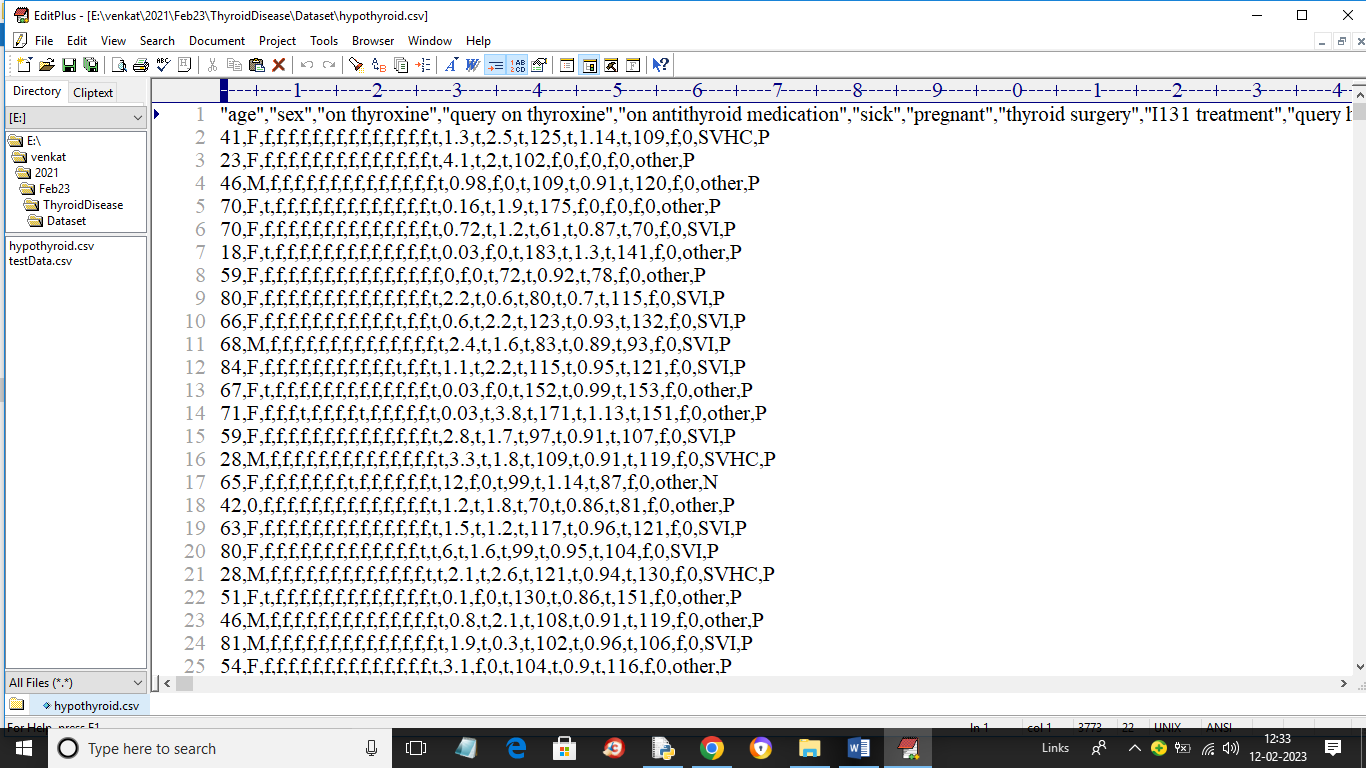
Thyroid Disease Detection using Machine Learning Algorithms

Every year worldwide millions of peoples are affected with thyroid disease and unattended or late diagnosis may worsen this disease and for timely prediction we are employing and evaluating performance of various machine learning algorithms such as SVM, Naïve Bayes and Random Forest and in all algorithms Random forest is giving best performance with an accuracy of 95 to 99%.

To implement this project we have used below thyroid dataset which contains more than 5000 records and below screen showing dataset details



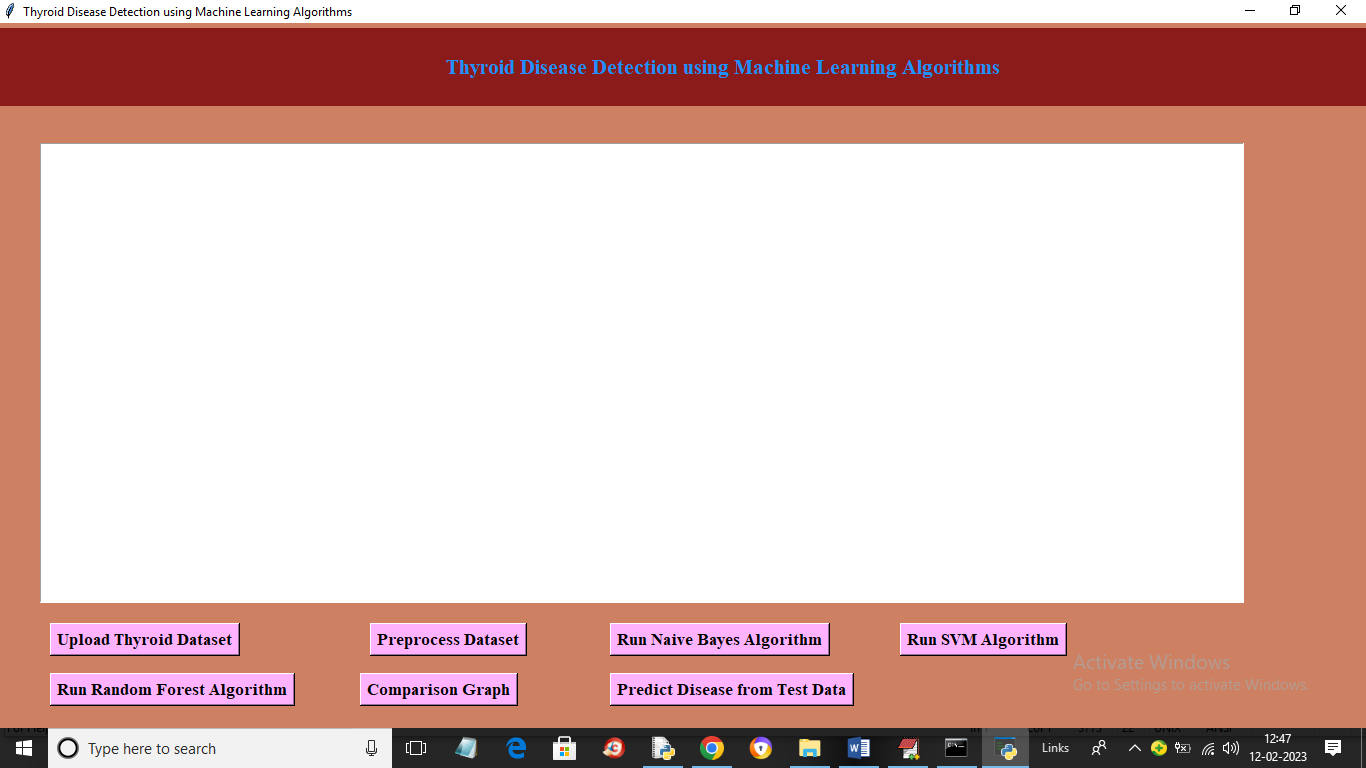
In above dataset screen first row contains dataset column names and remaining rows contains dataset values and in last column we have two class labels called ‘P’ (means presence of thyroid) and ‘N’ (means Normal). So by using above dataset we will train and evaluate performance of all algorithms.

To implement this project we have designed following modules

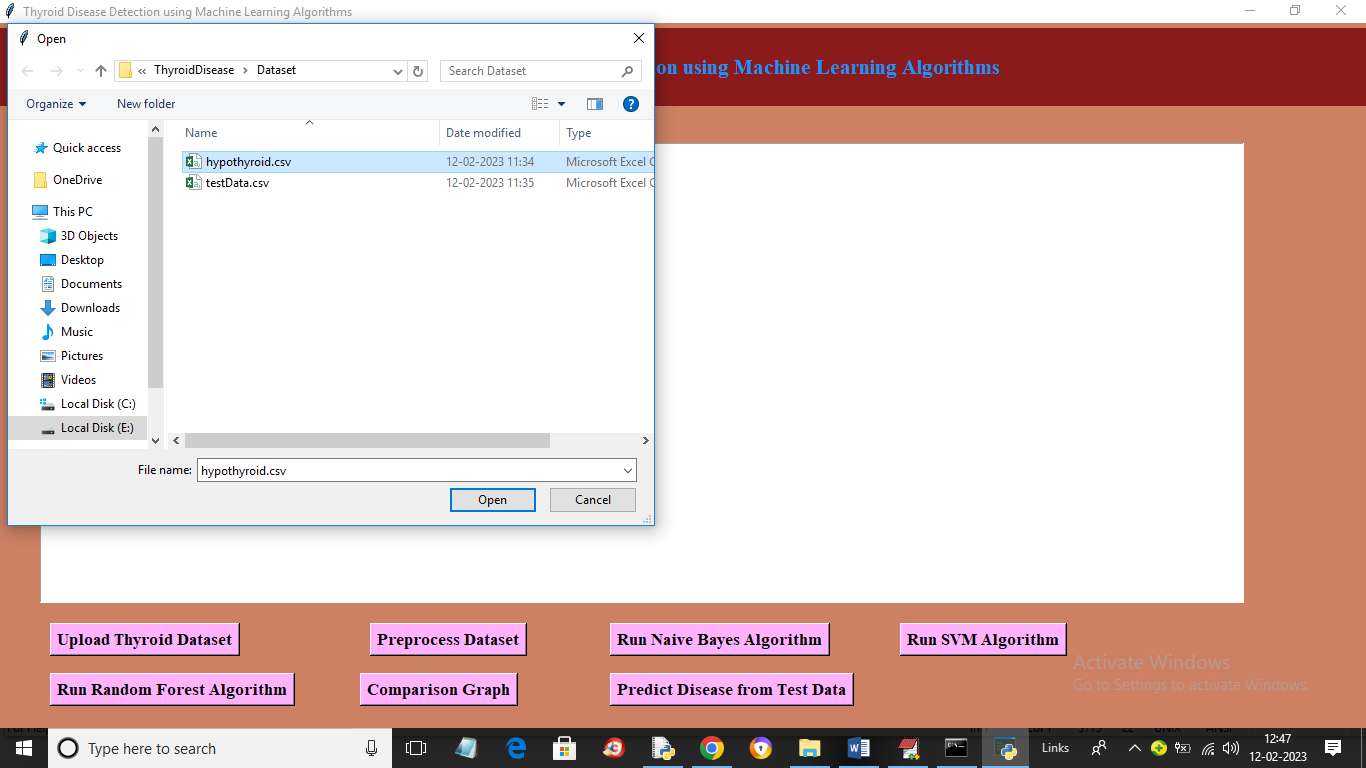
1. Upload Thyroid Dataset: using this module we will upload dataset details and then application will read and display dataset values and then find and plot graph of normal and thyroid patients count
2. Preprocess Dataset: using this module we will preprocess dataset and then remove missing values and then convert all non-numeric data into numeric data and then shuffle and then split dataset into train and test where application using 80% dataset for training and 20% for testing
3. Run Naive Bayes Algorithm: using this module we will input 80% dataset to Naïve Bayes to train a model and 20% test data will be applied on trained model to calculate prediction accuracy
4. Run SVM Algorithm: using this module we will input 80% dataset to SVM to train a model and 20% test data will be applied on trained model to calculate prediction accuracy
5. Run Random Forest Algorithm: using this module we will input 80% dataset to Random Forest to train a model and 20% test data will be applied on trained model to calculate prediction accuracy
6. Comparison Graph: using this module we will plot accuracy comparison graph between all algorithms
7. Predict Disease from Test Data: using this module we will upload test data and then random forest will predict weather test data is normal or contains thyroid disease

SCREEN SHOTS

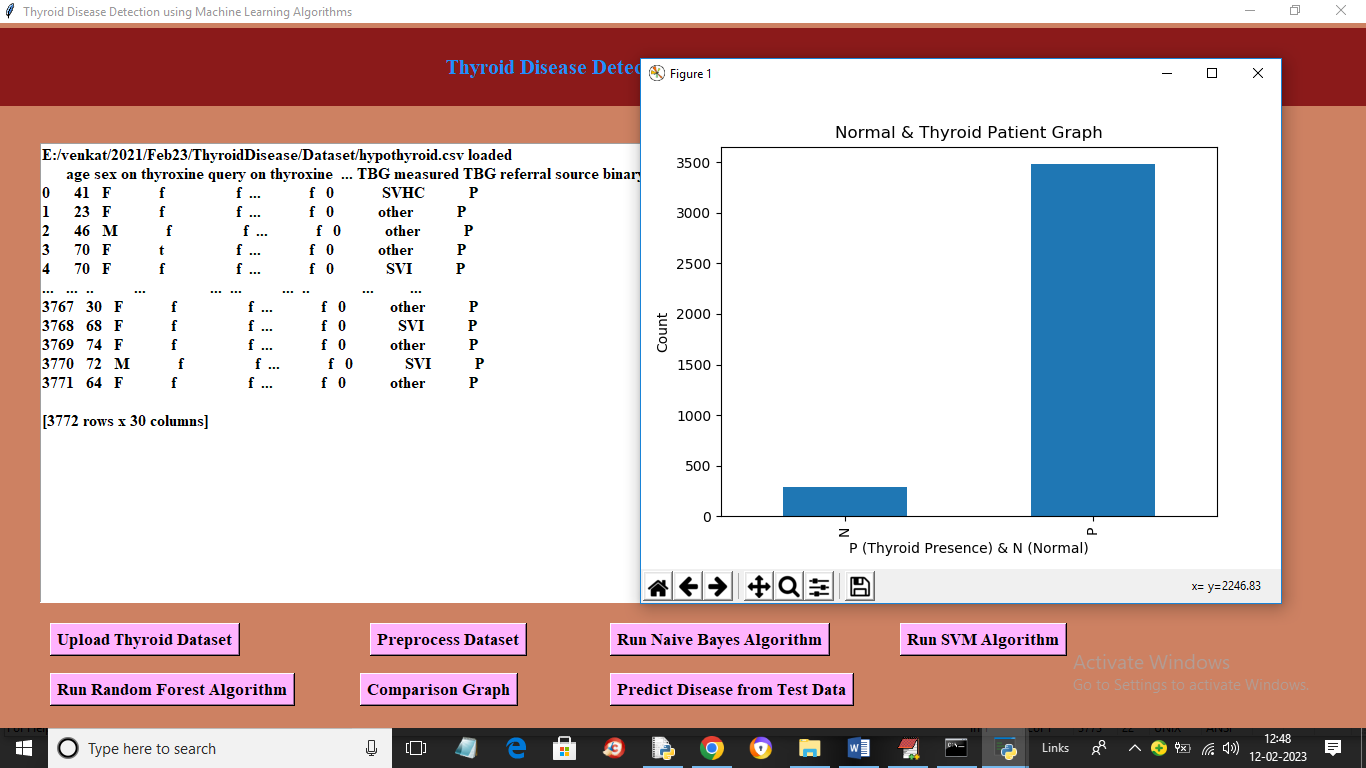
To run project double click on ‘run.bat’ file to get below screen



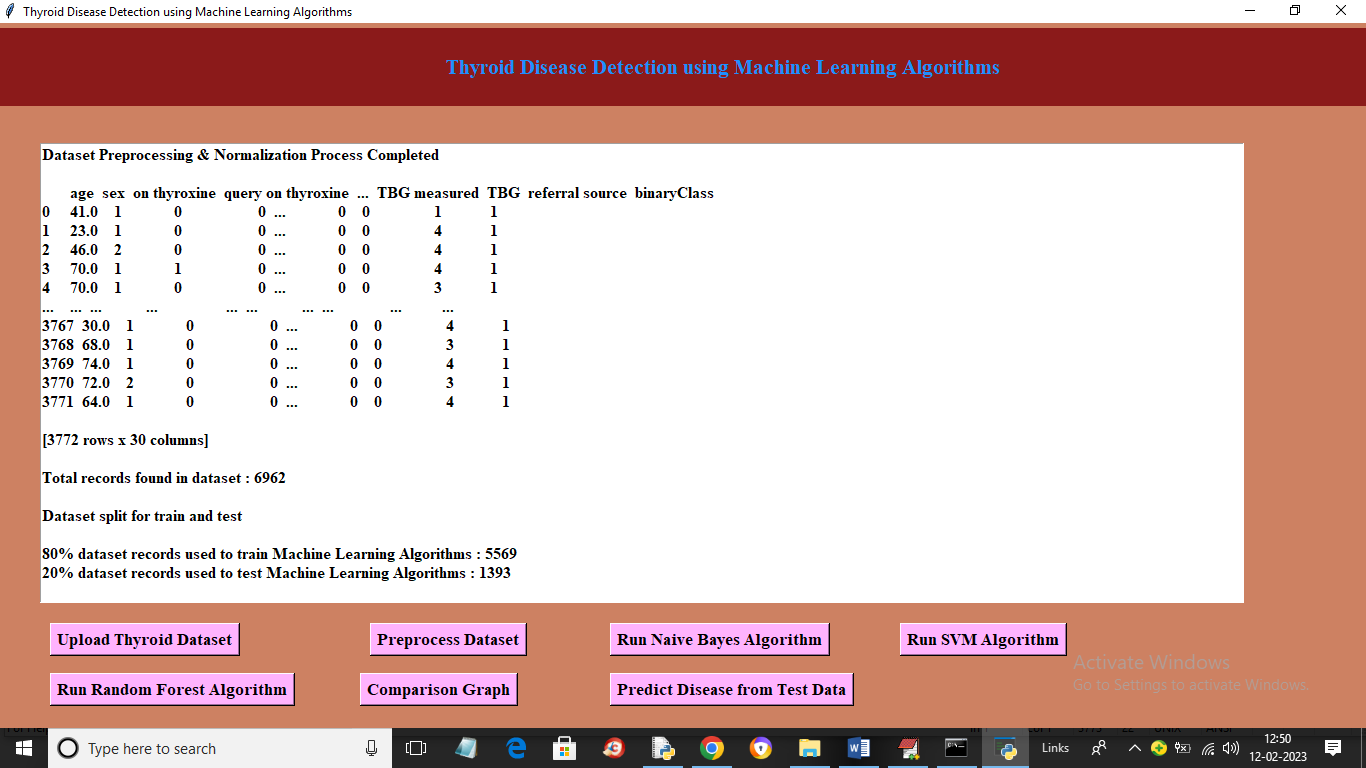
In above screen click on ‘Upload Thyroid Dataset’ button to load dataset and get below output



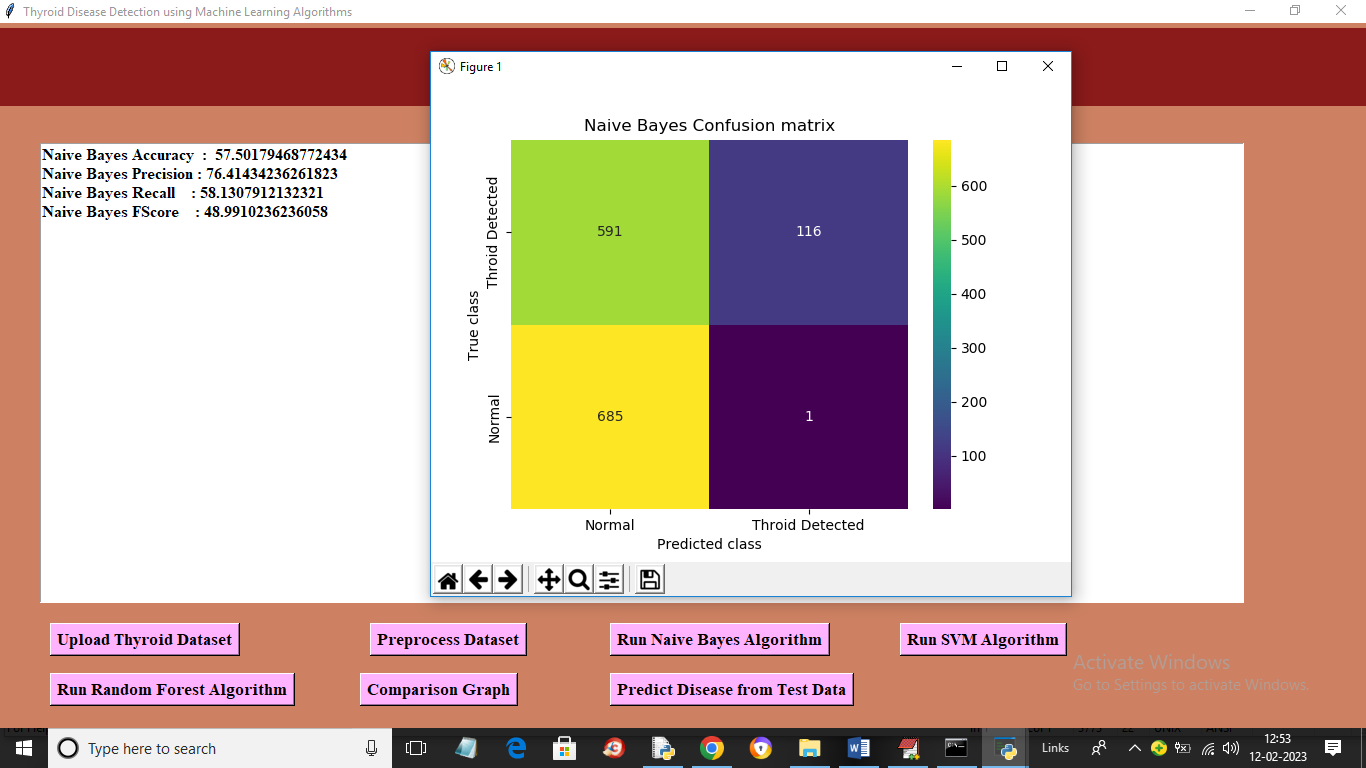
In above screen selecting and uploading thyroid dataset and then click on ‘Open’ button to load dataset and get below output



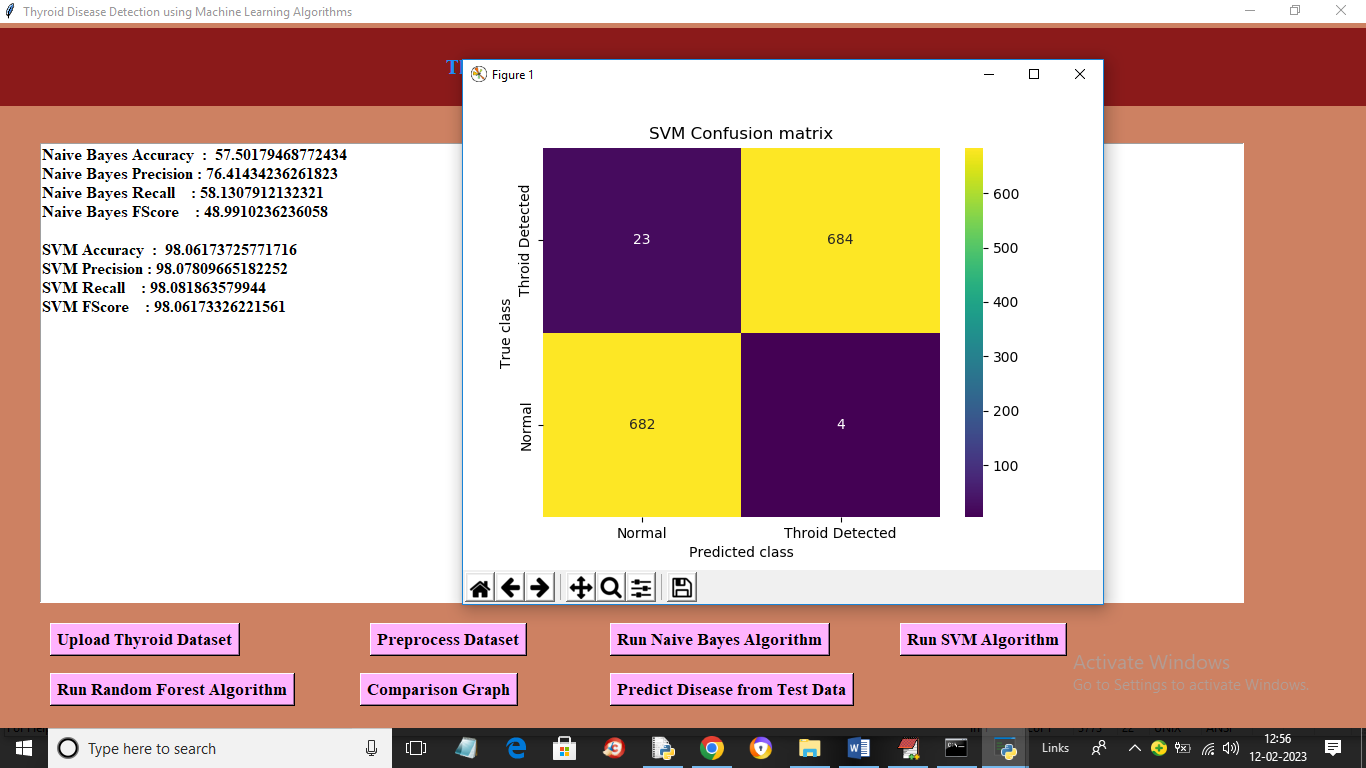
In above screen dataset loaded and in graph x-axis represents N (normal) and P (thyroid presence) and y-axis represents number of records and in above dataset values we can see some are non-numeric and some are numeric and machine learning algorithms accept only numeric values so we need to process dataset to encode non-numeric values to numeric values so click on ‘Preprocess Dataset’ button to get below output



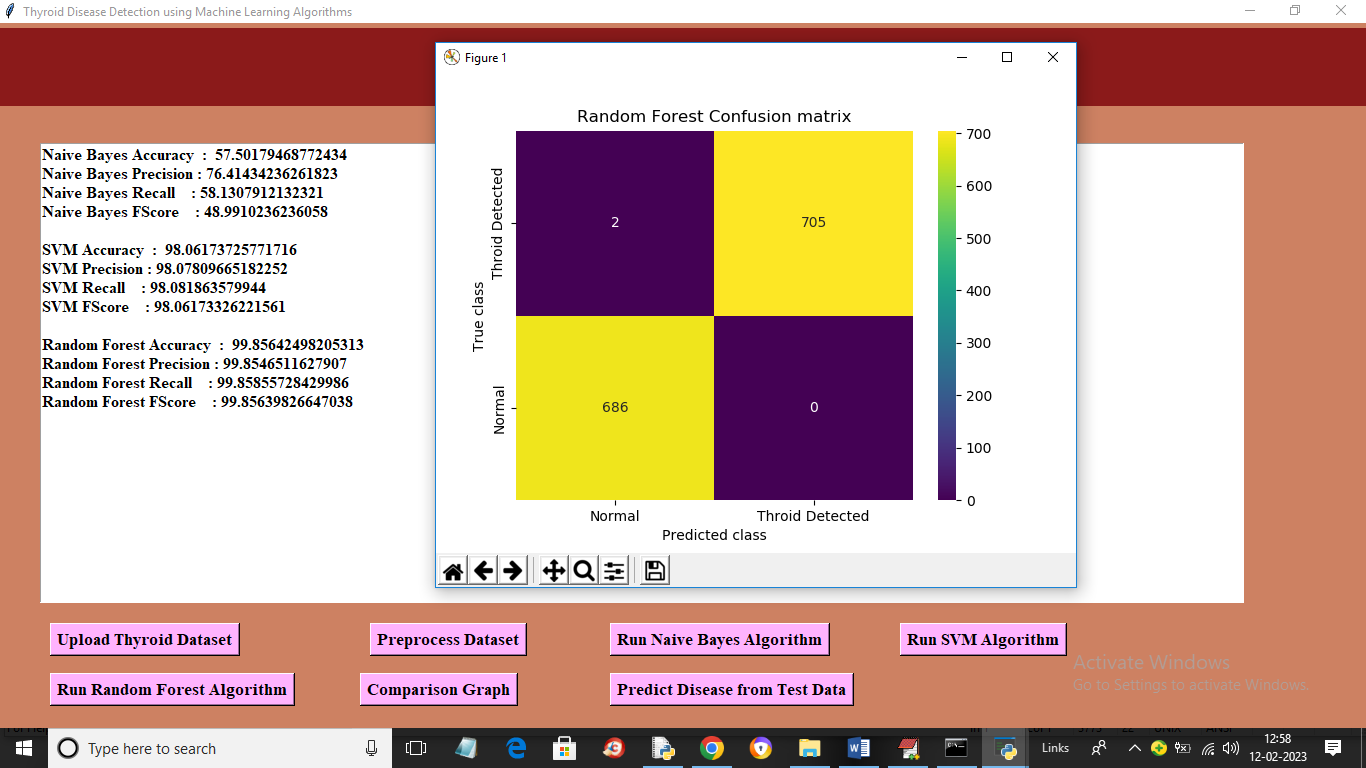
In above screen we can see all values are converted to numeric format and then we can see dataset contains 6962 records where application using 80% records (5569) for training and 1393 (20%) records for testing. Now click on ‘Run Naïve Bayes Algorithm’ button to train Naïve Bayes on 80% dataset and test on 20% data to get below prediction accuracy



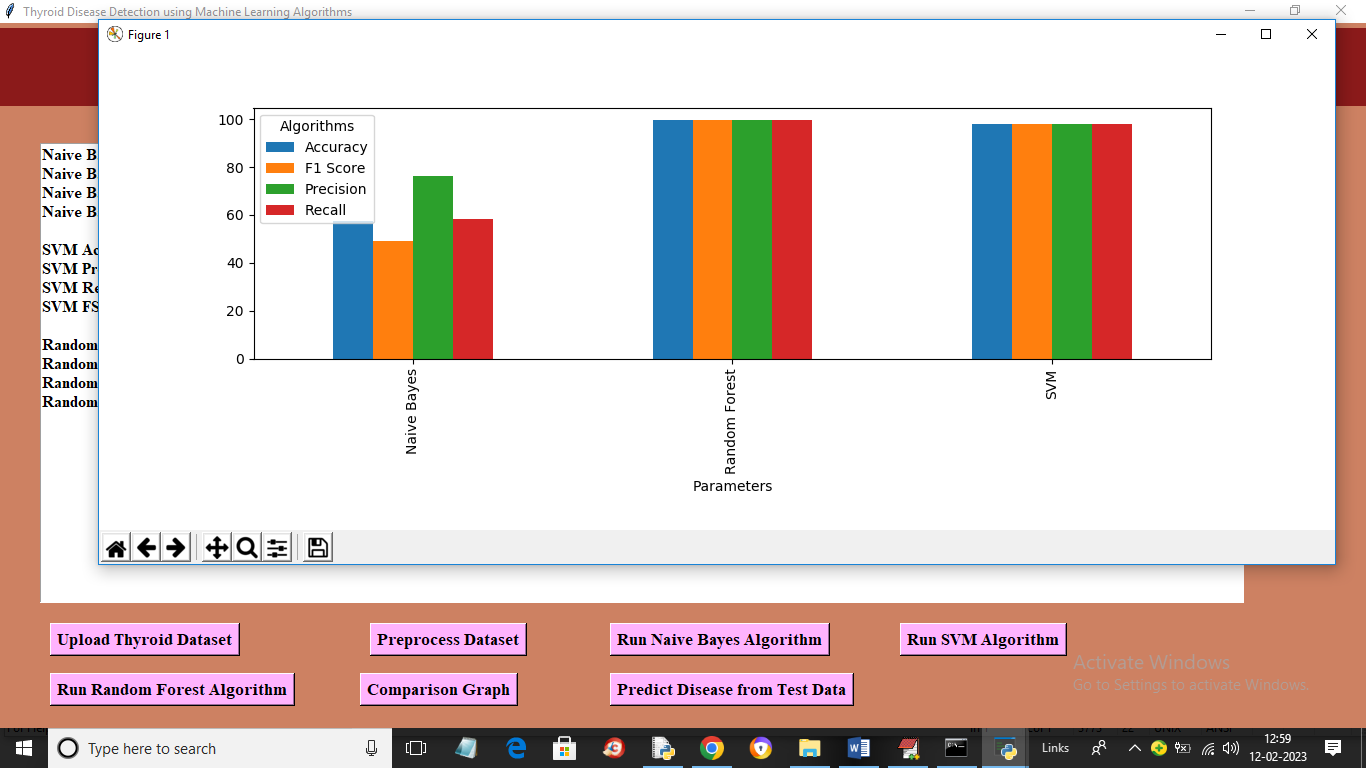
In above screen with Naïve Bayes we got 57% accuracy and we can see other metrics like precision, recall and FSCORE. In confusion matric graph x-axis represents Predicted Labels and y-axis represents True Labels and yellow and light blue colour in diagnol represents correct prediction and dark blue and green box contains incorrect prediction count. Now close above graph and then click on ‘Run SVM Algorithm’ button to get below output



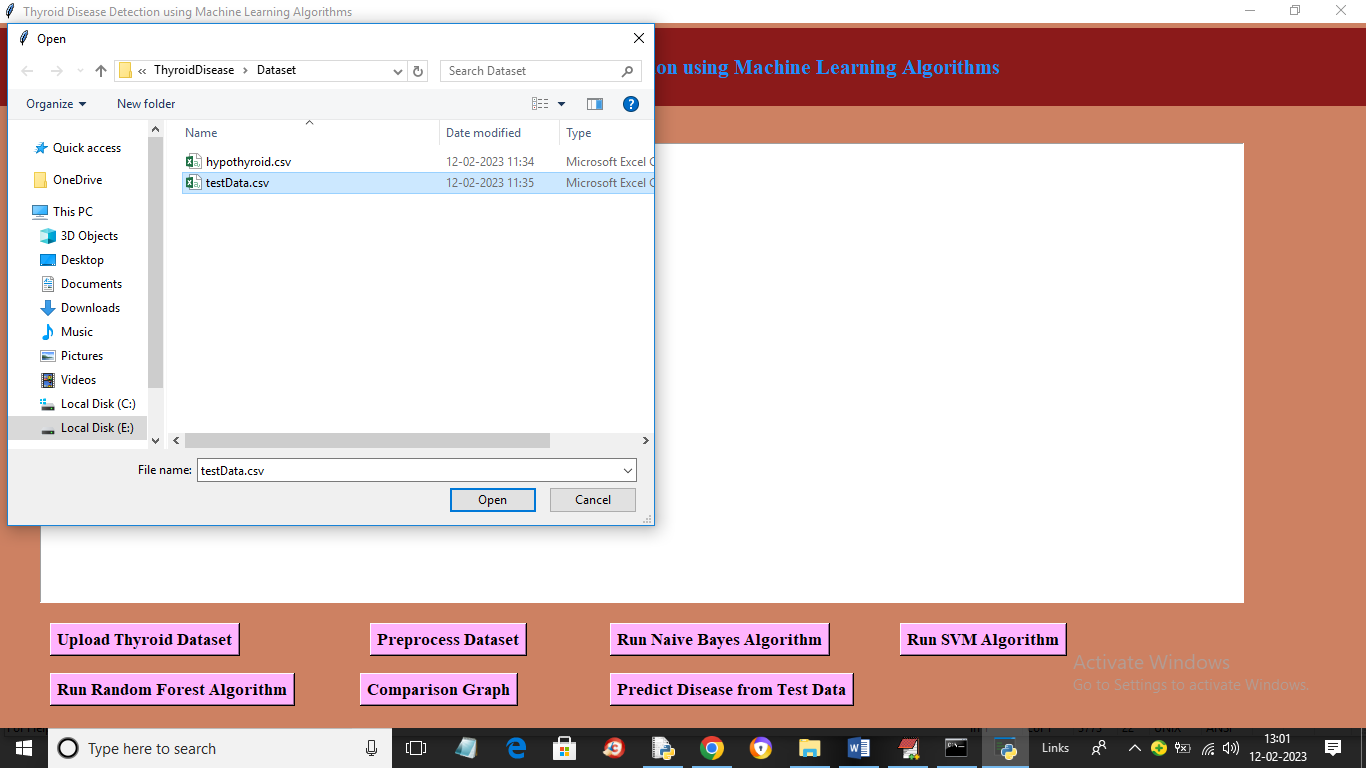
In above screen with SVM we got 98% accuracy and in confusion matrix graph yellow boxes contains correct prediction count and blue boxes contains incorrect prediction count. Now close above graph and then click on ‘Run Random Forest Algorithm’ button to get below output



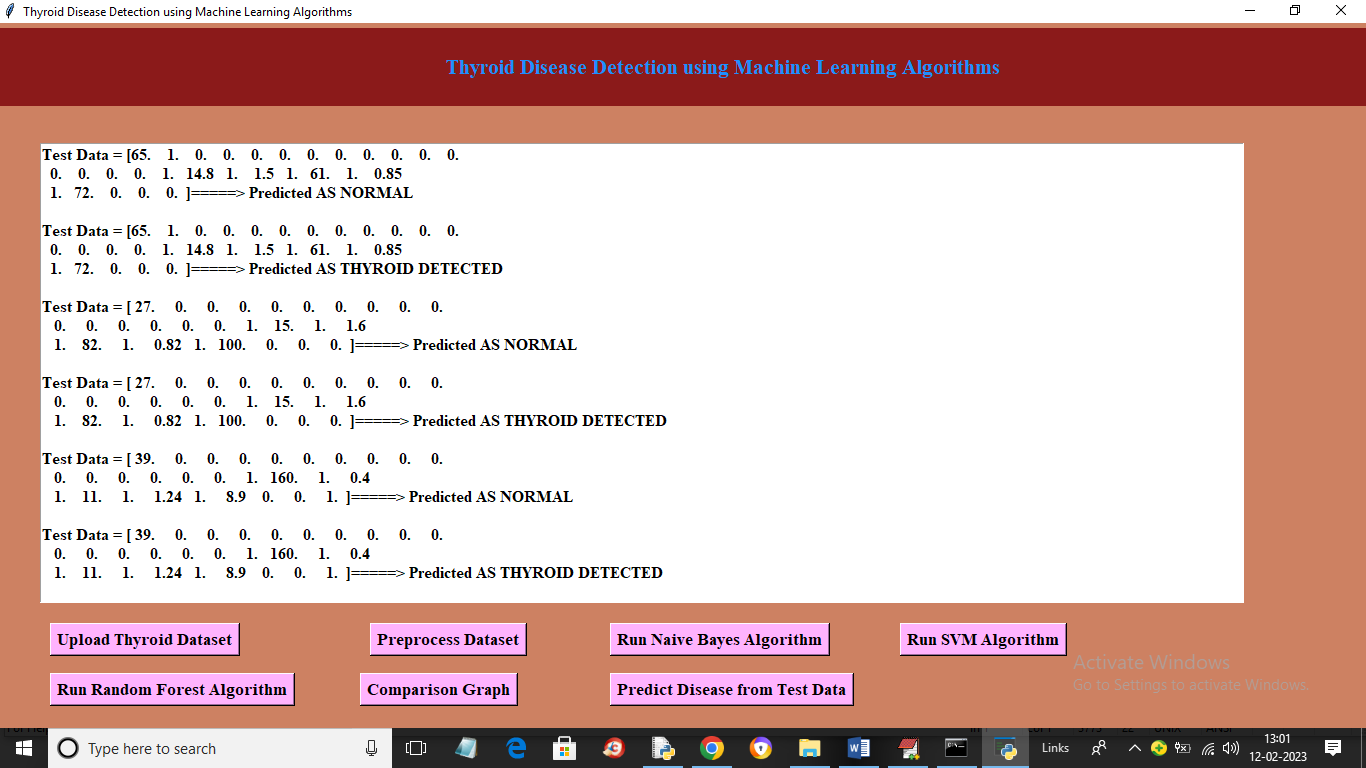
In above screen with Random Forest we got 99% accuracy and now click on ‘Comparison Graph’ button to get below graph



In above graph x-axis represents algorithm names and y-axis represents metric values like accuracy, precision, recall in different bar colour and in all algorithms Random Forest got high accuracy. Now close above graph and then click on ‘Predict Disease from Test Data’ to upload test data and get prediction output



In above screen selecting and uploading testData.csv file and then click on ‘Open’ button to get below output



In above screen in square bracket we can see test data and after arrow symbol we can see predicted output