Disease classification and prediction based on symptoms using Machine Learning

Janhavi Patil[1], Mrs. Varsha Pimprale[2]

[1] UG student

[1,2] Department of Computer Engineering, Cummins College of Engineering for Women, Pune, India

E-Mail: [1]janhavi.c.patil@cumminscollege.in, [2]varsha.pimprale@cumminscollege.in

Abstract

Today machine learning is used widely across all fields of science and technology. One of the most important among them is Health care sector. Health care sector is seeing a drastic and dynamic change post covid-19. Today, the need to visit and consult a doctor is reducing as many hospitals are providing online consultation. Thus, preliminary diagnosis and medication are provided to the patient without physically visiting the doctor. Classification and prediction are supervised machine learning techniques which have wide applications in health care sector such as preliminary diagnosis, quick medication and cost savings. This paper gives insights and results on three machine learning algorithms which classify and predict the disease/prognosis a patient may have based upon the given symptoms. Labelled Dataset is used for classification and prediction of disease based upon symptoms. This dataset is labelled based upon the binary values for each symptom. This system is used by end users like patients who will provide their symptoms and then will receive the most accurate prognosis. This paper also aims to provide a comparative study on 3 different ml classification algorithms based upon their results. The final prediction of the diagnosis is obtained from the cumulative results of 3 algorithms. Hence a more accurate and correct diagnosis is provided to customers in very cost -effective way.

Keywords: Classification, Machine Learning, Disease diagnosis, symptoms, labeled dataset.

1. Introduction

Due to covid, the world has witnessed the need of revival in healthcare systems. Traditional diagnosis and medication approaches can be improvised by incorporating ml techniques. Hence, this introduced the concept of virtual doctor. It means that no need to visit the doctor personally, instead consult doctors online, and get the required prescriptions and medication. Virtual doctors can also be used for early and preliminary diagnosis of symptoms. This is possible by using supervised learning, which is a subbranch of machine learning and artificial intelligence. Dataset in supervised learning is labelled to train the algorithms and test it. Classification and prediction techniques come under the category of supervised learning. However, is noticed that no perfect classifier exists. Hence, many researchers are trying to test and modify different ml algorithms such as K nearest neighbor (KNN), Random Forest classifier (RF), Support Vector machine (SVM), Naïve Bayes, Decision tree algorithm, etc. Many researchers have tried to predict severe diseases such as Parkinson's disease, cancer, tuberculosis, lung infections, heart disease, diabetes and classified them using X ray images, MRI reports, etc. According to previous research, disease classification and prediction is widely performed using deep learning, SVM, Random Forest classifier, fuzzy logic, Convolutional neural networks (CNN), and these are the most precise and efficient algorithms and techniques. The performance of different algorithms depends upon the quality of the dataset used. In this paper, a comparative study of the results (predicted disease/prognosis) using 3 algorithms which are Random Forest Classifier, Support Vector Machine, Gaussian Naïve bayes Classification is explained. All models are trained on the training dataset and then tested on the test dataset. The result of all 3 models is combined to get a more accurate and precise prediction(results).

2. Background

According to previous papers, a similar kind of system has been implemented previously. However, it was implemented as a preliminary level. Predictions were carried out by building models of Naïve Bayes, Decision Tree (DT), KNN, and Logistic regression (LR). Our proposed system implements Gaussian Naïve Bayes, Random Forest, and Support Vector Machine algorithms to predict the disease. Gaussian Naïve Bayes and Random Forest algorithms are advanced and modified versions of Naïve Bayes and Decision Tree algorithms. Moreover, these algorithms have higher accuracy and robust working. Hence, to increase the accuracy and reliability of prediction, the proposed model implements GNB, RF, SVM algorithms.

Citation: [C K Gomathy, “Prediction of disease using machine learning, 2021]

Another paper has done study about using different supervised ml algorithms for early detection of different chronic diseases like heart, kidney, Parkinson’s, breast, and brain diseases. Supervised algorithms such as SVM, KNN, RF, LR, NB, CNN, DT were implemented for every disease. All these algorithms were evaluated based upon their performance, accuracy, and processing time. It was concluded that SVM, RF, and LR were the most widely used algorithms. For predicting heart disease, LR proved to be the most effective and reliable algorithm. For breast cancer, RF showed more precision in probability for correct predictions. SVM model was best suited for kidney diseases and CNN for common diseases. In this way, different supervised ml algorithms were implemented to predict chronic diseases.

Citation: [Marouane Fethi Ferjani, “Disease prediction using Machine Learning,2020]

Another paper gave insights about the identification and prediction of disease using KNN, NB, SVM, RF for symptoms. Comparative analysis of the above algorithms was performed, and it was observed that all models gave 100 % accuracy. However, here, the combined model of all algorithms was not implemented. It was concluded that the selection of symptoms and their relevancy to each other also affects the accuracy of algorithms.

Citation: [Kunal Takke, Rameez Bhaijee, et all, “Medical Disease Prediction using machine Learning Algorithms, 2022]

One of the papers proposed a system of disease prediction using a decision classifier. Systems will recognize the disease based upon symptoms. However, when the same dataset was implemented using RF and GNB. It was observed that RF gave the highest accuracy with 97%. Hence, RF was suited well for a given dataset. However, if the combined model is implemented, an overall high accuracy can be obtained, which is explained through our proposed system.

Citation: [Raj H. Chauhan, Daksh N. Naik, et al, “Disease prediction using Machine Learning”,2020]

Study regarding the use of data mining algorithms like Weighted Associative Rule mining (WARM) is also performed for the diagnosis of heart disease. By using WARM, the strength of all the different features that contribute to heart disease is calculated. It required data preprocessing, feature selection, weight computation of features, applying model, and then evaluation. A set of crucial features with different weights was used for representing the strength and weightage of each of the feature that was needed in heart disease prediction. By assigning appropriate weights, high confidence prediction was possible.

Citation: [Armin Yazdani, Kasturi Dewi Varathan, et al, “A novel approach for heart disease prediction using strength scores with significant predictors,2021]

Therefore, a large amount of study and research is done and is currently undergoing in the field of disease prediction using ml algorithms and techniques. This study has led to tremendous advancements in the applications of ml in the healthcare systems. Based upon this literature review, a modified system for disease prediction is proposed.

3. Data Flow And architecture

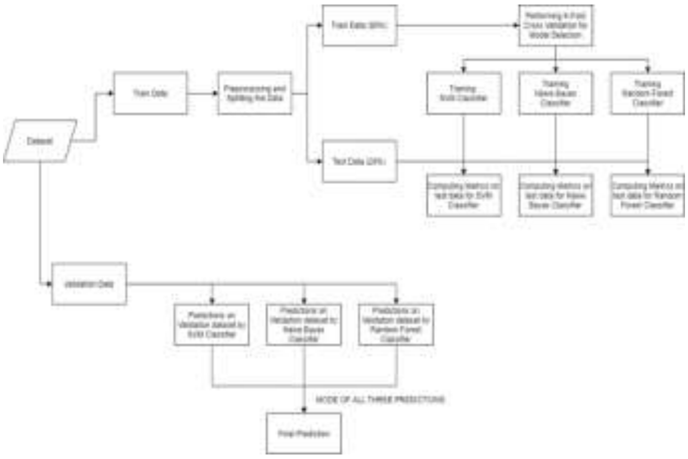

Dataset
Train Data
Preprocessing and Splitting the Data
Train Data (80%)
Performing k-Fold Cross Validation for Model Selection
Training SVM Classifier
Training Naive Bayes Classifier
Training Random Forest Classifier
Test Data (20%)
Computing Metrics on test data for SVM Classifier
Computing Metrics on test data for Naive Bayes Classifier
Computing Metrics on test data for Random Forest Classifier
Validation Data
Predictions on Validation dataset by SVM Classifier
Predictions on Validation dataset by Naive Bayes Classifier
Predictions on Validation dataset by Random Forest Classifier
MODE OF ALL THREE PREDICTIONS
Final Prediction

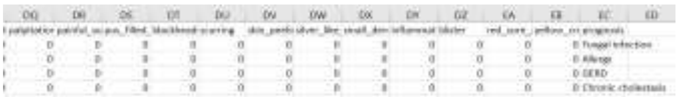
Figure 3. Architecture of working of model

4. Process flow

Preparing dataset

So, the approach for implementation of all algorithms starts with data preprocessing. Dataset used consists of two csv files. One is training and other is test dataset. All the models train on training datasets and then they are tested on test dataset. There are total 133 columns in dataset, out of which 132 are symptoms and last column is prognosis. As data available is not clean, we need to preprocess data before feeding it to model so that we obtain quality results. After cleaning dataset, it does not contain any null values. The prognosis column is label encoded as it is having string value. Rest of the columns have binary values 0 or 1 indicating absence and presence of symptom. Now this clean data is used to build and test different models.

This is sample of training data set.



DQ	DR	DS	DT	DU	DV	DW	DX	DY	DZ	EA	EB	EC	ED
palpitation	painful_wa	pus_filled_	blackhead	scurring	skin_peeli	silver_like	small_den	inflammat	blister	red_sore_	yellow_cr	prognosis	
0	0	0	0	0	0	0	0	0	0	0	0	Fungal infection	
0	0	0	0	0	0	0	0	0	0	0	0	Allergy	
0	0	0	0	0	0	0	0	0	0	0	0	GERD	
0	0	0	0	0	0	0	0	0	0	0	0	Chronic cholestasis	

Figure 1.1 Sample of training data set.

2. Implemented models

In implementation, three classification algorithms are used. They are as followed.

2.1 SVM (Support Vector Machine)

SVM algorithm comes under the supervised learning classification algorithms. This algorithm is used for building a best decision line on the given data to segregate and distinguish n-dimensional space into different classes, so that when a new data point has to be classified in future, it’s very easier and convenient to segregate the data point in correct expected category. This decision boundary is termed as a hyperplane. This algorithm chooses the points/vectors which are closest to the boundary (extreme) to create a hyperplane. These vectors present at boundary are called as support vectors and hence this algorithm is known as Support Vector Machine. This algorithm is used for**, text categorization face detection, image classification,** etc.

SVM can be of two types:

Linear SVM:

When it is possible to separate the data linearly, i.e. the dataset can be classified into two distinct categories, Linear SVM is used in such cases and it is known as Linear SVM classifier.

Non-linear SVM:

Non-Linear SVM is mostly used for the data which cannot be separated linearly into two distinct categories, which means that data cannot be classified using a single straight line, instead it needs a plane to classify the data. Such type of data is called as non-linear data and classifier used for non-linear data is called as Non-linear SVM classifier.

Support Vectors:

The data points or vectors which are closed to hyperplane are termed as Support Vector. They also affect the position of the hyperplane which distinguishes different classes. As these vectors support the hyperplane, hence called as Support vector.

2.2 Random Forest Classifier.

Random Forest is a supervised learning classification algorithm. It is implemented for both classification and regression problems. Random forest contains multiple decision trees of subsets on dataset. Average of predictions from multiple decision trees is taken into consideration for predicting final output. Therefore, it increases predictive accuracy of given dataset. Rather than relying on single decision tree, output of multiple decision trees is considered and based upon majority votes of predictions, the final output is predicted.

Accuracy is directly proportional to the number of trees in the forest. Greater the number of trees in the forest, higher is the accuracy obtained. Random forest consumes less training time as compared to the other algorithms. Even if the data is not pre-processed, it predicts output with high accuracy.

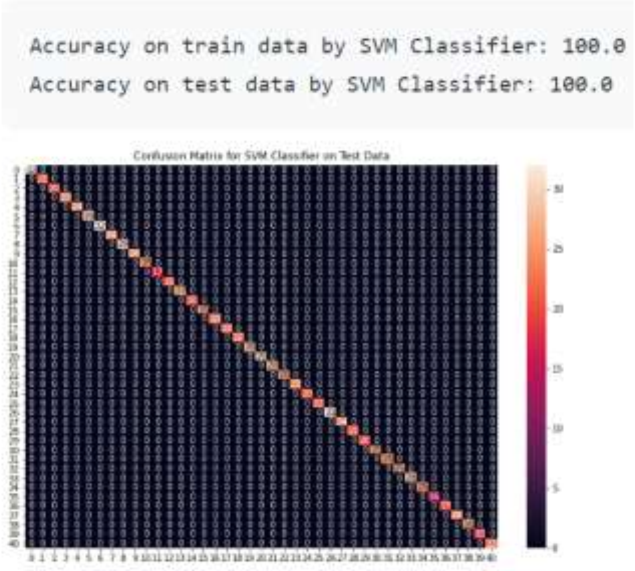
2.3. Gaussian Naïve Bayes

Gaussian Naïve Bayes is a classification algorithm which works based upon the calculations of probabilities by using Bayes theorem. This algorithm is perfect choice to deal with real world problems where there is less interdependency of features. Gaussian probability density function is used for predicting results after substituting parameters with new input values and final result is estimated for probability of new input value. In Gaussian Naïve Bayes algorithm, it is assumed that all features are independent of each other. Their values do not rely on each other. Training data is provided so that estimation of all the parameters which are required for classification is done. As Naïve Bayes has simple and efficient working, it is recommended to use this algorithm in solving many real-life problems and scenarios.

3. Model building

Training dataset is further divided into training and testing dataset as 80% training and 20 % test data set. Now K fold cross validation technique is used to split the training dataset into k subsets. Training is performed on k-1 subsets and one subset is used to evaluate performance of model and the combined mean score is calculated of all subsets. It is observed that all three algorithms i.e., SVM, Random Forest classifier and Gaussian NB performed well and mean scores are after cross validation are also very high. Therefore, to build accurate model we take predictions of all three models so that even one model gives wrong prediction, remaining two will give correct output and final output will be precise one. Doing so, we will get most accurate predictions on unseen data. To evaluate models, confusion matrix is used. Confusion matrix is defined as a technique which is used to define and analyze the performance of any algorithm. It shows the performance of classification algorithm with the help of a matrix. This matrix summarizes overall performance of the algorithm. Hence quality of models is tested using confusion matrix. After visualizing confusion matrix and its accuracy, all algorithms show 100% accuracy on train as well as test data. It means that all three algorithms predict the same prognosis.

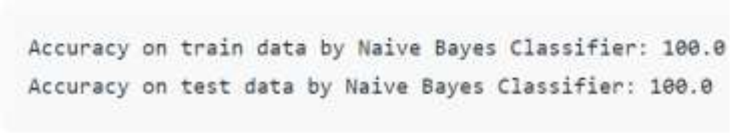
Below is confusion matrix of SVM classifier. It shows 100% accuracy.



```
Accuracy on train data by SVM Classifier: 100.0
Accuracy on test data by SVM Classifier: 100.0
```

Figure 3.1 Confusion matrix of SVM classifier.

Similarly, below is result for Gaussian Naïve Bayes Classifier.



```
Accuracy on train data by Naive Bayes Classifier: 100.0
Accuracy on test data by Naive Bayes Classifier: 100.0
```

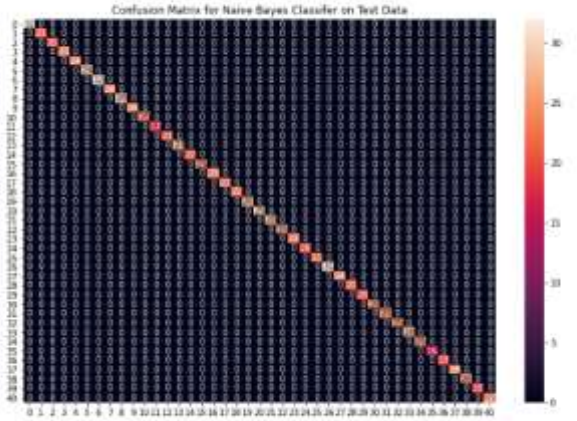
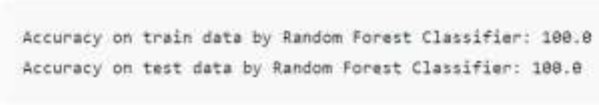


Figure 3.2 Confusion matrix of GNB.

It too gives 100% accuracy on both datasets.

Result for Random Forest Classifier is as follows.



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

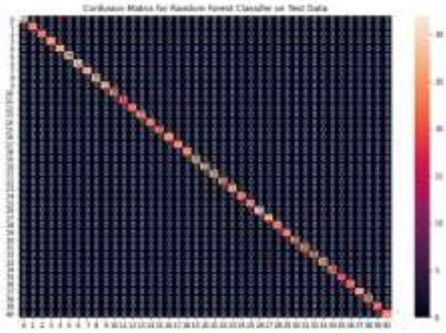
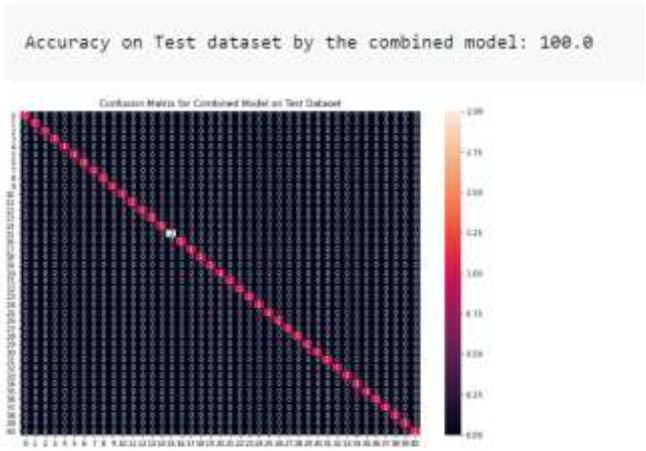


Figure 3.3 Confusion matrix of RF classifier.

Hence, above confusion matrices and results show that all three algorithms are giving 100% accuracy thereby predicting same prognosis. In this way model of algorithm is built and validated.

4. Fitting Model and validating on whole dataset.

From above results, it is observed that all models are working well on unseen data. Now these models will be trained on whole training data and then combined model will be tested on test dataset. After fitting models, the predicted result is combined, and confusion matrix is generated as follows:



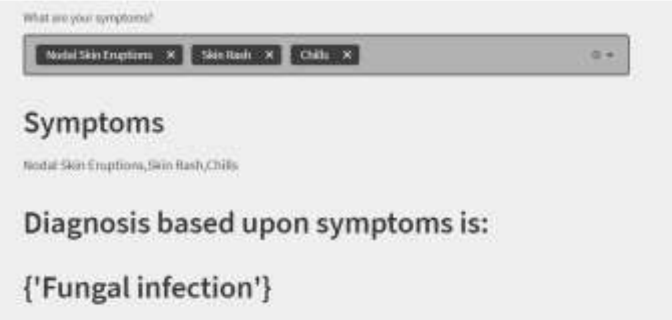
```
Accuracy on Test dataset by the combined model: 100.0
```

Figure 4.1 Confusion matrix of Combined model.

Above confusion matrix shows that combined model has classified all data points accurately.

5. Implementing function for accepting input.

Now to take inputs from user function is created. It takes symptoms separated by commas and predicts disease using combined model. The symptoms given as input are same as those present in dataset files. In this way, final output is predicted by implementing a function. Below is sample input and output given to function.



What are your symptoms?

Nodal Skin Eruptions × Skin Rash × Chills ×

Symptoms

Nodal Skin Eruptions,Skin Rash,Chills

Diagnosis based upon symptoms is:

{'Fungal infection'}

Figure 5.1 End user (patient) output window.

In this way, whole implementation of all three models is performed.

5. Results and Analysis

Findings shows that Random Forest algorithm performs better than KNN and Gaussian Naïve Bayes algorithm. But when features are independent, Gaussian Naïve Bayes outperforms Random Forest algorithm and KNN. Also, Gaussian Naïve Bayes and Random Forest are more robust, efficient and accurate than KNN algorithm and are more suitable for larger datasets whereas KNN is suitable for smaller datasets. Here, it is observed that accuracy of all three algorithms is 100%. Since dataset is preprocessed and uniform, it is difficult to distinguish and compare three algorithms based upon their accuracy and performance as all models are predicting correct and accurate prognosis. Experimental results prove that implemented model performs with 100% accuracy and hence it is more reliable.

6. Conclusion

It can be concluded that, using classification and prediction ml algorithms, we can develop a reliable, precise, accurate and cost-effective solution to predict diseases based upon symptoms. This system gives idea of preliminary diagnosis and medication to patients. It is believed that this proposed system will reduce seriousness of symptoms and disease by early prognosis. Early detection with high accuracy is possible with this model. Further users can consult doctors for treatment and medication. This model can be further modified to detect serious diseases like cancer, heart diseases etc. Future scope of this model includes other applications such as hand writing recognition, image recognition, X ray image classification etc. Hence proposed system proves to be very helpful, feasible and reliable solution and application of machine learning models.

References

[1] G. Battineni, G. G. Sagaro, N. Chinatalapudi, and F Amenta, “Applications of machine learning predictive models in the chronic disease diagnosis,” *Journal of Personalized Medicine*, vol. 10, no. 2, p. 21, 2020.

[2] Marouane Fethi Ferjani, Bournemouth University, England. ‘Disease Prediction using ML’.https://www.researchgate.net/publication/347381005_Disease_Prediction_Using_Machine_Learning

[3] Dr C K Gomathy, Mr. A. Rohith Naidu Sri Chandrasekharendra Saraswathi Viswa Mahavidyalaya , Kanchipuram Dr C K Gomathy, Mr. A. Rohit Naidu, Sri Chandrasekharendra Saraswati Viswa Mahavidyalaya, Kanchiouram, ‘Prediction of Disease using Machine Learning’. https://www.researchgate.net/publication/357449131_THE_PREDICTION_OF_DISEASE_USING_MACHINE_LEARNING

[4] Sneha Grampurohit, Chetan Sagarnal , K.L.E Institute of India, Hubli, ‘Disease Prediction using ML algorithms’,2020 International Conference for Emerging Technology (INCET).

[5] https://ieeexplore.ieee.org/abstract/document/9154130

[6] Rinkal Keniya, Aman Keniya, Vruddhi Shah, KJ Somaiya College of Engineering, ‘Disease Prediction from various symptoms using machine learning’

[7] https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3661426

[8] Shahadat Uddin, Arif khan, Md Ekramul Hossain, Mohammad Ali Moni, BMC Medical Informatics and Decision Making, ‘Comparing different supervised machine learning algorithms for disease prediction’ https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-1004-8

[9] M Vamshi Krishna Reddy, G V P Sai Abhijit , K Sai Nath, Mangali Sathyanarayana, ‘Disease Predictor Based on Symptoms using Machine Learning’.

[10] Anuj kumar, Mr Analp Pathak, ‘ A Machine Learning Model for Early Prediction of Multiple Diseases to Cure Lives’

[11] Kunal Takke, Rameez Bhaijee, Avinash Singh, Mr. Abhay Patil, ‘Medical Disease Prediction using Machine Learning Algorithms’

[12] Sneha Grampurohit, Chetan Sagarnal, ‘Disease Prediction Using Machine Learning

Algorithms'.

[13] P.Hamsagayathri, S Vigneshwaran, 'Symptoms based Disease Prediction Using Machine Learning Techniques'.

[14] Kommineni Sai Kumar, Maganti Sai Sathya, Abdul Nadeem, Singamaneni Rajesh, 'Disease Prediction based on Symptoms using Database and GUI'.

[15] Mayur Gadekar, Soyeb Jamadar, Prajak Pachpute, Sanket Shinde, Swati Bhosale, 'Symptoms based disease Prediction'

[16] Nivethitha. A, Pramoth Krishnan. T, Narendran. G, 'Smart disease Prediction Using Machine Learning'

[17] Raj H. Chauhan, Daksh N. Naik, Rinal A. Halpati, Sagarkumar J. Patel, Mr. A. D. Prajapati, 'Disease Prediction using Machine Learning'