VIT
Vellore Institute of Technology

A project on

"Disease Prediction using Machine Learning"

Submitted with the requirements for the degree of

M.TECH INTEGRATED (SOFTWARE ENGINEERING)

By

B DEVI PRASAD – 19MIS1018
VIGNESH KUMAR REDDY – 19MIS1085
M SAI BENARJY – 19MIS1121

FOR THE SUBJECT

ESSENTIALS OF DATA ANALYTICS (CSE3506)

SLOT

G2

Under the guidance of

Dr. RAJALAKSHMI R

ACKNOWLEDGEMENT

We wish to express our sincere thanks and deep sense of gratitude to our project guide, **Dr. R. Rajalakshmi,** School of Computer Science and Engineering for her consistent encouragement and valuable guidance offered to us throughout the course of the project work.

We are extremely grateful to **Dr. R. Ganesan, Dean,** School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Chennai, for extending the facilities of the School towards our project and for his unstinting support.

We express our thanks to our **Head of the Department** for his support throughout the course of this project.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the courses.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

19MIS1018 – B DEVI PRASAD
19MIS1085 – VIGNESH KUMAR REDDY
19MIS1121 – M SAI BENARJY

BONAFIDE CERTIFICATE

Certified that this project report entitled **Disease Prediction using Machine Learning** is a bona-fide work of **B DEVI PRASAD 19MIS1018, VIGNESH KUMAR REDDY 19MIS1085, M SAI BENARJY 19MIS1121** carried out the "J"-Project work under my supervision and guidance for **CSE3506 – ESSENTIALS OF DATA ANALYTICS**.

Dr. R. Rajalakshmi

SCOPE

TABLE OF CONTENTS

Abstract:

Our Disease Prediction system predicts the disease of the user on the basis of symptoms provided by the user, the symptoms are given as an input to the system. The system analyses the symptoms provided by the user as input and gives the probable disease as the output. Disease Prediction is done by implementing 4 Classifiers. Our model uses 4 classifier algorithms and takes the mode of the 4 algorithms to provide us the optimal solution.

Keywords: Classifiers, Machine Learning, Disease.

Introduction:

It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Various techniques in data mining and neural networks have been employed to find out the severity of heart disease among humans. The severity of the disease is classified based on various methods like Support vector classifier, Gaussian Naive Bayes Classifier, Random Forest Classifier, K-Nearest Neighbours, Decision tree and Logistic Regression. The nature of heart disease is complex and hence, the disease must be handled carefully. Not doing so may affect the heart or cause premature death. The perspective of medical science and data mining are used for discovering various sorts of metabolic syndromes. Data mining with classification plays a significant role in the prediction of heart disease and data investigation. Proposed system for disease prediction. The doctor may not be available always when needed. But, in the modern time scenario, according to necessity one can always use this prediction system anytime. The symptoms of an individual along with the age and gender can be given to the ML model to further process. After preliminary processing of the data, the ML model uses the current input, trains and tests the algorithm resulting in the predicted disease.

Dataset Description:

The hospital data will be in the form of textual format or in the structural format. The dataset used in this project is real-life data. The structural data contains symptoms of patients while unstructured data consist of textual format. The dataset used is contains real-life hospital data, and data stored in data center. The data provided by the hospital contains symptoms of the patients.

Complete Dataset consists of 2 CSV files. One of them is training and other is for testing your model. Each CSV file has 133 columns. 132 of these columns are symptoms that a person experiences and last column is the prognosis. These symptoms are mapped to 42 diseases you can classify these set of symptoms to. The dataset used is contains real-life hospital data, and data stored in data centre. The data provided by the hospital contains symptoms of the patients.

Related work:

1. “Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare Ayman Mir, Sudhir N. Dhage –IEEE@2020”.

These research hopes to recommend the best algorithm based on efficient performance result for the prediction of diabetes disease. Experimental results of each algorithm used on the dataset will be evaluated. The four classifiers have been compared based on training time, testing time and accuracy value. The overall performance of Support Vector machine to predict the diabetes disease is better than Naive Bayes, Random Forest and Simple Cart.

2. “Latest trends on heart disease prediction using machine learning and image fusion Manoj Diwakar a, Amrendra Tripathi, Kapil Joshi c, Minakshi Memoria c”.

A review of the classification methods for machine learning and image fusion that have been demonstrated to help healthcare professionals identify heart disease. An appropriate algorithm must be used to develop a predictive model that delivers accurate results. We observe that ANN has good effects for predicting heart disease in most models. Finally, the use of machine learning and image fusion to detect heart disease is an essential activity, and it can be of assistance to both healthcare authorities and patients.

3. “Multiple disease prediction using Machine learning algorithms K. Arumugam , Mohd Naved, Priyanka P. Shinde , Orlando Leiva-Chauca,Antonio Huaman-Osorio, Tatiana Gonzales-Yanacd”

For prediction of diseases, different machine learning algorithms are used to ensure quick and accurate predictions. We find-tuned the decision tree model for optimum performance in forecasting the chance of heart disease in diabetic patients since it consistently outperformed the naive Bayes and support vector machine models.

4. “ Disease Prediction Using Machine Learning Over Big Data Vinitha S, Sweetlin S, Vinusha H and Sajini S”.

Due to big data progress in biomedical and healthcare communities, accurate study of medical data benefits early disease recognition, patient care and community services. In this paper, it bid a Machine learning Decision tree map algorithm by using structured and unstructured data from hospital. It also uses Map Reduce algorithm for partitioning the data. the scheming accuracy of our proposed algorithm reaches 94.8% with an regular speed.

5. “Disease Prediction by Machine Learning Over Big Data From Healthcare Communities MIN CHEN (Senior Member, IEEE), YIXUE HAO, KAI HWANG (Life Fellow, IEEE),LU WANG, AND LIN WANG”.

The big data growth in biomedical and healthcare communities, accurate analysis of medical data benefits early disease detection, patient care, and community services. In this paper, they proposed a new convolutional neural network based multimodal disease risk prediction (CNN-MDRP) algorithm using structured and unstructured data from hospital, Compared to several typical prediction algorithms, the prediction accuracy of our proposed algorithm reaches 94.8% with a convergence speed which is faster than that of the CNN-based unimodal disease risk prediction(CNN-UDRP) algorithm.

6. “Development of Big Data Predictive Analytics Model for Disease Prediction using Machine learning Technique R. Venkatesh & C. Balasubramanian & M. Kaliappan”.

Now days, health prediction in modern life becomes very much essential. Big data analysis plays a crucial role to predict future status of health and offers preeminent health outcome to people. They designed a map reduce algorithm for Naive Bayes technique that integrated with Apache Spark framework for performing big data predictive analytics that helps to reduce the computation complexity because of its parallelism. The proposed BPA-NB scheme classify and predict the future status from the heart disease data set effectively that discussed in result and discussion section.

7. “Application of Machine Learning in Disease Prediction Pahulpreet Singh Kohli & Shriya Arora”.

Applied different classification algorithms, each with its own advantage on three separate databases of disease (Heart, Breast cancer, Diabetes) available in UCI repository for disease prediction. The feature selection for each dataset was accomplished by backward modeling using the p-value test. The application of machine learning in the field of medical diagnosis is increasing gradually. The results of this study confirm the application of machine learning algorithms in prediction and early detection of diseases. To our best understanding, the model built according to the proposed method exhibits better accuracy than the existing ones [13,14,16,17,19,20]. The prediction accuracy of our proposed method reaches 87.1% in Heart Disease detection using Logistic Regression, 85.71% in Diabetes prediction using Support Vector Machine (linear kernel) and 98.57% using AdaBoost classifier for Breast Cancer detection.

8. “Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis Prof. Dhomse Kanchan B & Mr. Mahale Kishor M.”

The worldwide study on causes of death due to heart disease/syndrome has been observed that it is the major cause of death. If recent trends are allowed to continue, 23.6 million people will die from heart disease in coming 2030. The healthcare industry collects large amounts of heart disease data which unfortunately are not “mined” to discover hidden information for effective decision making. In this paper, study of PCA has been done which finds the minimum number of attributes required to enhance the precision of various supervised machine learning algorithms. The purpose of this research is to study supervised machine learning algorithms to predict heart disease.

9. “ Disease prediction from various symptoms using machine learning Rinkal Keniya, Aman Khakharia ,Vruddhi Shah ,Vrushabh Gada ,Ruchi Manjalkar ,Tirth Thaker ,Mahesh Warang , Ninad Mehendale”.

Accurate and on-time analysis of any health related problem is important for the prevention and treatment of the illness. The traditional way of diagnosis may not be sufficient in the case of a serious ailment. Developing a medical diagnosis system based on machine learning (ML) algorithms for prediction of any disease can help in a more accurate diagnosis than the conventional method. Based on the symptoms, age, and gender of an individual, the diagnosis system gives the output as the disease that the individual might be suffering from. The weighted KNN algorithm gave the best results as compared to the other algorithms. The accuracy of the weighted KNN algorithm for the prediction was 93.5 %. This diagnosis model can act as a doctor for the early diagnosis of a disease to ensure the treatment can take place on time and lives can be saved.

10. “Disease Prediction using Machine Learning Algorithms Sneha Grampurohit & Chetan Sagarnal”.

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. This research work carried out demonstrates the disease prediction system developed using Machine learning algorithms such as Decision Tree classifier, Random forest classifier, and Naïve Bayes classifier. The paper presents the comparative study of the results of the above algorithms used.

Motivation:

There is a need to study and make a system which will make it easy for an end user to predict the chronic diseases without visiting physician or doctor for diagnosis. To detect the Various Diseases through the examining symptoms of patient's using different techniques of Machine Learning Models. To Handle text data and Structured data is no Proper method. The Proposed system will consider both structure and unstructured data. The Predictions Accuracy will Increase using Machine Learning.

Proposed Methodology:

ML process starts from a pre-processing data phase followed by feature selection based on Probability in each algorithm, classification of modelling performance evaluation, and the results with improved accuracy. The feature selection and modelling keep on repeating for various combinations of attributes.

- preprocessing of the dataset
- then we discuss briefly with the proposed methodology,
- Identifying the main features or variables which is suitable or the project.
- followed by model comparison and selection of best model
- finally they summarize the main finding and
- We find out the highest accuracy among the algorithms and we select the algorithm and then started separating the data and label and then

And then we find the accuracy on test data then finally we will build a predictive system.

- We take our dataset and split it in the ratio of 80 and 20 for training and testing respectively
- And we label encode "prognosis" column which contains our disease names which will be our target area
- And then we have trained, predicted and analysed the ML models using these algorithms:

SVM

KNN

RANDOM FOREST CLASSIFIER

Gaussian Naive Bayes Classifier

Logistic Regression

Decision Tree

At last we take the MODE of these 6 algorithms to give our answer

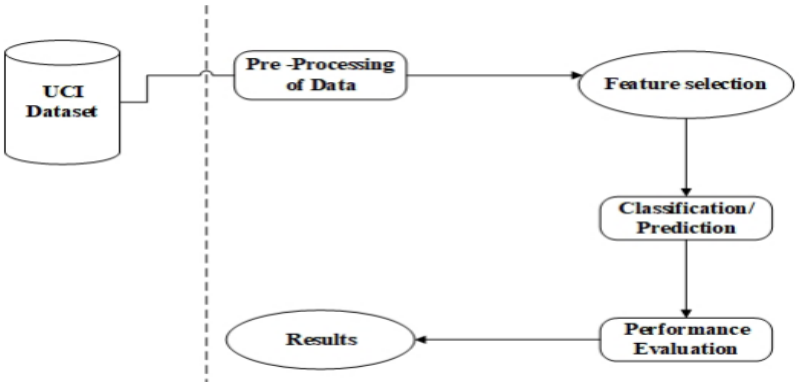


Fig. 1

Algorithms Used:

1. Naïve Bayes
2. Support Vector Classification
3. K-Nearest Neighbour
4. Decision Tree
5. Random Forest

Results and Discussion:

We have used the all classification Machine learning algorithms.
#Load the necessary libraries for data analysis and visualization:
library(tidyverse) # for data manipulation and visualization
library(ggcorrplot) # for visualizing correlation matrix

- First we did the visualization

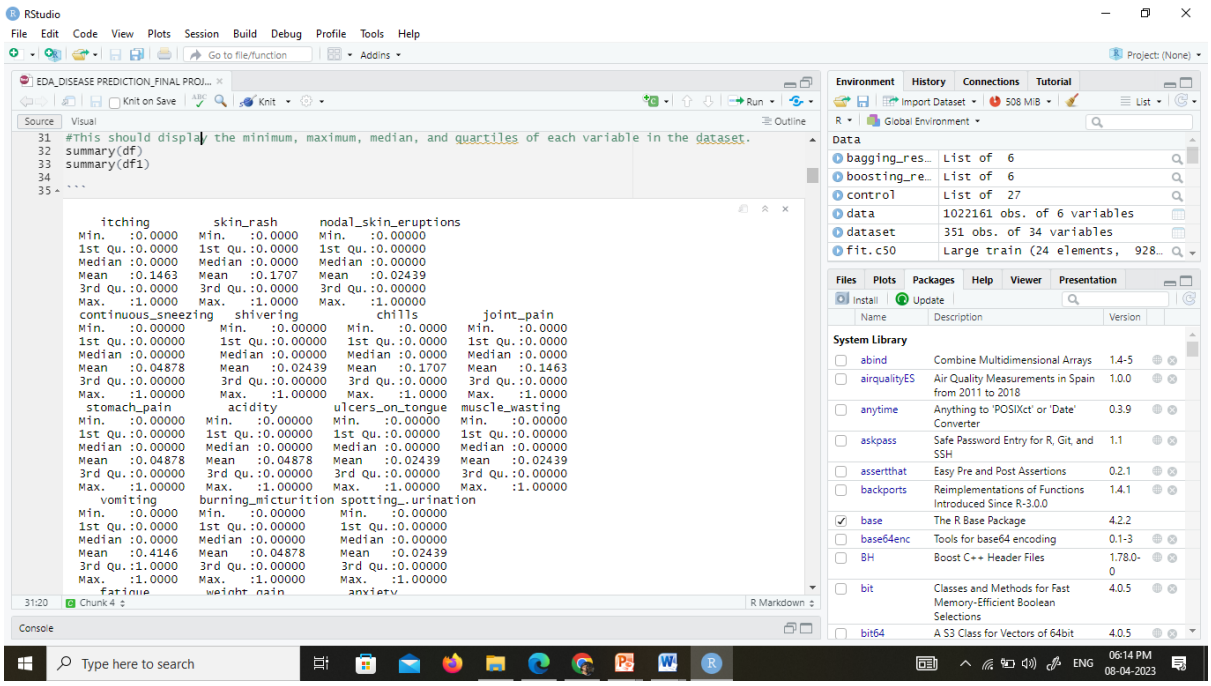


Fig.2 Summary

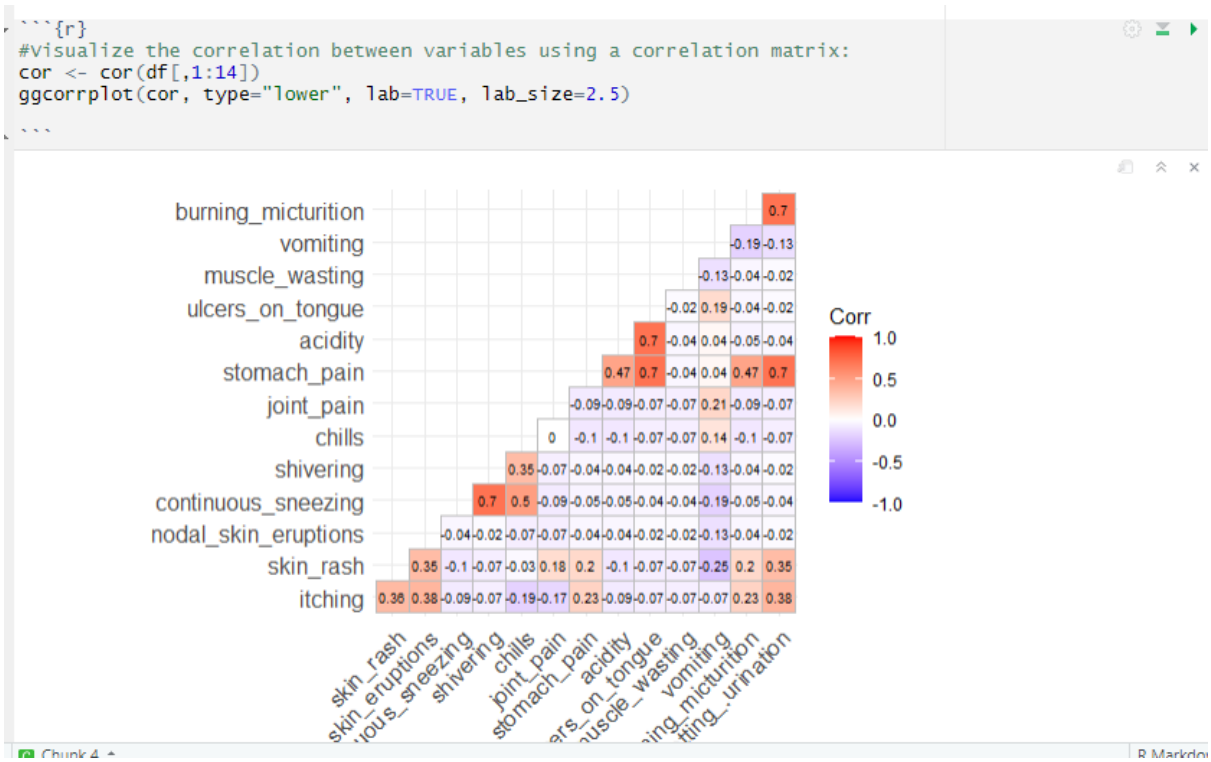


Fig.3 Confusion Matrix

```
ggplot(df, aes(x=itching, y=skin_rash)) +
  geom_point(color="blue", alpha=0.5) +
  labs(title="itching vs skin_rash", x="itching", y="Skin_rash")
```

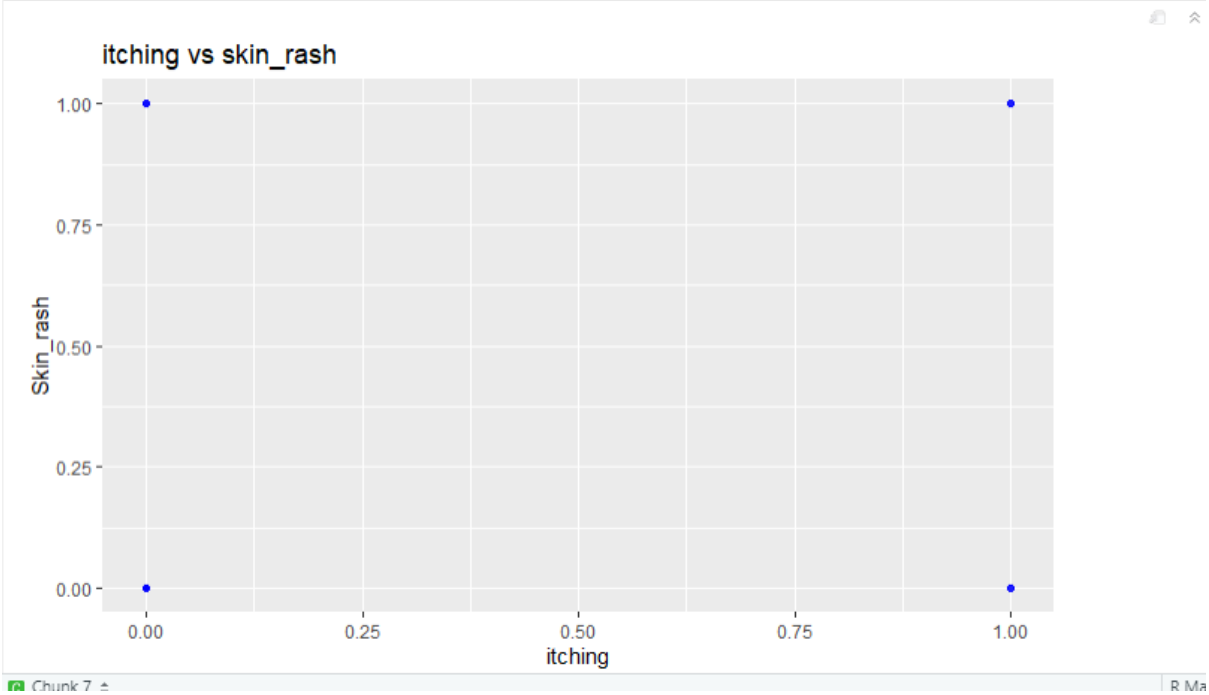


Fig.4 Geometric Point

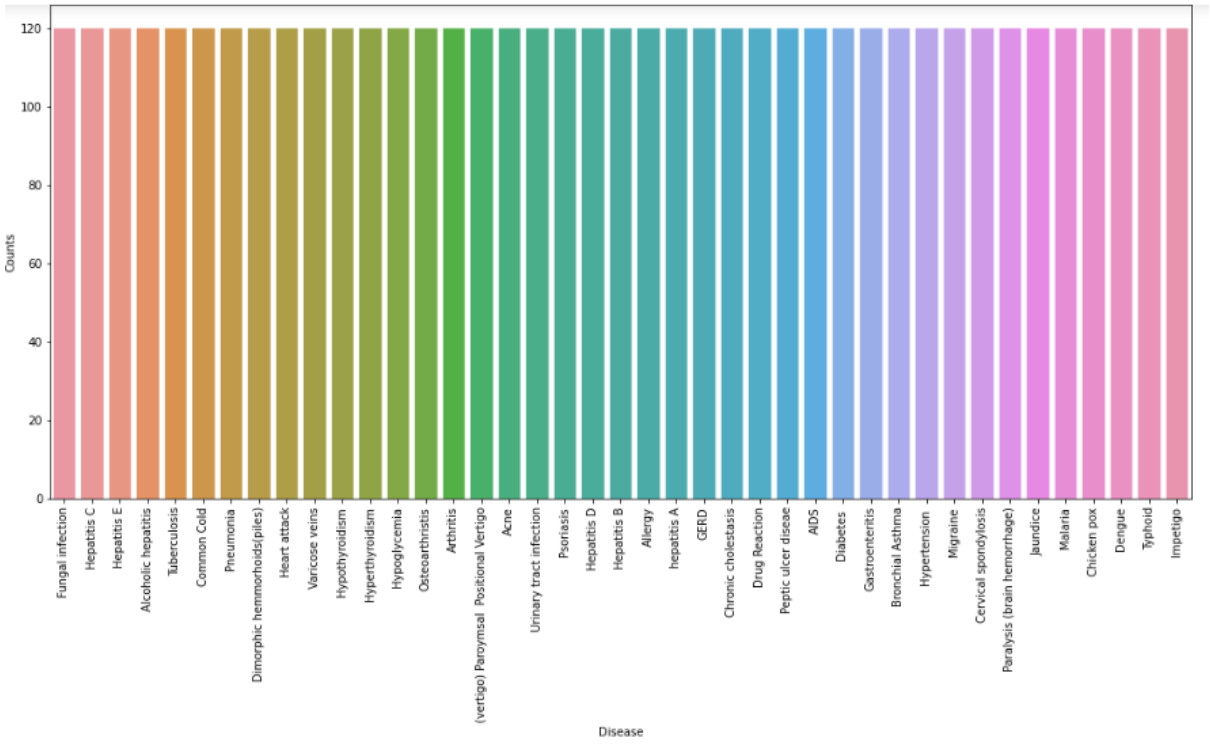
Splitting The Data For Training And Testing The Model:

We have split the data into 80:20 format i.e. 80% of the dataset will be used for training the model and 20% of the data will be used to evaluate the performance of the model.

Steps for the algorithm implementation and model processing:

- Load the required libraries
- Load the training and testing data sets
- Pre-process the data sets
- We can now train the algorithm on the training set using the function from the suitable package.
- We can then use the trained model to predict the outcomes of the validation set using the predict() function.
- We can evaluate the accuracy of the model by comparing the predicted outcomes to the actual outcomes in the validation set.
- Finally, we can use the trained model to predict the outcomes of the testing set and evaluate the accuracy on this set as well.

Prognosis Disease:



Classification Algorithms:

1. Naïve Bayes:

```
#Finally, we can use the trained model to predict the outcomes of the testing set and evaluate the
accuracy on this set as well.
test_pred <- predict(nb_model, newdata = test_data[,-133])
test_accuracy <- mean(test_pred == test_data$prognosis)
print(paste0("Test Accuracy: ", test_accuracy))


```

```
[1] "Accuracy: 1"
[1] "Test Accuracy: 1"
```

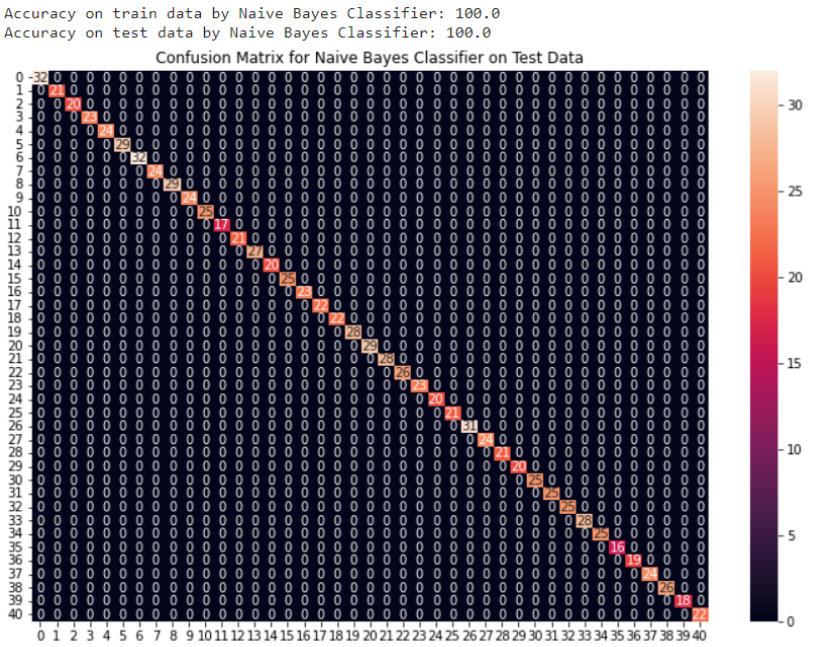


Fig.5 Naïve Bayes Results

2. Support Vector Machine:

```
# Make predictions using the trained SVM model
predictions <- predict(svm_model, test_X)

# Calculate accuracy and precision of predictions
accuracy <- confusionMatrix(predictions, test_Y)$overall["Accuracy"]
precision <- confusionMatrix(predictions, test_Y)$byClass["Precision"]
```

```
                 Accuracy : 1
                   95% CI : (0.914, 1)
    No Information Rate : 0.0244
    P-Value [Acc > NIR] : < 2.2e-16

                    Kappa : 1
```

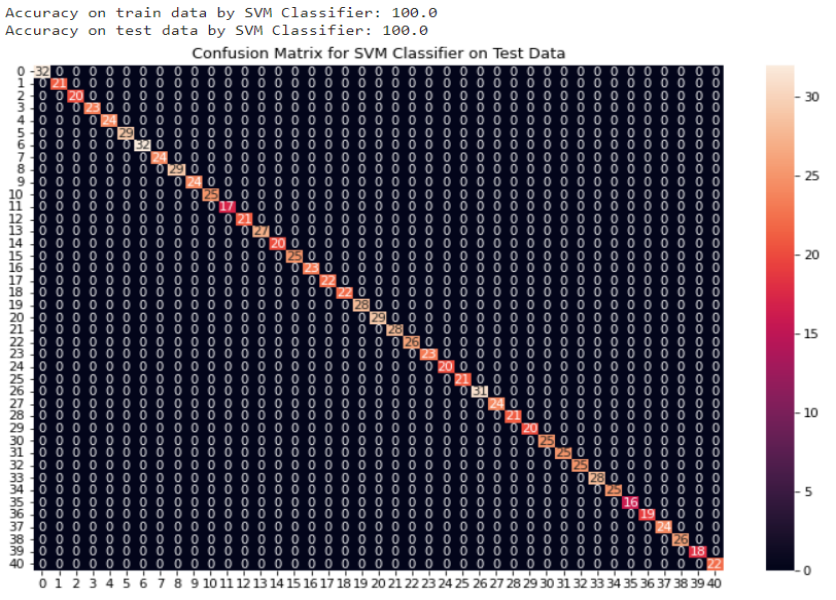


Fig. 6 Support vector Machine

3. K-Nearest Neighbour

```
# Train the K-Nearest Neighbors model
knn_model <- knn(train_x, test_x, train_y, k=5)

# Evaluate the model accuracy
accuracy <- mean(knn_model == test_y)
print(paste0("Accuracy: ", round(accuracy * 100, 2), "%"))


```

```
[1] "Accuracy: 100%"
```

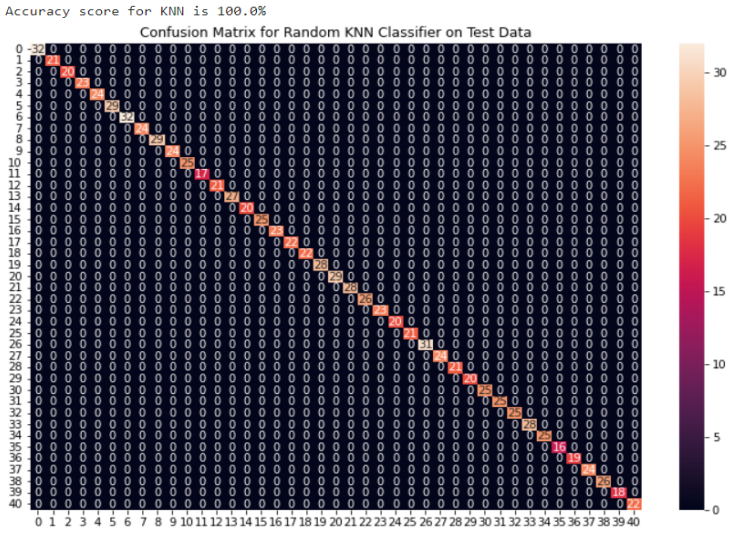
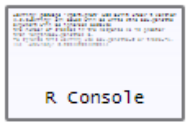


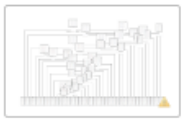
Fig.7 KNN classifier

4. Decision Tree:

```
# Evaluate the accuracy of the model
accuracy <- sum(predictions == testing_data$prognosis) / nrow(testing_data)
print(paste("Accuracy:", accuracy))

```





```
Warning: package 'rpart.plot' was built under R version 4.2.3Warning: All boxes will be
white (the box.palette argument will be ignored) because
the number of classes in the response 41 is greater than length(box.palette) 6.
To silence this warning use box.palette=0 or trace=-1.[1] "Accuracy: 0.902439024390244"
```

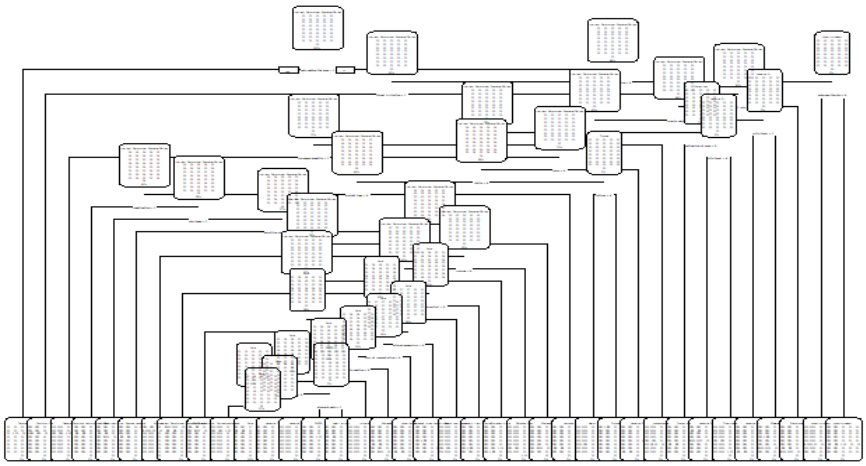


Fig.8 Decision Tree classifier

5. Random Forest:

```
# Evaluate Model Performance
confusion_matrix <- table(predictions, testing_data$prognosis)
accuracy <- sum(diag(confusion_matrix))/sum(confusion_matrix)
precision <- diag(confusion_matrix)/rowSums(confusion_matrix)
recall <- diag(confusion_matrix)/colSums(confusion_matrix)
```

```
[1] "Accuracy: 1"
 [1] "Precision: 1"
```

```
Accuracy on train data by Random Forest Classifier: 100.0
Accuracy on test data by Random Forest Classifier: 100.0
```

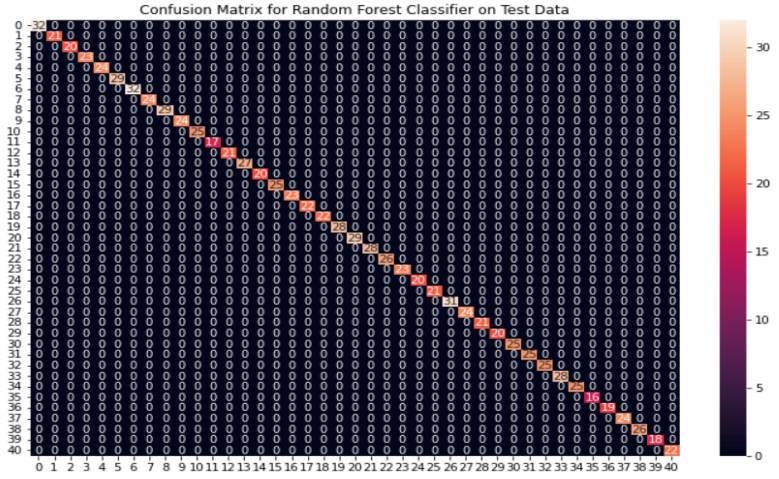


Fig.9 Random Forest

Predict the Results:

```
# Testing the function
print(predictDisease("Itching,Skin Rash,Nodal Skin Eruptions"))
```

```
{'rf_model_prediction': 'Fungal infection', 'naive_bayes_prediction': 'Fungal infection', 'svm_model_
prediction': 'Fungal infection', 'final_prediction': 'Fungal infection'}
```

Fig.10 Prediction

If we give the functions it will predict the disease

Table Experimental Results:

Sl.No	Algorithm	Accuracy	Test Accuracy	F1-score	Precision
1.	Naïve Bayes	100	100	0.65	100
2.	Support Vector Machine	100	100	0.863	100
3.	K-NN	100	100	0.667	100
4.	Decision Tree	0.9024	0.9130	0.53	0.90
5.	Random Forest	100	100	0.83	100

Discussion:

Both of those metrics take class predictions as input so you will have to adjust the threshold regardless of which one you choose. Remember that the F1 score is balancing precision and recall on the positive class while accuracy looks at correctly classified observations both positive and negative.

The highest possible value of an F-score is 1.0, indicating perfect precision and recall, and the lowest possible value is 0, if either precision or recall are zero.

Therefore, accuracy does not have to be greater than F1 score. Because the F1 score is the harmonic mean of precision and recall, intuition can be somewhat difficult.

F1 score	Interpretation
> 0.9	Very good

Here, the best algorithm is **SUPPORT VECTOR MACHINE**.

Comparative Study:

ML process starts from a pre-processing data phase followed by feature selection based on Probability in each algorithm , classification of modeling performance evaluation, and the results with improved accuracy. The feature selection and modeling keep on repeating for various combinations of attributes.

Conclusion:

The main aim of this disease prediction system is to predict the disease on the basis of the symptoms. This system takes the symptoms of the user from which he or she suffers as input and generates final output as a prediction of disease. Average prediction accuracy probability of 100% is obtained.

This project aims to predict the disease on the basis of the symptoms. The project is designed in such a way that the system takes symptoms from the user as input and produces output i.e., predict disease. In conclusion, for disease risk modelling, the accuracy of risk prediction depends on the diversity feature of the hospital data.

References:

[1] Pingale, Kedar, et al. "Disease prediction using machine learning." International Research Journal of Engineering and Technology (IRJET) 6 (2019): 831-833.

[2] Pingale, K., Surwase, S., Kulkarni, V., Sarage, S., & Karve, A. (2019). Disease prediction using machine learning. International Research Journal of Engineering and Technology (IRJET), 6, 831-833.

[3] Pingale K, Surwase S, Kulkarni V, Sarage S, Karve A. Disease prediction using machine learning. International Research Journal of Engineering and Technology (IRJET). 2019 Dec;6:831-3.

[4] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." IEEE access 7 (2019): 81542-81554.

[5] Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making, 19(1), 1-16.

[6] Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making. 2019 Dec;19(1):1-6.

[7] Dahiwade, D., Patle, G. and Meshram, E., 2019, March. Designing disease prediction model using machine learning approach. In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1211-1215). IEEE.

[8] Patel J, TejalUpadhyay D, Patel S. Heart disease prediction using machine learning and data mining technique. Heart Disease. 2015 Sep;7(1):129-37.

[9] Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2021). Multiple disease prediction using Machine learning algorithms. Materials Today: Proceedings.

[10] Kohli, P. S., & Arora, S. (2018, December). Application of machine learning in disease prediction. In 2018 4th International conference on computing communication and automation (ICCCA) (pp. 1-4). IEEE.
