Comparative Analysis on Machine Learning Algorithms for Multiple Disease Prediction

aSivadi Balakrishna, bYerrakula Gopi

a,bDepartment of Computer Science & Engineering, Vignan’s Foundation for Science, Technology & Research

(Deemed to be University), Vadlamudi, Guntur, A.P., India.

adrsivadibalakrishna@gmail.com

bygopi091@gmail.com

Abstract

Keywords

K-Nearest Neighbor (KNN),

Random Forest,

Decision Tree,

Naïve Bayes,

Disease Prediction, Machine Learning

These days, majority of the humans are suffered from multiple diseases because of eating habits and environmental situations. Henceforth, the prediction of these multiple diseases become a challenging and critical task these days. Machine Learning (ML) algorithms becomes more popular to predict multiple diseases. In order to predict and analyze multiple diseases, in this paper, we investigated and examined various ML algorithms such as Decision Tree, Random Forest, Naïve Bayes, K-Nearest Neighbor (KNN) used for accurate prediction of disease. For analysis of the ML-based classification algorithms, this paper intently used Accuracy as a performance metric and tested it on the DiseaseSymptomKB dataset. The accuracy of general disease prediction by using Decision Tree is 95%, Random Forest is 95%, Naïve Bayes is 95% and KNN is 92%.

1. Introduction

Computerized Medical Information (CMI) has grown in popularity in many locations throughout the world, such as EMR for replacing paper medical records, and online medical records. There are a lot of reports and appointments, thus there's a lot of healthcare. Data has accumulated because of the fast expansion of the internet [1]. It has been described that Health care analysis has a massive impact on medical information, which they have, analyzes the data through machine learning inputs. This provides a clear view to patients about the disease. It deals with the overall performance report through deep learning techniques. For any development of these areas, they are dependent upon machine learning algorithms [2]. Over recent years Machine learning is one of the most widely used approaches in a variety of computer engineering fields, including natural language processing, image processing, pattern identification, cyber security, and many more. Healthcare is one of the most active areas in machine learning research [3]. The extracts from numerous papers, as well as general observations in a hospital setting, demonstrate that there is a need for cheap healthcare services for the huge Indian population, particularly for those living in distant areas. To diagnose, new techniques are required. As a result, less costly, simple, and reliable approaches for diagnosing disease and ensuring treatments should be adopted. Noninvasive diagnostic procedures, on the other hand, must be researched. Healthy persons are classified using machine learning techniques [4-5].

Nowadays, Multiple Disease Prediction is the same as disease prediction but here it contains data of different diseases and their symptoms. It predicts whether the symptoms match with the symptoms of disease present in the data. Some symptoms ignored now may lead to fatal disease in a later stage. Therefore, the prediction of disease at an earlier stage becomes an important task.

There is a general image of the doctor using technology is shown in the below Fig 1.

****

Figure 1. The general image of a doctor using technology

The features of Disease Prediction Using Machine Learning are as follows.

* We can easily prepare the data and transform that data into an algorithm, which may reduce the overall work of the system
* To make the user more application friendly rather than discussing with others for their disease.
* It provides the necessary options to choose from the drop-down.
* Here the user’s name and Symptoms entered and the disease predicted are stored in the database for future reference.
* User must enter the name and at least 2 symptoms to the max of 5 symptoms to predict the disease.

The remaining sections of this paper have been organized as follows: Section 2 describes the literature survey related to the multiple disease predictions based on machine learning algorithms. Section 3 deals with the proposed methodology for multiple disease detection. Section 4 handles the Results and Analysis with obtained ones. Finally, section 5 concludes this paper with future enhancements.

1. Literature Survey

In this section, the authors are discussed on multi-disease prediction based on machine learning algorithms of several researchers over the last years.

Initially, Karim et al. [6] created a method based on DSAE and ESD. Through extracting features that are more efficient in an amount of time in 2018. Unlike traditional frameworks, this framework builds functionality quickly. By using DSAE architecture, it increases the proportions of data. Low-proportional data is poor in the DSAE and ESD. Abdullah and Rajalaxmi [7] have proposed Decision trees to forecast the accuracy of events connected to heart disease. Jun Wu was the associate editor in charge of calculating the evaluation of this work and approving its publication. Then predicting cardiac disease, various methods of knowledge abstraction have been used, including established data mining methods. Hongmei et al. [8] the work was titled "A multilayer perceptron-based medical decision support system for heart disease diagnosis." High blood pressure, Coronary heart disease, Rheumatic valve heart disease, Chronic, and Congenital heart disease are the five heart disorders identified using heart auscultation in this study. Kar and Maity [9] had described a model consisting of four stages: preprocessing, identification of potential lesion, removal of optic disc and vessels, and post-production. Tao et al. [10] took the machine-learning model for IHD detection. Then, three categories of characteristics were retrieved and fed into separate classifiers. The last step is to evolve the prediction results of all three classifiers.

Jiang et al. [11] described individuals who are on the go and have the same symptoms. Patients who are members of the same group can safely communicate PHI or their personal medical experiences with one another by utilizing the AA's system settings and protected information. They did not explore the possibility of sharing a secret with others without sacrificing generality, because this form of active assault is impossible to prevent in practically all security systems. Das et al. [12] predicted heart illness using ANN's backpropagation multilayer perception (MLP). When compared to the findings of current models in the same area, the acquired results are shown to be superior. Xiao et al. [13] described the MRLDC computational method for identifying illness circRNA candidates. We start by detailing the data sources and how we built a multilayer circRNA-disease heterogeneous network that comprises circRNA-circRNA, disease-disease, and circRNA-disease networks. A combined objective optimization approach is then created to anticipate disease-associated circRNAs. After deleting duplicate correlations from various pieces of evidence, we kept 331 associations between 312 circRNAs and for humans, there is 40 illness. Mirjalili et al. [14] proposed the Whale Optimization Algorithm that is inspired by humpback whales. On 29 well-known test functions, the WOA method is evaluated. The results on unimodal functions demonstrate WOA's superior exploitation. The results on multimodal functions validate WOA's exploring capabilities. The results on structural design challenges back up WOA's performance in the field. Arora et al. [15] provide a novel nature-inspired algorithm, the butterfly optimization algorithm (BOA), that solves global optimization problems by mimicking the food seeking and mating behavior of insects. On a set of 30 benchmark test functions, the proposed method is evaluated and validated, and its performance is compared to that of competing metaheuristic algorithms.

The following list of limitations are observed or inferred from the literature survey

* As it is observed from the survey, very few types of research are based on the prediction of multiple diseases in healthcare.
* The data produced in healthcare is very huge which leads to prediction difficulty.

This gave us the motivation leading to the interest in taking up this research, where multi-disease prediction is the latest approach that is being utilized and satisfactory analysis of data needs to be performed.

1. Proposed Methodology for Multi Disease Detection

In this section, the proposed methodology has been discussed and along with used Machine Learning algorithms for comparative analysis on multiple disease detection predictions. The following Machine Learning Algorithms were used for the comparative analysis of multiple disease prediction.

* Random Forest: Random Forest is a well-known supervised learning algorithm that is based on ensemble learning notion, is a method in which multifarious classifiers are amalgamated to address a complex issue and to enhance model efficiency.
* Naive Bayes: It is one of the most popular classification algorithms used to classify the variables based on their conditional probability values. For computation purposes, it uses the Bayes theorem and the used class labels can be used for the prediction of classified data.
* Decision Tree: A decision Tree is the technique of supervised learning, which is used for both Regression and classification; however, it is often used to solve classification problems. The classification is tree-structured, where inner denotes the characteristics of a set of data, branches indicate decision rules and each leaf node denotes the results. It is a schematic representation to get all the potential results to a decision/problem on given criteria
* K-Nearest Neighbor: K-Nearest Neighbor is a classical Machine Learning algorithm that uses the Supervised Learning approach. K-NN assumes the new data/case and available data’s similarity and the new data is put into the data into the most comparable group.

Figure2 represents the steps that are followed during the implementation of the multiple disease prediction. The system consists of steps like Data Collection, Pre-processing the data, taking the Training Data and applying ML algorithms, analysis, and predicting the results. At first, the dataset was preprocessed with preprocessing techniques and divided into training and testing. The training dataset is sent to machine learning algorithms and testing data is sent to the prediction model to test the model, the user interacts with the GUI interface and enters Symptoms and the result is predicted and displayed the Disease. As it is shown the detailed overview is given below:

1. Disease Dataset: The first step in the model includes the collection of datasets.
2. Data Pre-processing: The second step does the pre-processing which includes handing the missing values, dimensionality reduction, and vectorization, which is used to speed up the computation.
3. Splitting the data: The data is divided into a training set and a testing set. By default, the ratio of training to testing split is 80:20 (percent) respectively. i.e., 0.8 data is taken into training and 0.2 is taken into testing.
4. Machine Learning algorithms: The algorithms that were used for training our model are discussed in detail below.
5. Evaluation of the model: once the training is done, the model is evaluated and the testing set is evaluated later based on how the model is trained. The performance can be measured using performance metrics that were discussed in the Experiment Evaluation section.

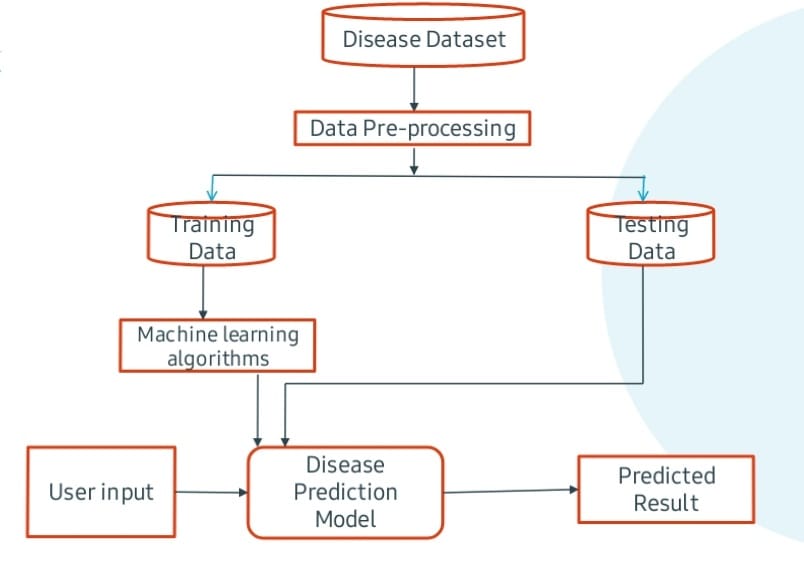


Figure 2. Workflow of the proposed system

1. Results and Analysis

In this section, we will discuss some of the classification metrics for the evaluation of the machine learning models we used in this paper. After that, we will provide the performance of the machine learning models and the GUI output results used in the Multiple Disease Prediction. Finally, we will discuss the performance of each algorithm in terms of accuracy and confusion matrix.

* 1. Dataset and Implementation Specifications
* Dataset for this work was collected from a study of the University of Columbia performed at Newyork Presbyterian Hospital during2004.
* http://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html
* Training dataset size:4920 X 133
* Testing dataset size:41 x 133

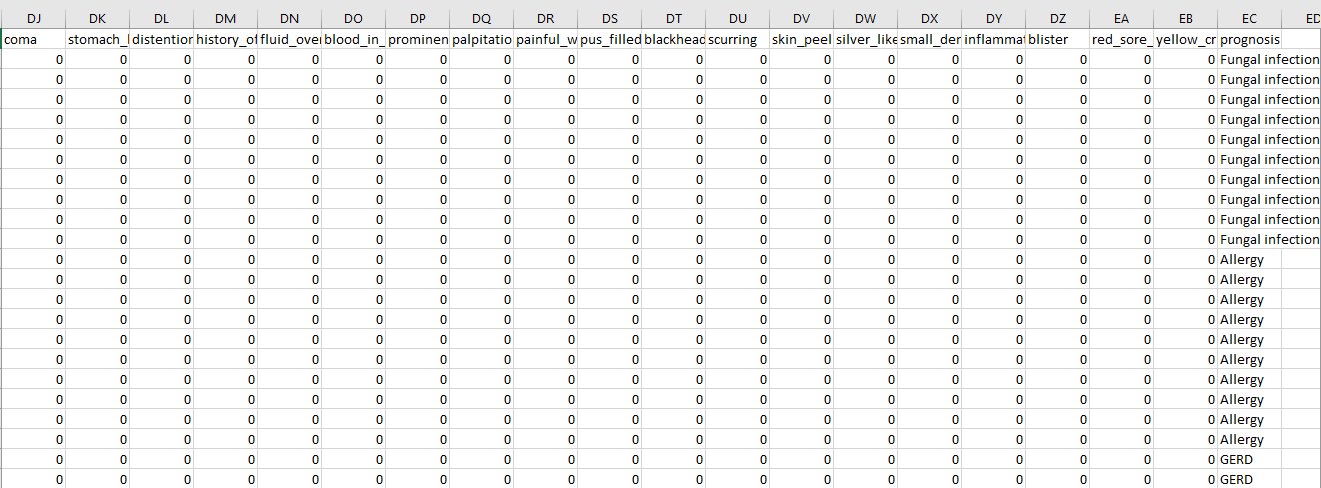
****

Figure 3. Sample image of the dataset.

The dataset consists of symptoms and diseases data. Here columns represent the symptoms (133) and row represent Disease and its symptoms. The data is in the form of 0’s and 1’s if the particular symptom is there for a particular disease, it is 1 or else it is 0. The sample dataset image is shown in Figure 3.

The proposed multiple disease prediction mechanism was built using the Anaconda Navigator (Jupiter notebook) platform. This has been written python code on I7 Intel Pentium Processor DELL laptop, 8 GB RAM, and 1 TB HDD on Windows 10 platform.

* 1. Evaluation metrics

In this section, the following evaluation metrics are used for evaluating the performance of the ML algorithms.

* + 1. Accuracy

The Accuracy is measured as correctly classified predicted class samples over all the given testing class data samples.

Accuracy = (correctly predicted class / total testing class) × 100%

These screenshots show the interface of the output screen using the Tkinter library in python. Fig 4 shows the interface to enter symptoms and get the prediction. The user enters the symptoms and clicks the predict buttons. The user should select the symptoms from the given dropdown as shown in Figure 5. Similarly, Figure 6 shows the predicted disease.

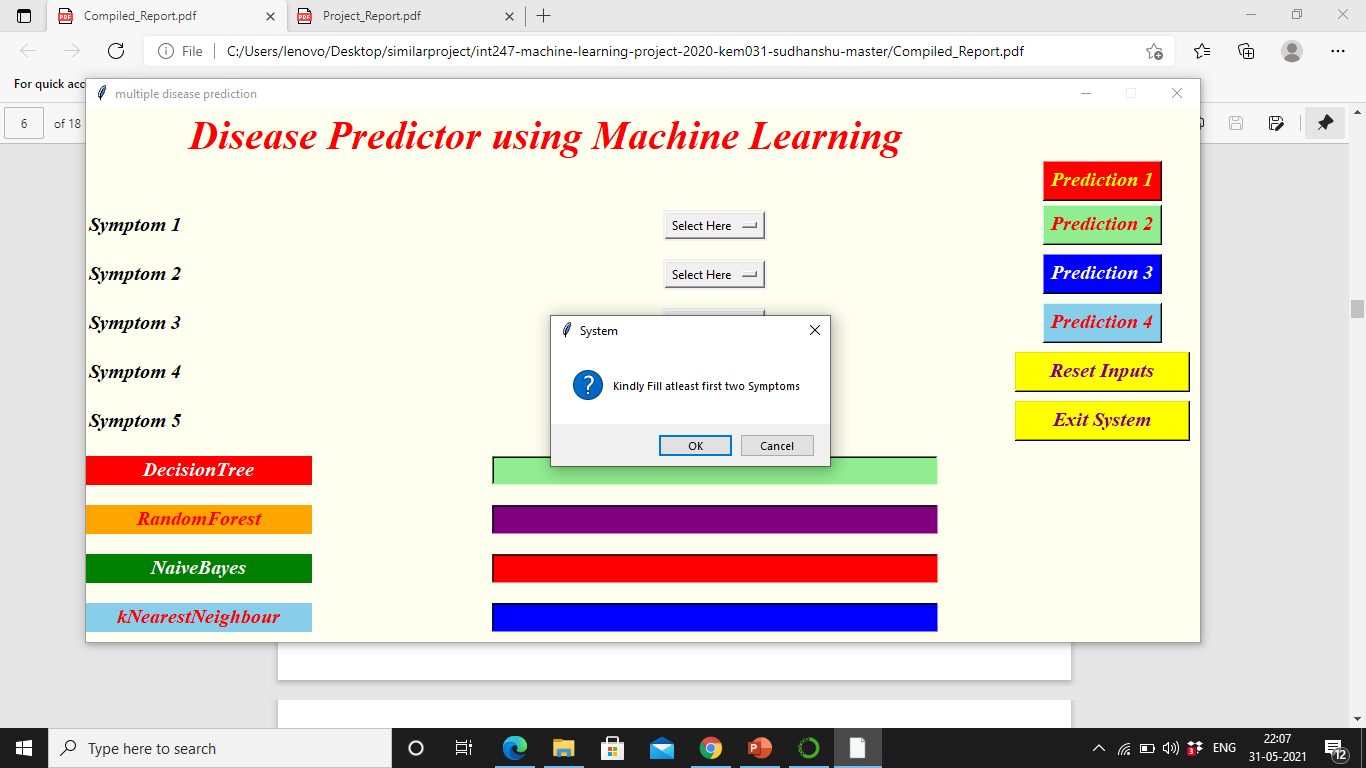
****

Figure 4. Interface to select the symptoms and get the prediction.

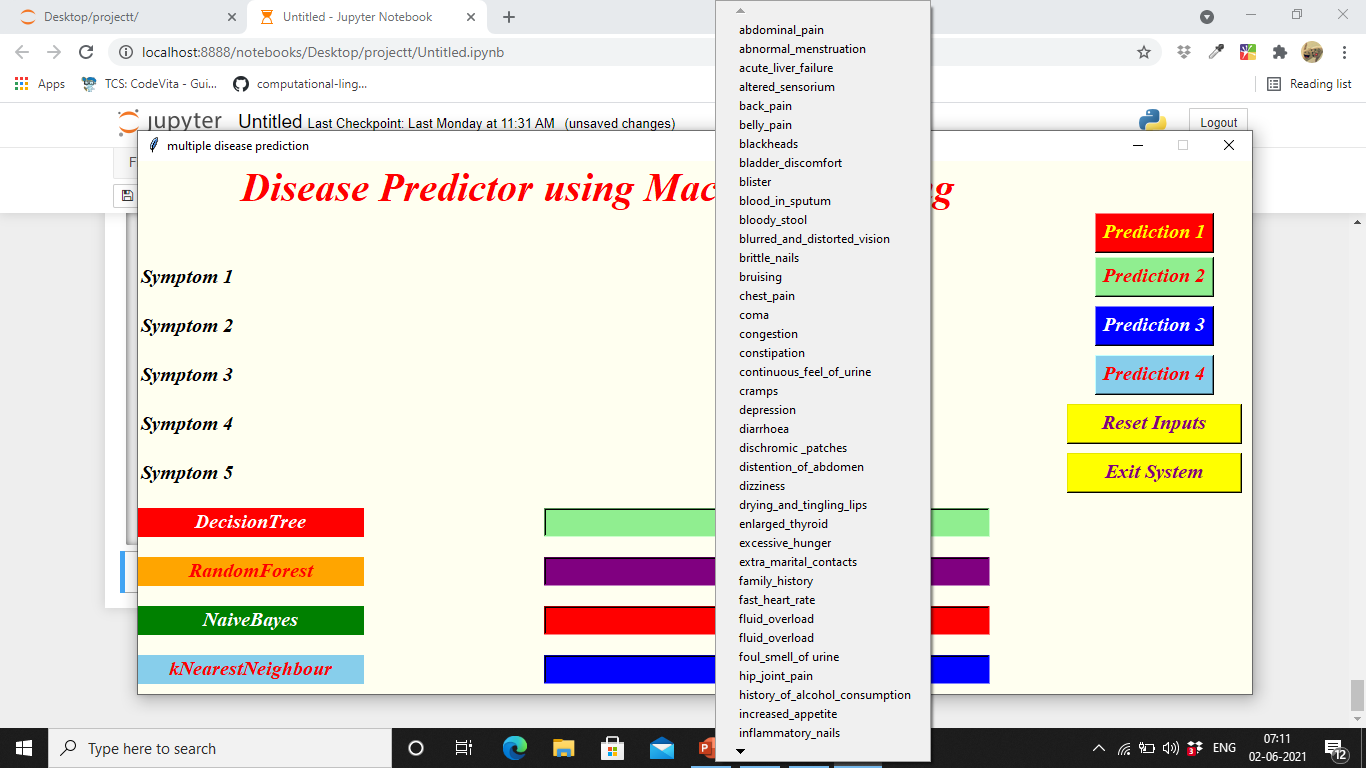
****

Figure 5. Interface to select the symptoms from the dropdown

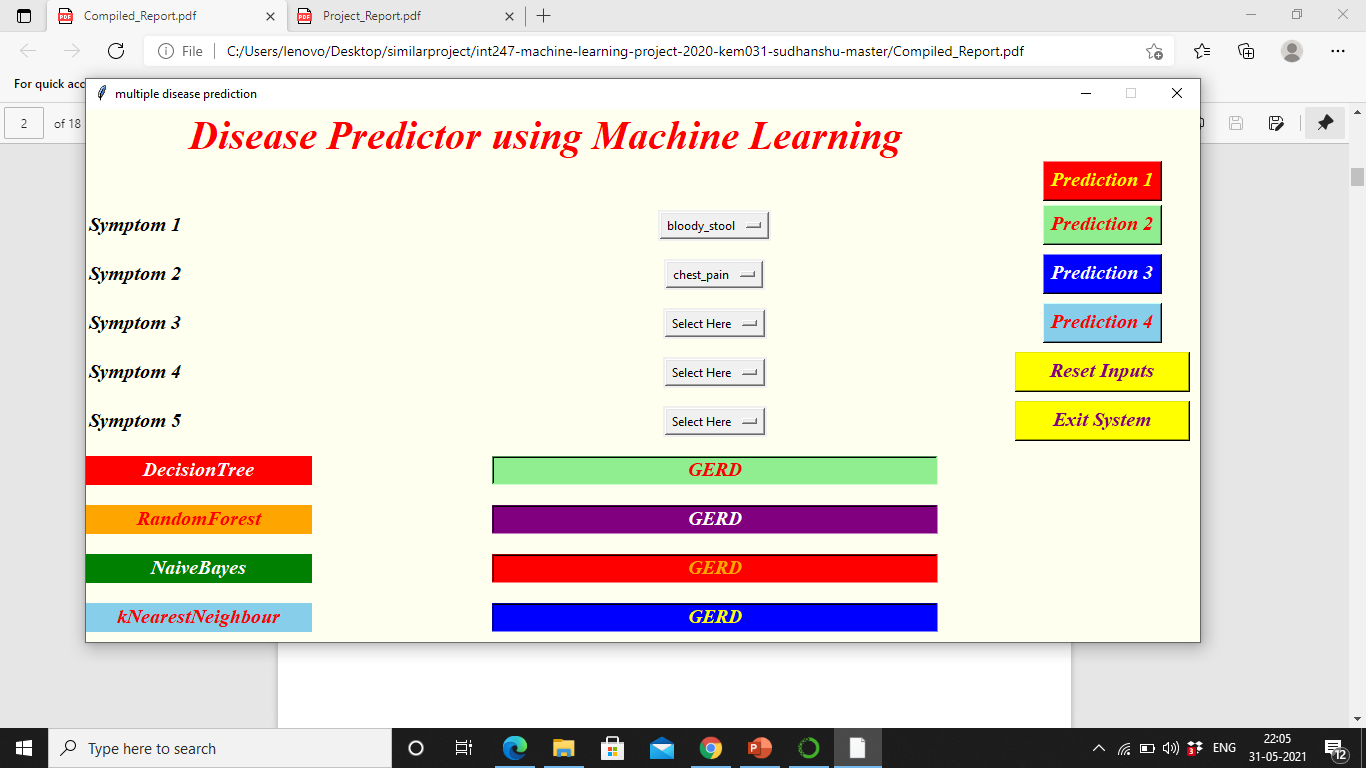
****

Figure 6. Interface to show the predicted disease

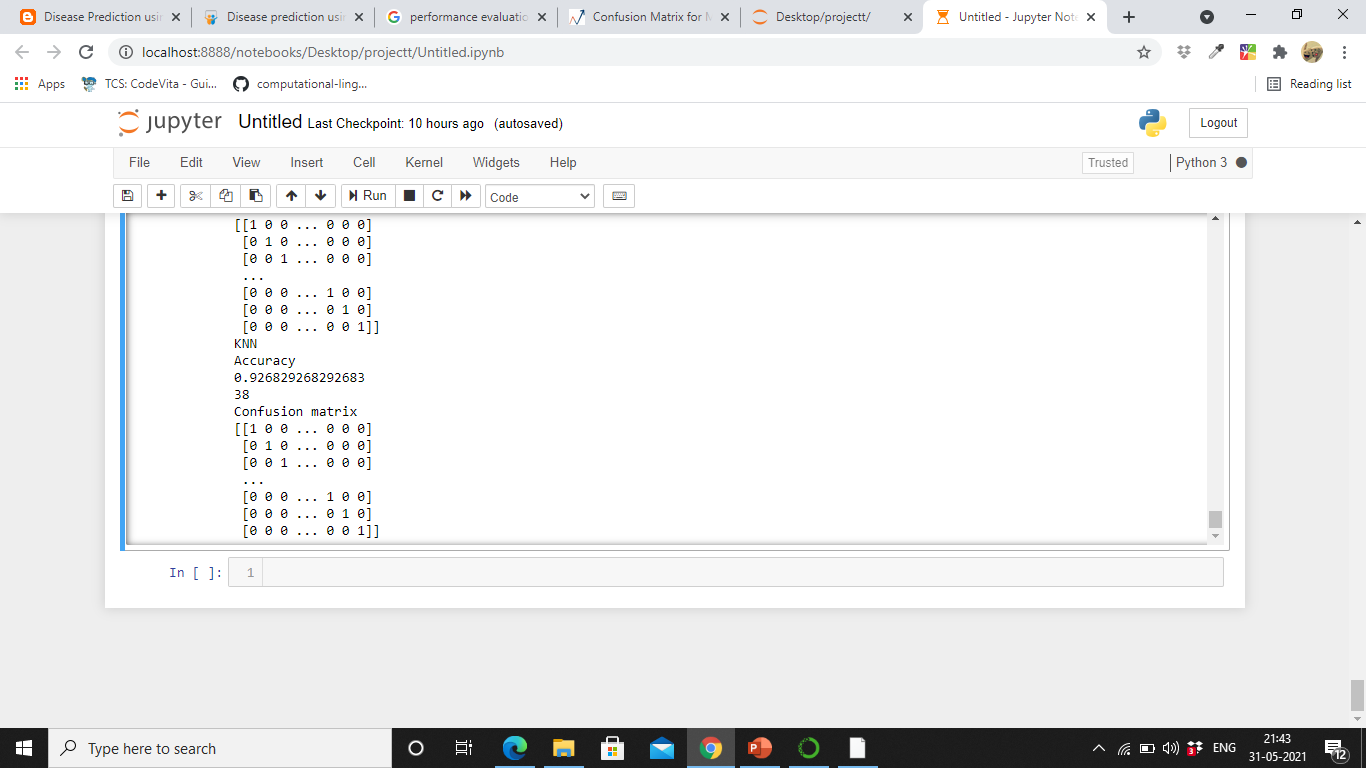
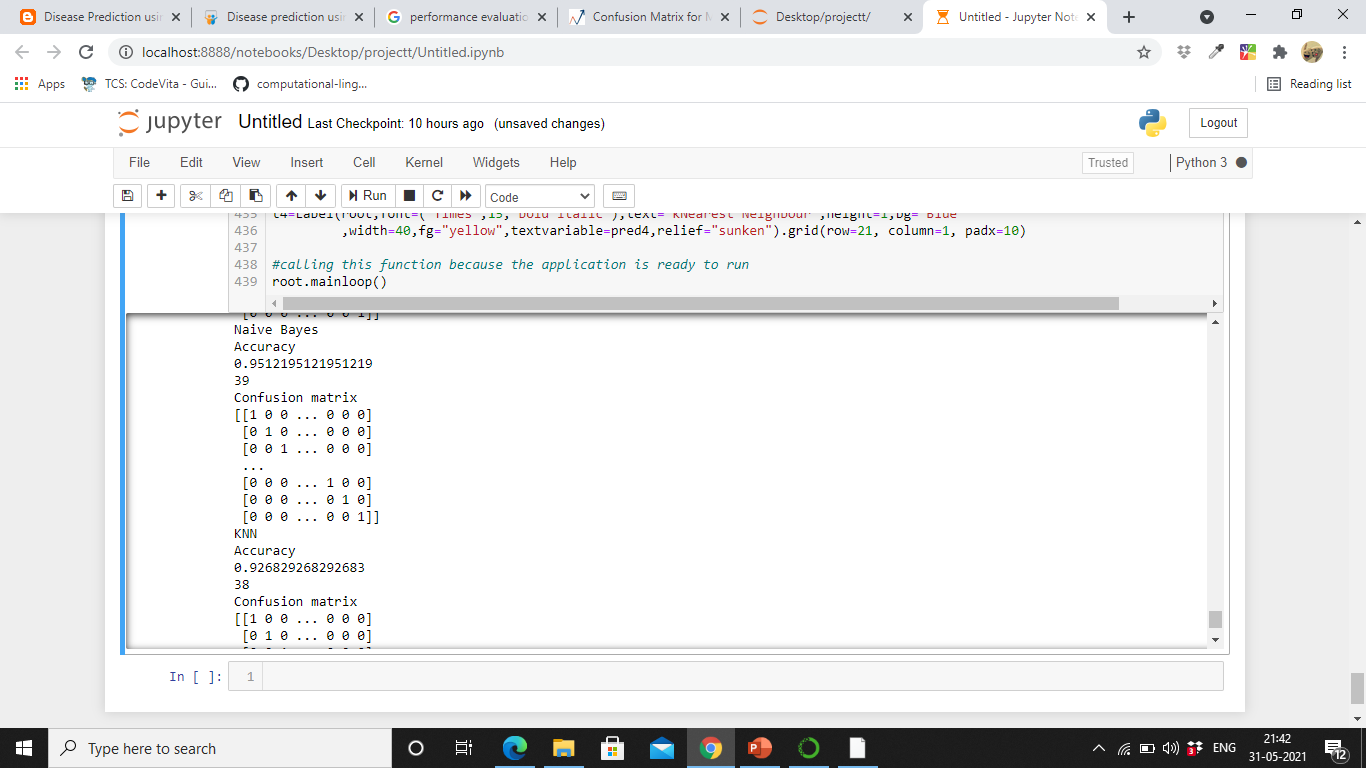
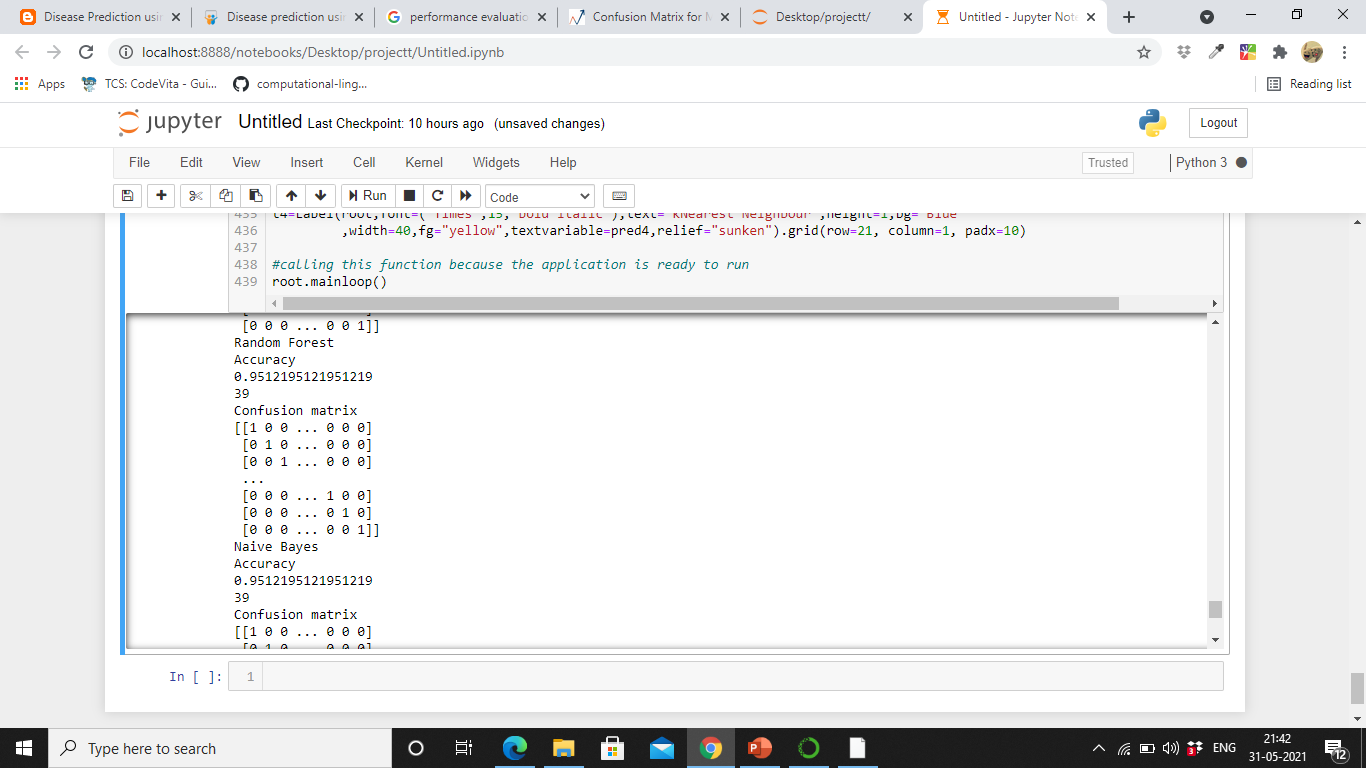
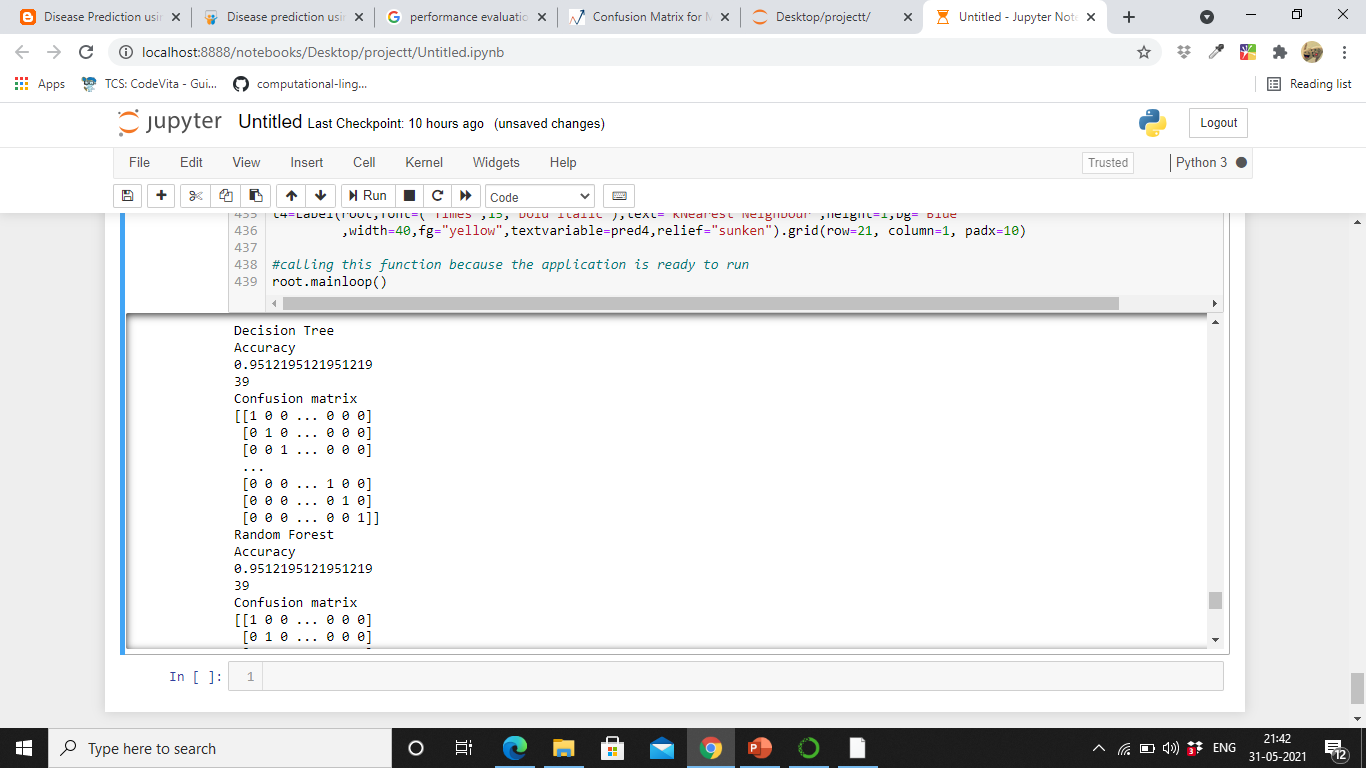
****

Figure 7. Accuracy and the confusion matrix for ML algorithms

Figure 7 shows the accuracy and the confusion matrix of all four i.e. Decision Tree, Random Forest, Naïve Bayes, and KNN algorithms. The database stores all the information regarding the users.

1. Conclusion and Future Work

This paper proposed and implemented a “Multiple Disease prediction” is to predict the disease based on the data/symptoms entered by the user using machine learning. We were successful in creating such a system that use 4 different algorithms to do so. The results obtained after performing machine-learning algorithms such as Decision Tree, Naïve Bayes, Random Forests accuracy is 95% and accuracy of KNN is 92%. In continuation of this work, it may extent by considering several diseases and various Deep Learning models.

References

1. Hong, Wenxing, Ziang Xiong, Nannan Z1. heng, and Yang Weng. "A medical-history-based potential disease prediction algorithm." IEEE Access 7 (2019): 131094-131101.
2. Peng, Yonghong. "A novel ensemble machine learning for robust microarray data classification." Computers in Biology and Medicine 36, no. 6 (2006): 553-573.
3. Saranya, G., and A. Pravin. "A comprehensive study on disease risk predictions in machine learning." International Journal of Electrical and Computer Engineering 10, no. 4 (2020): 4217.
4. Vijayalaxmi, A., S. Sridevi, N. Sridhar, and Sateesh Ambesange. "Multi-disease prediction with artificial intelligence from core health parameters measured through non-invasive technique." In 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1252-1258. IEEE, 2020.
5. Haq, Amin Ul, Jian Ping Li, Muhammad Hammad Memon, Asad Malik, Tanvir Ahmad, Amjad Ali, Shah Nazir, Ijaz Ahad, and Mohammad Shahid. "Feature selection based on L1-norm support vector machine and effective recognition system for Parkinson’s disease using voice recordings." IEEE Access 7 (2019): 37718-37734.
6. Karim, Ahmad M., Mehmet S. Güzel, Mehmet R. Tolun, Hilal Kaya, and Fatih V. Celebi. "A new framework using deep auto-encoder and energy spectral density for medical waveform data classification and processing." Bio-cybernetics and Biomedical Engineering 39, no. 1 (2019): 148-159.
7. A. S. Abdullah and R. R. Rajalaxmi, ``A data mining model for predicting the coronary heart disease using random forest classier,'' in Proc. Int. Conf. Recent Trends Compute. Methods, Commun. Controls, Apr. 2012, pp. 2225.
8. http://www.calsci.com/heartattacks.html
9. Kar, Sudeshna Sil, and Santi P. Maity. "Automatic detection of retinal lesions for screening of diabetic retinopathy." IEEE Transactions on Biomedical Engineering 65, no. 3 (2017): 608-618.
10. Tao, Rong, Shulin Zhang, Xiao Huang, Minfang Tao, Jian Ma, Shixin Ma, Chaoxiang Zhang et al. "Magnetocardiography-based ischemic heart disease detection and localization using machine learning methods." IEEE Transactions on Biomedical Engineering 66, no. 6 (2018): 1658-1667
11. Jiang, Shunrong, Xiaoyan Zhu, and Liangmin Wang. "EPPS: Efficient and privacy-preserving personal health information sharing in mobile healthcare social networks." Sensors 15, no. 9 (2015): 22419-22438.
12. Das, Resul, Ibrahim Turkoglu, and Abdulkadir Sengur. "Effective diagnosis of heart disease through neural networks ensembles." Expert systems with applications 36, no. 4 (2009): 7675-7680.
13. Xiao, Qiu, Jiawei Luo, and Jianhua Dai. "Computational prediction of human disease-associated circRNAs based on manifold regularization learning framework." IEEE Journal of biomedical and health informatics 23, no. 6 (2019): 2661-2669.
14. Mirjalili, Seyedali, and Andrew Lewis. "The whale optimization algorithm." Advances in engineering software 95 (2016): 51-67.
15. Arora, Sankalap, and Satvir Singh. "Butterfly optimization algorithm: a novel approach for global optimization." Soft Computing 23, no. 3 (2019): 715-734.

**Author’s Biography**

|  |  |
| --- | --- |
| **Description: F:\Teaching\VIGNAN'S\VIGNAN'S Research\Single author\T1\Cluster Computing\Dr-SBK-Photo-min.jpg** | **Dr. Sivadi Balakrishna** is currently working as an Associate Professor in the Department of Computer Science and Engineering at Vignan’s Foundation for Science, Technology & Research (VFSTR). He has pursued Full-time Ph.D. in the Department of Computer Science and Engineering, Pondicherry University (A Central University) in 2020, Puducherry, India. He received his Bachler of Technology (B.Tech) in the Department of Computer Science and Engineering from Jawaharlal Nehru Technological University (JNTU) in 2010 and Master of Technology (M.Tech) in the Department of Computer Science and Engineering from Jawaharlal Nehru Technological University (JNTU) in 2013, Kakinada, AP, India. He has qualified NET (National Eligibility Test) in Dec-2018, which was conducted by UGC. He has published more than 20 research articles in international journals and contributed chapters to several books. He has also presented papers at several international conferences. His current research interests are Machine Learning, Computer Vision, and Artificial Intelligence. |
| Description: C:\Users\Admin\Downloads\WhatsApp Image 2022-01-05 at 9.11.02 AM.jpeg | **Yerrakula Gopi** is currently working as Lab assistant in Department of Computer Science and Engineering in Vignan’s Foundation for Science, Technology & Research (VFSTR). He received his Bachelor of Science (BSc) in the Department of Bachelor of Science in Computer from Acharya Nagarjuna University (ANU) in 2018, Guntur, AP, India. His current research interests are Machine Learning, Computer Vision and Artificial Intelligence. |