Acropolis Institute of Technology and Research

Disease Prediction

*Aman Patel, Anurag Yadav, Bhomik Gahlot, Chandransh Singh Chouhan*

INTRODUCTION

Humans are now afflicted with a variety of ailments as a result of the current state of the environment and their lifestyle choices. It is critical to dtect and forecast such diseases at an early stage to prevent them from progressing to their final stages. Most of the time, doctors find it challenging to precisely identify ailments by hand. There are multiple techniques in machine learning that do predictive analytics on large amounts of data that are used in a variety of industries. Predictive analytics assist doctors in making accurate decisions regarding health and treatment of a patient based on the vast data available, but altogether it is a very difficult process. Several diseases like cancer, diabetes etc. are becoming the cause of deaths globally a the lack of early detection of these diseases. The lack of good medical infrastructure ad most common reason is d a low ratio of doctors to the population is also a major fact r. Based on the recommendation of WHO, the ratio of doctors to patients t should be 1:1000 but in India the ratio is 1:1456, which indicates the shortage of doctors in India. Thus to save a lot of lives, early recognition and diagnosis of these diseases is very crucial. This work is all about predicting diseases using machine learning algorithms.

OBJECTIVES

There is a need to study and make a system which will make it easy for an end users 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.
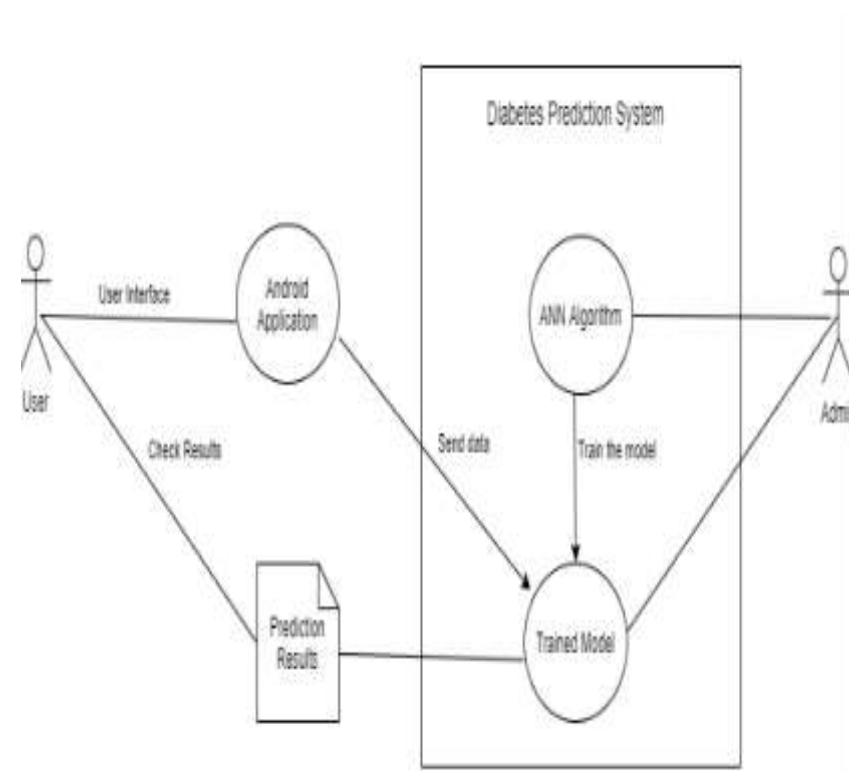
The user gives input to the system and system provides the user with a set of probable diseases

METHOD

In supervised machine learning, it is important to train an estimator on balanced data ,so the model is evenly informed on all classes. The observation of the dataset and its visualization leads us to the conclusion that the data is balanced and there's no imbalance in the data, which means that training and testing will give real exactness

It is a straightforward method for building classifiers, which are models that give class labels to problem cases represented as vectors of characteristic values. Here the class labels are selected from a finite set. For training such classifiers, there is no one algorithm, but rather a variety of algorithms based on the same principle: all naive Bayes classifiers assume that the value of one feature is independent of the value of any feature, given the class variable.

For example, if the fruit is red, round, and around 10 diameter, it is termed an apple. A naive Bayes cla examines each of these characteristics to cont independently to the likelihood that this fruit is an regardless of any possible confounding variables.
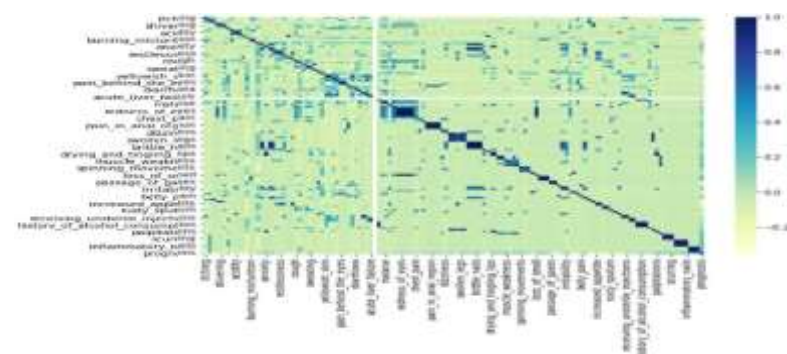
Multinomial Naive Bayes: The naive Bayes classification strategy works with Bayes' theorem, which is known to be used in arithmetic and computer science probabilistic analysis. This algorithm considers each property as an autonomous property that contributes to the final classification. For this reason, it has been praised in various studies for its accuracy and unwavering quality.
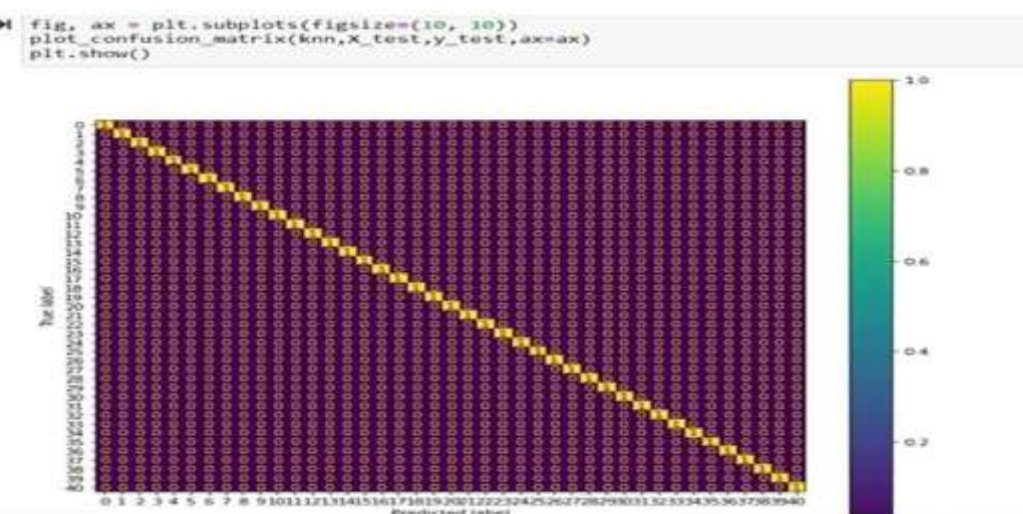
Data collection and Data analysis:

Dataset: The data is collected from Columbia University and this study of diseases and their corresponding symptoms is available on Kaggle.

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Results

Selection of less symptoms: The project is set up in such a way that the user is introduced to the chatbot system and a preference is asked whether the user would like to get an estimation or prediction of his/her disease based on the input into the chatbot the user is taken to the prediction system that contains machine learning algorithms like Naïve Bayes, Random Forest Classifier ,K- Nearest Neighbors and Support Vector Machine classifier which takes the user's symptoms as input and the prediction labelsbased on individual accuracies. The user has the option of selecting one to five symptoms, if lesssymptoms are entered, the accuracy will be lower hence, the larger the number of

```
In [25]: fig, ax = plt.subplots(figsize=(10, 10))
         plot_confusion_matrix(knn,X_test,y_test,ax=ax)
         plt.show()
```

symptoms, the more accurate the diagnosis.
Selecting of symptoms which are not relevant to each other: If the user selects the symptoms which are not relevant to each other in the diagnosis of the disease then the machine learning models may give less accuracy on the classification problem.

```
Out[43]: <matplotlib.legend.Legend at 0x1bfa99c6700>
```
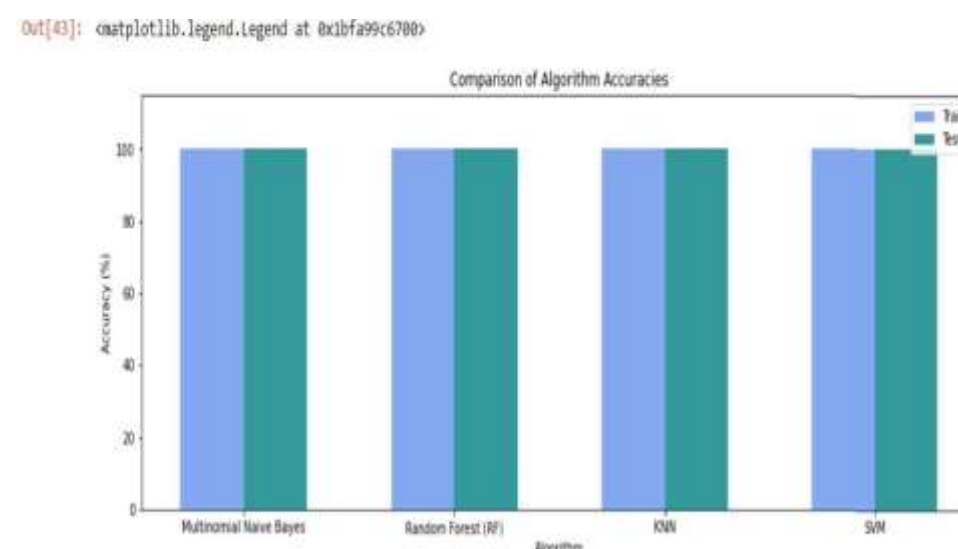
The F1 score will only be 1 if both the fit and recall are 1. The F1 score is high only if both the fit and recall are high. The F1 score is the harmonic mean of precision and recall and is a better measure than precision.

Conclusion

This paper proposed a method of identification and prediction of the presence of a disease in an individual using the machine learning algorithms like Naïve Bayes, Random Forest Classifier, K-Nearest Neighbors and Support Vector Machines given that the user have given the maximum of five symptoms. Thus, the given methods also lead to a comparative analysis of various machine learning algorithms for multiclass classification. It is highly believed that the proposed system can reduce the risk of diseases by diagnosing them earlier and also reduces the cost of diagnosis, treatment, and doctor consultation, however the selection of symptoms does play an important role in the accuracy of the disease prediction

REFERENCE

International Journal of Emerging Technologies and Innovative Research (www.jetir.org | UGC and issn Approved), ISSN:2349-5162, Vol.8, Issue 3, page no. pp18-20, March-2021

issn Approved), ISSN:2349-5162, Vol.8, Issue 3, page no. pp18-20, March-2021

Catano, V. M., Wisner, W. H., & Hackett, R. D. (2016). Recruitment and selection in Canada (6th Ed.). Toronto, ON: Nelson Education Ltd

Chen, C., Huang, Y., & Lee, M. (2011). Test of a model linking applicant resume information and hiring

Recommendations. International Journal of Selection and Assessment, 19, 374-387.

Derous, E., & Ryan, A. M. (2012). Documenting the adverse impact of resume screening: Degree of

ethnic identification matters. International Journal of Selection and Assessment, 20, 464-474.

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

CONTACT

1.Aman patel,

2. Anurag yadav,

3. Bhomik gahlot,

4. Chandransh singh chouhan   CS1 Year

amanPatelcs21@acropolis.in

anuragyadav20388@acropolis.in

Chandranshchouhan20665@acropolis.in

Bhomikgahlotcs21@acropolis.in

Output

```
#Output
rf_model_prediction --> Fungal infection
naive_bayes_prediction --> Fungal infection
svm_model_prediction --> Fungal infection
final_prediction --> Fungal infection
```