")
    print(accuracy_score(y_test, y_pred))
#     print(accuracy_score(y_test, y_pred,normalize=False))
#     print("Confusion matrix")
#     conf_matrix=confusion_matrix(y_test,y_pred)
#     print(conf_matrix)
    print('Recall:',recall_score(y_test, y_pred,average='macro'))
    print('F1:',f1_score(y_test, y_pred,average='macro'))
    print('Precision:',precision_score(y_test, y_pred,average='macro'))
randomforest()
```

```
Random Forest
Accuracy
0.9024390243902439
Recall: 0.9024390243902439
F1: 0.8780487804878049
Precision: 0.8658536585365854
```



```
def KNN():
    from sklearn.neighbors import KNeighborsClassifier
    knn=KNeighborsClassifier(n_neighbors=5,metric='minkowski',p=2)
    knn=knn.fit(X,np.ravel(y))

    from sklearn.metrics import classification_report,confusion_matrix,accuracy_score
    y_pred=knn.predict(X_test)
    print("kNearest Neighbour")
    print("Accuracy")
    print(accuracy_score(y_test, y_pred))
#     print(accuracy_score(y_test, y_pred,normalize=False))
#     print("Confusion matrix")
#     conf_matrix=confusion_matrix(y_test,y_pred)
#     print(conf_matrix)
    print('Recall:',recall_score(y_test, y_pred,average='macro'))
    print('F1:',f1_score(y_test, y_pred,average='macro'))
    print('Precision:',precision_score(y_test, y_pred,average='macro'))
KNN()
```

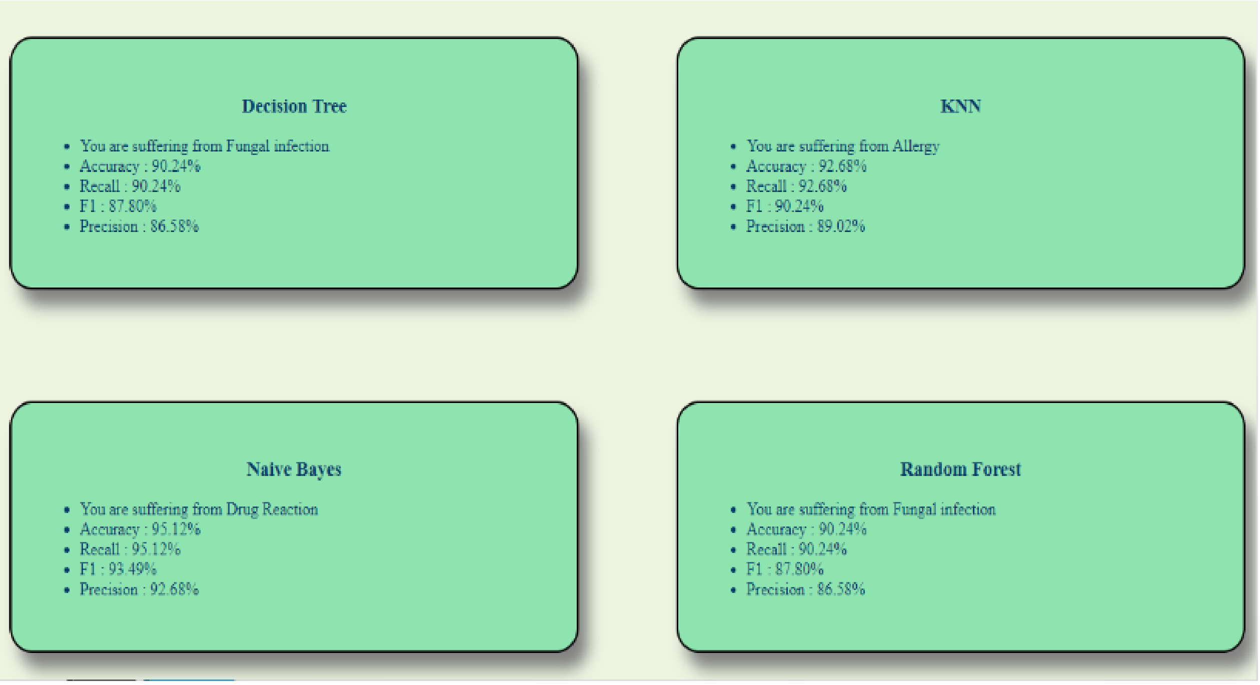
```
kNearest Neighbour
Accuracy
0.926829268292683
Recall: 0.926829268292683
F1: 0.9024390243902439
Precision: 0.8902439024390244
```



```
pred3=StringVar()
def NaiveBayes():
    from sklearn.naive_bayes import GaussianNB
    gnb = GaussianNB()
    gnb=gnb.fit(X,np.ravel(y))

    from sklearn.metrics import classification_report,confusion_matrix,accuracy_score
    y_pred=gnb.predict(X_test)
    print("Naive Bayes")
    print("Accuracy")
    print(accuracy_score(y_test, y_pred))
#     print(accuracy_score(y_test, y_pred,normalize=False))
#     print("Confusion matrix")
#     conf_matrix=confusion_matrix(y_test,y_pred)
#     print(conf_matrix)
    print('Recall:',recall_score(y_test, y_pred,average='macro'))
    print('F1:',f1_score(y_test, y_pred,average='macro'))
    print('Precision:',precision_score(y_test, y_pred,average='macro'))
NaiveBayes()
```

```
Naive Bayes
Accuracy
0.9512195121951219
Recall: 0.9512195121951219
F1: 0.9349593495934958
Precision: 0.926829268292683
```

5. Results and Discussion



Decision Tree

- You are suffering from Fungal infection
- Accuracy : 90.24%
- Recall : 90.24%
- F1 : 87.80%
- Precision : 86.58%

KNN

- You are suffering from Allergy
- Accuracy : 92.68%
- Recall : 92.68%
- F1 : 90.24%
- Precision : 89.02%

Naive Bayes

- You are suffering from Drug Reaction
- Accuracy : 95.12%
- Recall : 95.12%
- F1 : 93.49%
- Precision : 92.68%

Random Forest

- You are suffering from Fungal infection
- Accuracy : 90.24%
- Recall : 90.24%
- F1 : 87.80%
- Precision : 86.58%

The disease is predicted and the accuracy score for the respective algorithms are also shown to the user.

Naive Bayes is a linear classifier which tends to be faster when applied to big data

Naive Bayes offers you two hyperparameters to tune for smoothing: alpha and beta. A hyperparameter is a prior parameter that is tuned on the training set to optimize it.

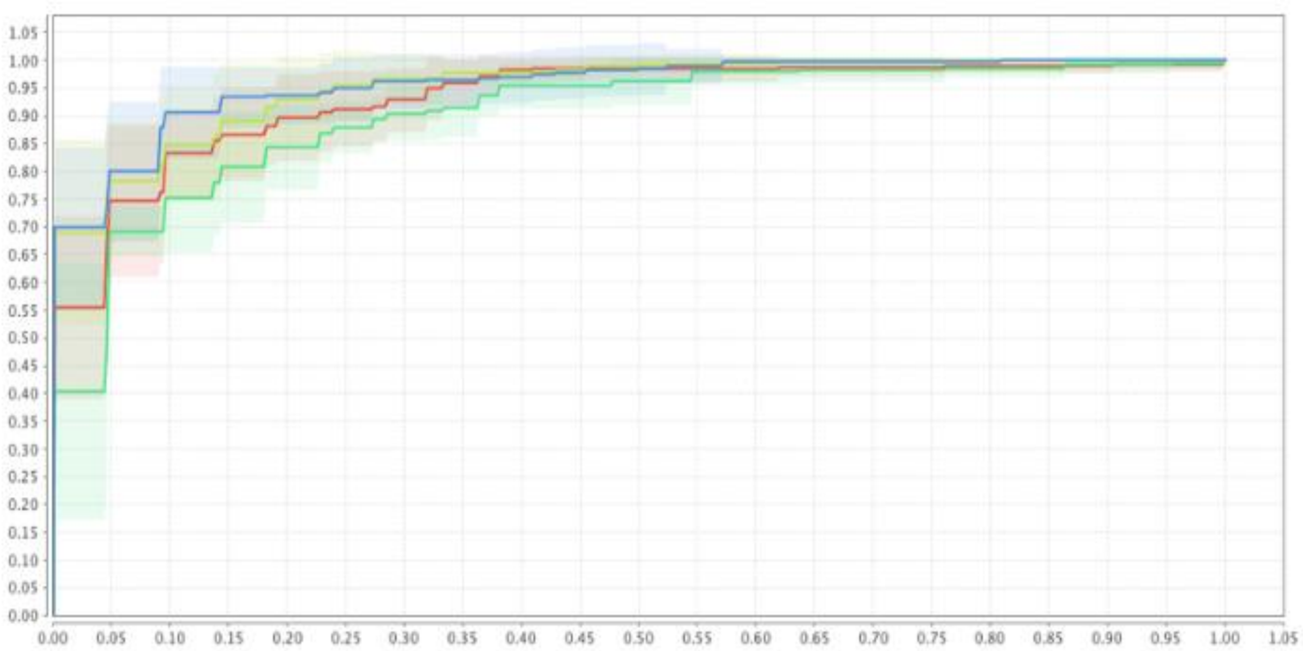
Overfitting is a major problem with decision trees

K-NN (and Naive Bayes) outperform decision trees when it comes to rare occurrences.

Building up of clutter to the point where rooms become unusable

Random Forest is comparatively less impacted by noise.

ROC curve



- Naive Bayes
- KNN
- Random forest
- Decision tree

1. Here Naive Bayes is the best algorithm for prediction
2. KNN is a better algorithm for prediction
3. Both decision tree and random forest algorithms are good for predicting the disease

6. Conclusion and Future Work

In this project, we predict the disease based on the given symptoms by using four supervised classification algorithms which are Naive Bayes, K Nearest Neighbors, Decision tree and Random Forest. The performance of these algorithms is measured for the given dataset in the table. Also from the ROC curve we can infer that Naive Bayes is the best algorithm for prediction of diseases amongst the four.

Algorithms	Accuracy(%)	Precision(%)	F1 score(%)	Recall(%)
Naive Bayes	95.12	92.68	93.49	95.12
KNN	92.68	89.02	90.24	92.68
Decision tree	90.24	86.58	87.80	90.24
Random forest	90.24	86.58	87.80	90.24

As a part of future work, we can test these algorithms for the larger datasets. And also we can use the other existing classification algorithms for testing this dataset which can perform even better.

References:

[1] Hasan, M., Islam, M.M., Zarif, M.I.I. and Hashem, M.M.A., “Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches”, Internet of Things Journal, Volume 7, Page no.10-15.

[2] Lippi, G. & Plebani, M. Procalcitonin in patients with severe coronavirus disease 2019 (COVID-19): A meta-analysis. Clin. Chim. Acta Int. J. Clin. Chem. 505, 190 (2020).

[3] H. H. Patel and P. Prajapati, “Study and analysis of decision tree based classification algorithms,” International Journal of Computer Science and Engineering, vol. 6, no. 10, pp. 74–78, 2018.

[4] M. J. Aitkenhead, “A co-evolving decision tree classification method,” Expert Systems with Applications, vol. 34, no. 1, pp. 18–25, 2008.

[5]Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering and Naïve Bayes algorithm. International Journal of Computer Science and Information Technologies, 6(1), 637-9.

APPENDIX:

Plagiarism report