MULTIPLE DISASE PREDICTION USING MACHINE LEARNING

With the growing demand for accessible healthcare solutions, machine learning has emerged as a powerful tool for early disease prediction and diagnosis. This project presents a **Multiple Disease Prediction System** that utilizes machine learning algorithms to predict various diseases based on user-provided symptoms. Three popular classification techniques—**Decision Tree**, **Support Vector Machine (SVM)**, and **Random Forest**—are implemented and evaluated using a labeled symptom-disease dataset. These models are assessed using metric such as **accuracy**.

Among the three, the **Random Forest classifier** demonstrates superior performance with the highest prediction accuracy and generalization capability. Owing to its robustness and ability to handle diverse inputs, Random Forest is selected for the final deployment phase.

To ensure user accessibility and real-time interaction, the final model is integrated into a web application developed using **Streamlit**, a lightweight and efficient Python-based framework. The web app provides an intuitive interface where users can select their symptoms from a list and instantly receive the most probable disease prediction along with basic information about the condition.

This system serves as a valuable aid for preliminary diagnosis, increasing health awareness and encouraging timely medical consultation, especially in remote or underserved areas.

Keywords: Diabetes, Machine Learning, Prediction, Dataset, Ensemble, Scikit-Learn, Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF).

1. INTRODUCTION

In the modern world, multiple diseases like heart disease, kidney failure, and respiratory issues are becoming common and life-threatening. These diseases often occur due to changing lifestyles, poor diet, pollution, and genetic factors. Early detection of such diseases is essential to reduce the risk of complications and improve treatment outcomes. According to various global health reports, millions of people suffer from multiple health conditions, especially in developing countries. The increasing burden on healthcare systems highlights the need for automated solutions to assist in disease prediction and awareness. To address this, machine learning techniques are widely used in the healthcare sector due to their ability to learn from medical data and make accurate predictions. In this work, we explore the prediction of multiple diseases by considering various symptoms and health-related attributes. For this purpose, a symptom-disease dataset is used, and multiple classification techniques such as Decision Tree, Support Vector Machine (SVM), and Random Forest are applied. Machine learning helps in training the model from this data to identify patterns and predict possible diseases. Though various techniques exist, it is challenging to choose the most suitable one. Therefore, we evaluate and compare the performance of each model. The main contributions of this work include: using symptom-based input for predicting multiple diseases, comparing different machine learning models, and identifying the most accurate model for practical use.

• Implementation and evaluation of traditional and ensemble machine learning models to predict eight complications in diabetic patients by utilizing a comprehensive UAE based dataset.

• Identification of the dominant characteristics that may lead to disease complications using feature selection methods.

1.1 AN OVERVIEW OF MULTIPLE DISEASE PREDICTION

It follows eight steps:

Step1: Import required libraries, Import diabetes dataset.

Step2: Pre-process data to remove missing data.

Step3: Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.

Step4: Select the machine learning algorithm i.e. Support Vector Machine, Decision Tree and Random Forest algorithm.

Step5: Build the classifier model for the mentioned machine learning algorithm based on training set.

Step6: Test the Classifier model for the mentioned machine learning algorithm based on test set.

Step7: Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

Step8: After analysing based on various measures conclude the best performing algorithm.

1.2. MULITPLE DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS-

Sentiment Analysis is a Machine Learning tool that analyses texts for polarity, from positive to negative. By training machine learning tools with examples of emotions in text, machines automatically learn how to detect sentiment without human input. There are a number of techniques and complex algorithms used to command and train machines to perform sentiment analysis. There are pros and cons to each. But, used together, they can provide exceptional results.

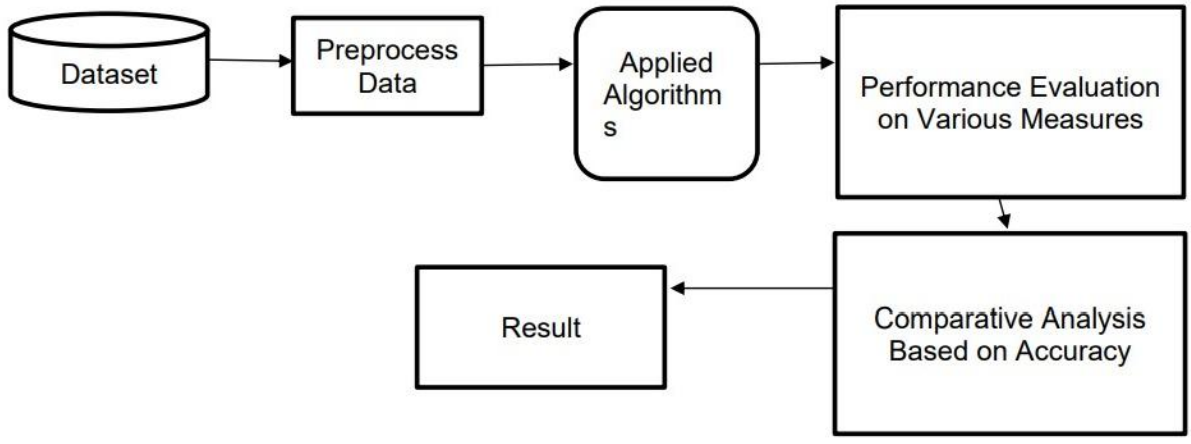


Figure1. Multiple Disease Prediction using Machine Learning

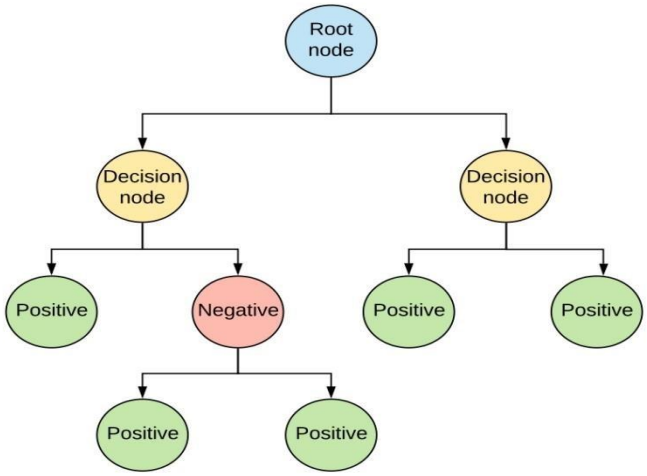
Machine Learning Algorithms for Multiple Disease Prediction:

1.2.1. Support Vector Machine:

Support Vector Machine also known as svm is a supervised machine learning algorithm. Svm is most popular classification technique. Svm creates a hyperplane that separate two classes. It can create a hyperplane or set of hyper plane in high dimensional space. This hyper plane can be used for classification or regression also. Svm differentiates instances in specific classes and can also classify the entities which are not supported by data. Separation is done by through hyperplane performs the separation to the closest training point of any class.

1.2.2. Decision Tree:

Decision tree is a basic classification method. It is supervised learning method. Decision tree used when response variable is categorical. Decision tree has tree like structure based model which describes classification process based on input feature. Input variables are any types like graph, text, discrete, continuous etc. Steps for Decision Tree Algorithm-

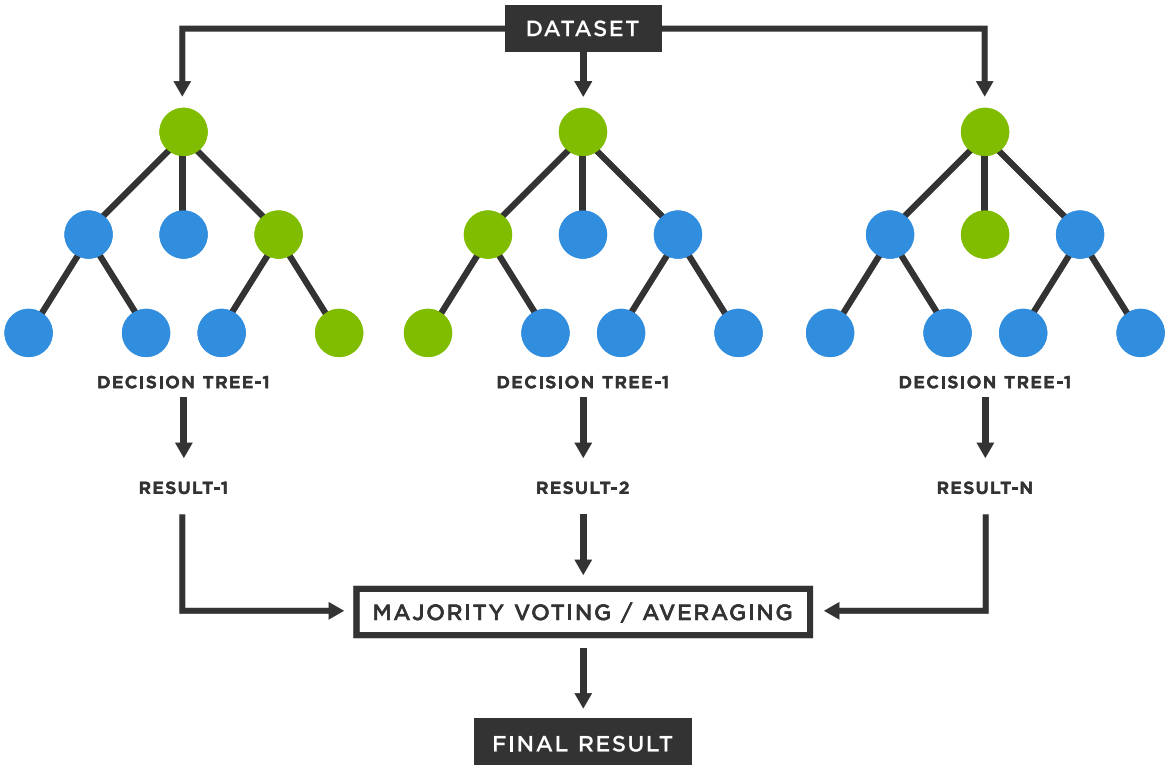


- Construct tree with nodes as input feature.
- Select feature to predict the output from input feature whose information gain is highest.

- The highest information gain is calculated for each attribute in each node of tree.
- Repeat step 2 to form a subtree using the feature which is not used in above node.

1.2.3. Random Forest (RF):

Random Forest It is type of ensemble learning method and also used for classification and regression tasks. The accuracy it gives is grater then compared to other models. This method can easily handle large datasets. Random Forest is developed by Leo Bremen. It is popular ensemble Learning Method. Random Forest Improve Performance of Decision Tree by reducing variance. It operates by constructing a multitude of decision trees at training



time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

2. DESIGN

2.1 ARCHITECTURE:

Steps for training a classifier for Multiple Disease prediction. Firstly, data have to be prepared in order to obtain a data set namely, the training set, by means of pre-processing and feature selection methods. Then, such a data set is involved in the learning step, which uses ML algorithms and yields a trained classifier. Finally, the classifier has to be tested on a different data set namely, the test set.

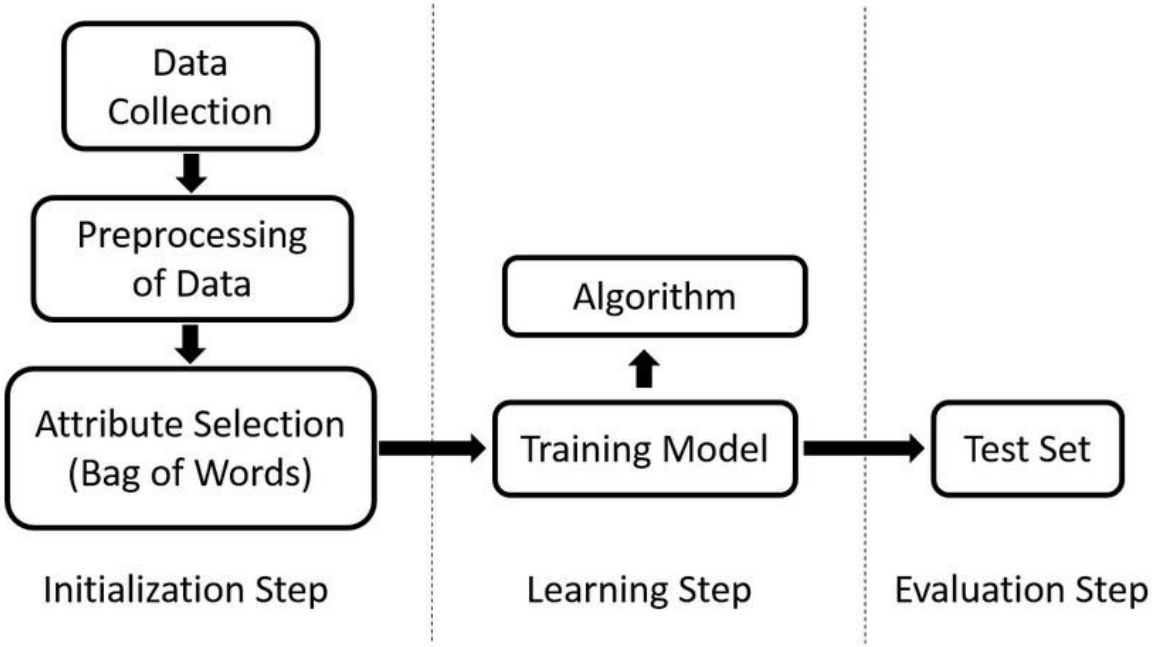
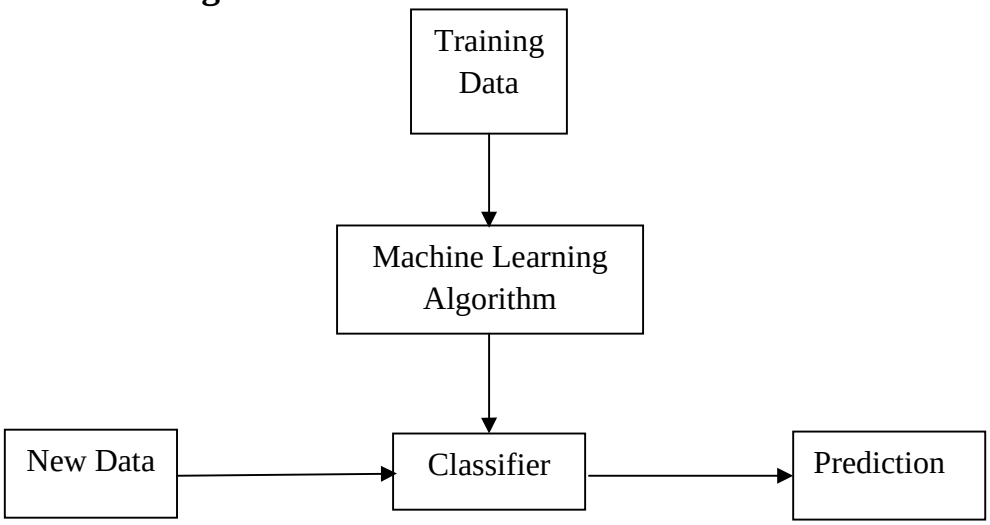


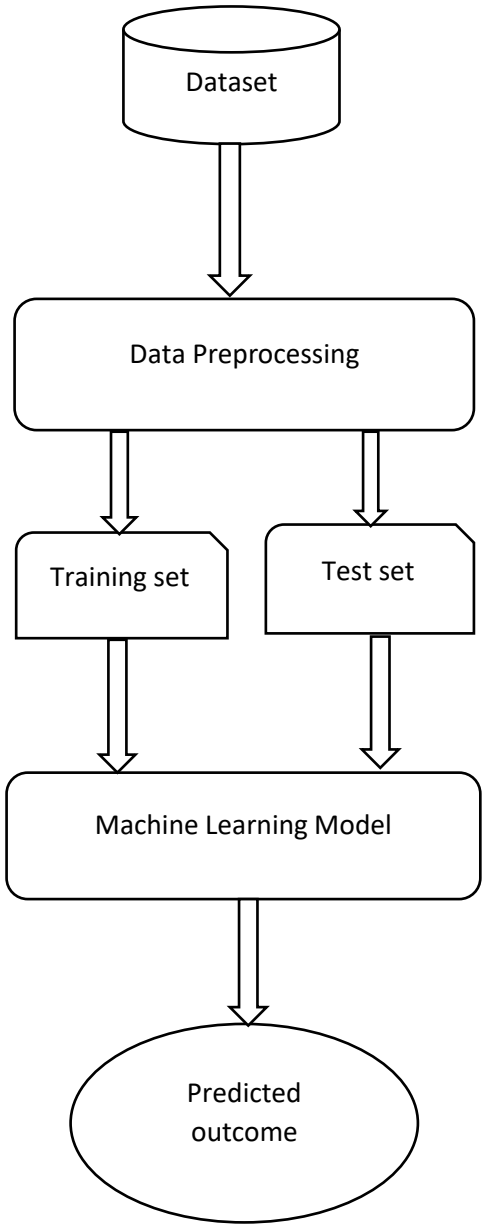
Figure 2 Steps to evaluate Sentiment Analysis

2.2 System Flow Diagram:



.

2.3 block diagram:



3. IMPLEMENTATION

We experimented with different classification and ensemble algorithms to predict diabetes. In the following, we briefly discuss the phase.

3.1. Dataset Description- the data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 768 patients.

```
multiple_disease_dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 45 entries, 0 to 44
Data columns (total 25 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Fever                  45 non-null     int64
 1   Shortness of breath    45 non-null     int64
 2   Cough                  45 non-null     int64
 3   Chest pain             45 non-null     int64
 4   Nausea                 45 non-null     int64
 5   Vommitting             45 non-null     int64
 6   Lightheadness          45 non-null     int64
 7   Sweating               45 non-null     int64
 8   Sudden weakness        45 non-null     int64
 9   Numbness               45 non-null     int64
 10  Confusion              45 non-null     int64
 11  Headache               45 non-null     int64
 12  Lump                   45 non-null     int64
 13  Weight loss            45 non-null     int64
 14  Fatigue                45 non-null     int64
 15  Bleeding               45 non-null     int64
 16  Seizures               45 non-null     int64
 17  Swelling               45 non-null     int64
 18  Conjuctivities         45 non-null     int64
 19  Diarrhea               45 non-null     int64
 20  Liver Damage           45 non-null     int64
 21  Cancer                 45 non-null     int64
 22  Stiff Neck             45 non-null     int64
 23  Pain in upper abdomen  45 non-null     int64
 24  Disease                45 non-null     object
dtypes: int64(24), object(1)
memory usage: 8.9+ KB
```

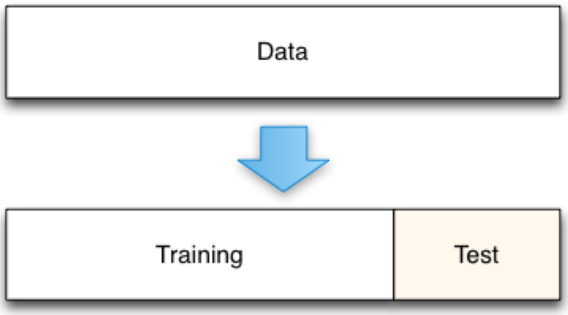
The 25th attribute is class variable of each data points. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics.

3.2 Data Preprocessing- Data preprocessing is most important process. Mostly healthcare related data contains missing vale and other impurities that can cause effectiveness of data. To improve quality and effectiveness obtained after mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively this process is essential for accurate result and successful prediction. For Pima Indian diabetes dataset we need to perform pre processing in two steps.

0. Missing Values removal- Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces dimensionality of data and help to work faster.

1. Splitting of data- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically aim of normalization is to bring all the attributes under same scale.

3.3.SPLITTING OF DATA: The Processed data is splitted into the training set and testing set based on the size of the dataset.



3.4.Apply Machine Learning- When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective to apply Machine Learning Techniques to analyse the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction. The Techniques are follows-

3.4.1.Support Vector Machine- Support Vector Machine also known as svm is a supervised machine learning algorithm. Svm is most popular classification technique. Svm creates a hyperplane that separate two classes. It can create a hyperplane or set of hyper plane in high dimensional space. This hyper plane can be used for classification or regression also. Svm differentiates instances in specific classes and can also classify the entities which are not supported by data. Separation is done by through hyperplane performs the separation to the closest training point of any class.

Algorithm-

- Select the hyper plane which divides the class better.
- To find the better hyper plane you have to calculate the distance between the planes and the data which is called Margin.
- If the distance between the classes is low then the chance of miss conception is high and vice versa. So we need to
- Select the class which has the high margin. Margin = distance to positive point + Distance to negative point.

3.4.2.Decision Tree- Decision tree is a basic classification method. It is supervised learning method. Decision tree used when response variable is categorical. Decision tree has tree like structure based model which describes classification process based on input feature. Input variables are any types like graph, text, discrete, continuous etc. Steps for Decision Tree Algorithm-

- Construct tree with nodes as input feature.

- Select feature to predict the output from input feature whose information gain is highest.
- The highest information gain is calculated for each attribute in each node of tree.
- Repeat step 2 to form a subtree using the feature which is not used in above node.

3.4.3. Random Forest- It is type of ensemble learning method and also used for classification and regression tasks. The accuracy it gives is grater then compared to other models. This method can easily handle large datasets. Random Forest is developed by Leo Bremen. It is popular ensemble Learning Method. Random Forest Improve Performance of Decision Tree by reducing variance. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction (regression) of the individual trees.

- The first step is to select the R features from the total features m where R<<M.
- Among the R features, the node using the best split point.
- Split the node into sub nodes using the best split.
- Repeat a to c steps until l number of nodes has been reached.
- Built forest by repeating steps a to d for a number of times to create n number of trees.

  The random forest finds the best split using the Gin-Index Cost Function which is given by:

  The first step is to need the take a glance at choices and use the foundations of each indiscriminately created decision tree to predict the result and stores the anticipated outcome at intervals the target place. Secondly, calculate the

votes for each predicted target and ultimately, admit the high voted predicted target as a result of the ultimate prediction from the random forest formula. Some of the options of Random Forest does correct predictions result for a spread of applications are offered.

4.RESULTS

By comparing all models on each other, RANDOM FOREST MODEL has more accuracy in predicting diabetes.

Classifier	**Accuracy**
Random forest	0.73
Decision Tree	0.73
Support Vector Machine	0.63

Table 1: Evaluation of Models

5.CONCLUSION AND FUTURE SCOPE

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning method in which SVM, Random Forest, Decision Tree classifiers are used. And 77% classification accuracy has been achieved. The Experimental results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.