ISSN No:-2456-2165

Disease Prediction using Machine Learning

Anika Shreya Pawar
Dept. of CSE, BNMITBangalore, India

Abstract:- Machine learning (ML) is an artificial intelligence (AI) technique that facilitates the improvement of predictability in software applications without requiring explicit programming. Data from the past is used to predict new outcomes using machine learning algorithms. During the course of this paper, we used four machine learning algorithms: Logistic Regression, Support Vector Machine, Decision Tree, and Gradient Boosting. Our chosen algorithms were applied to five datasets from the healthcare domain, in which we were able to predict kidney disease, liver disease, breast cancer predictions, heart diseases, and diabetes predictions.

Keywords:- *Machine Learning; kidney; heart; liver; breast- cancer; diabetes; svm; linear regression; decision tree; gradient boosting.*

I. INTRODUCTION

Machine learning can be defined as the use of statistical models and probabilistic algorithms to answer questions so we can make informative decisions based on our data. Machines make all these things possible by filtering useful pieces of information and piecing them together based on patterns to get accurate results. The different machine learning algorithms are supervised, unsupervised and reinforcement learning algorithms. In supervised learning, we use known or labeled data for the training data. Since the data is known, the learning is, therefore, supervised, i.e., directed into successful execution. Two types of supervised learning algorithms are regression and classification. Examples of this division of algorithm include decision tree, linear regression logistic regression, and support vector machine, just to name a few. The second type of machine learning algorithm is the unsupervised machine learning algorithm. In unsupervised learning, the training data is unknown and unlabeled - meaning that no one has looked at the data before. Clustering is a type of unsupervised learning algorithm. Examples of this division of algorithm include singular value decomposition, K-means clustering, Apriori, Hierarchical clustering, Principal component analysis, just to name a few. The third type of machine learning algorithm is the reinforcement learning. The algorithm discovers data through a process of trial and error and then decides what action results in higher rewards. The agent, the environment, and the actions are the three major components that make up reinforcement learning. Examples of this division of algorithm include Q-learning, R-learning and TD-learning.

Machine learning helps the healthcare specialists like doctors and physicians to identify the healthcare needs and solutions much faster and more accurately. Such early and accurate predictions help in making the health of individual much better. Some of the domains of healthcare where machine learning is being used are cancer detection, laser treatment, heart disease prediction, liver disease prediction, kidney disease prediction and whether the person is diabetic or not.

II. DATASET

In this paper, five datasets are being used, each for one disease prediction. These datasets are available with no cost on the Kaggle website.

A. Liver Disease Prediction Dataset

The liver dataset contains 11 attributes and 583 rows of data. This dataset contains column like the gender of the patient, their age and other attributes like insulin level, protein intake just to name a few. To predict whether the patient has a liver disease or no there is a column attribute with the title ‘Dataset’. This attribute is used in the prediction process.

B. Diabetes Prediction Dataset

The diabetes dataset contains 9 attributes and 768 rows of data. This dataset contains column like pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, age and outcome. To predict whether the patient has diabetes or no there is a column attribute with the title ‘outcome’.

C. Kidney Disease Prediction Dataset

The kidney dataset contains 26 attributes and 400 rows of data. This dataset contains column like age, blood pressure, sugar levels, red blood cells, just to name a few. To predict whether the patient has a kidney disease or no there is a column attribute with the title ‘class’, which defines the class of the disease.

D. Breast Cancer Prediction Dataset

The breast cancer dataset contains 33 attributes and 570 rows of data. This dataset contains column that describe the size, shape and location of the cancer. To predict whether the patient is suffering from breast cancer or no there is a column attribute with the title ‘diagnosis’, which describes whether the cancer is benign or malignant.

E. Heart Disease Prediction Dataset

The heart dataset contains 14 attributes and 303 rows of data. This dataset contains column like age, gender, chest pain type, etc., just to name a few. To predict whether the patient has a heart disease or no there is a column attribute with the title ‘output’.

ISSN No:-2456-2165

III. METHODOLOGY

We proposed a methodology that contains the following steps data collection, data pre-processing, exploratory data analysis, feature selection, model selection and prediction. First step is data collection. In this step, the five different datasets are collected and then the data is loaded. The second step is data pre-processing step. In this step, the data that was collected is pre-processed like checking for null values, checking for duplicate values and removing them and dropping the unnecessary columns. The third step is exploratory data analysis; basically it refers to understanding the relationship between different columns of the dataset. Nowfourth step is feature selection, it refers to the variable or attributes selection. In this step, after selecting the most significant features we split the dataset into test and train data. The next step is model selection. In this step, we select a model or a group of models that will predict the disease for us. The last step is prediction. Prediction refers to predict the output after the model has been trained on the trained data and apply to the test data to observe the prediction. The figure (figure 1) below describes the architecture of the proposed system.

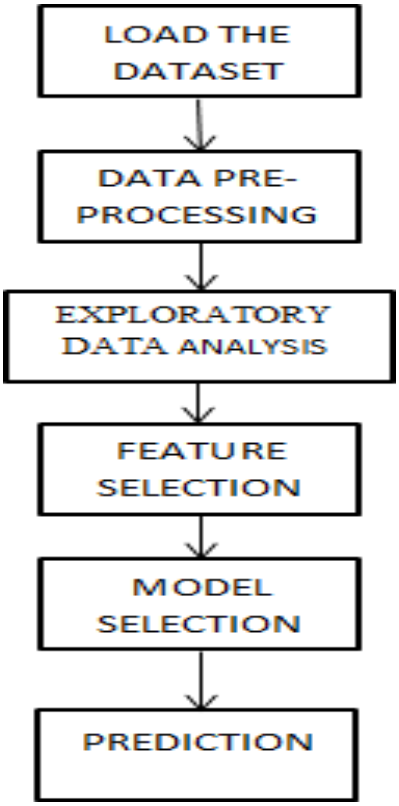

Fig. 1: System Architecture

A. *Logistic Regression*

Logistic regression is one of the most popular machine learning algorithms. It is a type of supervised learning algorithm. It is used for predicting dependent variables using aset of independent variables. It predicts the output of categorical dependent data. These categorical values could be 'Yes' or 'no', or 'true' or 'false', or a probabilistic value between 0 and 1. Logistic regression is a classification based algorithm.

B. *Support Vector Machine*

Support Vector Machine (SVM) is one of the types of supervised machine learning algorithm. It can be used for both regression and classification problems, but it is primarily used for classification problems in machine learning. In SVM algorithms, the main goal is to create the best line or decision boundary that will segregate n-dimensional space into classes so that it is easy to categorize a new data point. The best decision boundary is called a hyperplane. Some domains where SVM can be applied are image classification, text categorization and face detection.

C. *Decision Tree Classifier*

Decision Tree is a supervised learning algorithm which is preferably used for classification problems when compared to regression problems. It has a tree like structure which follows the top-down approach. This decision tree consists of internals nodes which represent the features of the dataset, each leaf node represent a particular outcome and branches of the tree which represents the decision or rules. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into sub trees.

D. *Gradient Boosting Classifier*

Gradient boosting classifier is a supervised learning algorithm that is used to combine multiple weak models which will help in strengthening the learner. One of the main advantages of gradient boosting machines comes from the fact that they can be used on more on multi-classification problems than on binary classification problems.

IV. EXPERIMENTAL RESULTS

We applied Logistic Regression, Support Vector Machine, Decision Tree and Gradient Boosting Classifier on all the five chosen datasets of the liver, kidney, heart, breast-cancer and diabetes. The following results were observed individually:

A. *Experimental results for the Breast Cancer Prediction*

SL NO.	MODEL	SCORE
1.	Logistic Regression	95.91
2.	Support Vector Machine	97.66
3.	Decision Tree Classifier	91.81
4.	Gradient Boosting Classifier	97.66

Table 1: Results for Breast Cancer Prediction

As shown above, Gradient boosting classifier and supportvector machine showed the highest accuracy score.

ISSN No:-2456-2165

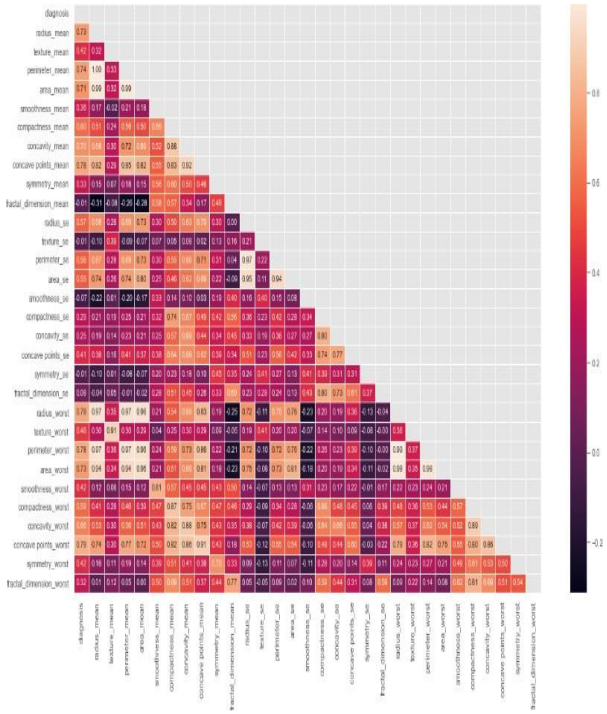

Fig. 2: Correlation Matrix of Breast Cancer

B. Experimental results for the Heart Disease Prediction

SL NO.	MODEL	SCORE
1.	Logistic Regression	80.22
2.	Support Vector Machine	51.65
3.	Decision Tree Classifier	76.92
4.	Gradient Boosting Classifier	79.12

Table 2: Results for heart disease prediction

As shown above, Logistic Regression algorithm showed thehighest accuracy.

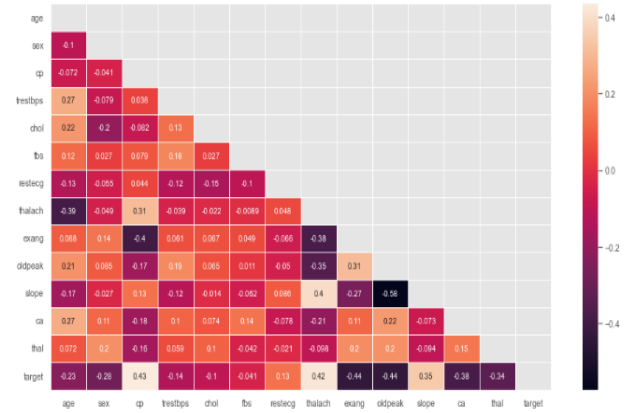

Fig. 3: Correlation Matrix of Heart disease

C. Experimental results for the Kidney Disease Prediction

SL NO.	MODEL	SCORE
1.	Logistic Regression	90.00
2.	Support Vector Machine	72.50
3.	Decision Tree Classifier	97.50
4.	Gradient Boosting Classifier	97.50

Table 3: Results for Kidney disease prediction

As shown above, Decision Tree Classifier and Gradient Boosting Classifier showed the highest accuracy.

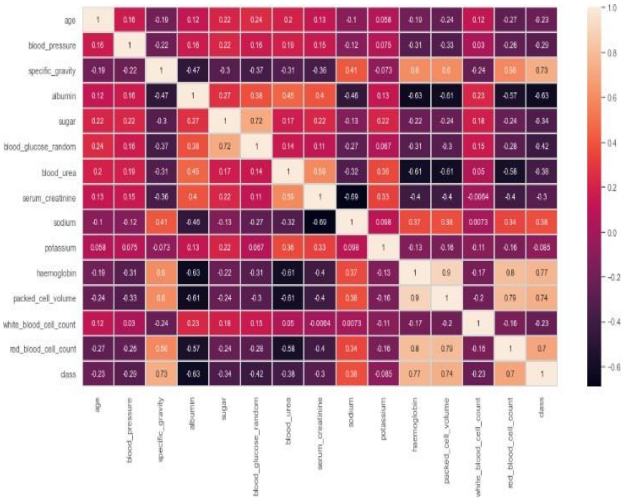

Fig. 4: Correlation Matrix of Kidney Disease

D. Experimental results for the Diabetes Prediction

SL NO.	MODEL	SCORE
1.	Logistic Regression	81.14
2.	Support Vector Machine	83.33
3.	Decision Tree Classifier	82.89
4.	Gradient Boosting Classifier	86.84

Table 4: Results for Diabetes prediction

As shown above, Gradient Boosting Classifier showed thehighest accuracy.

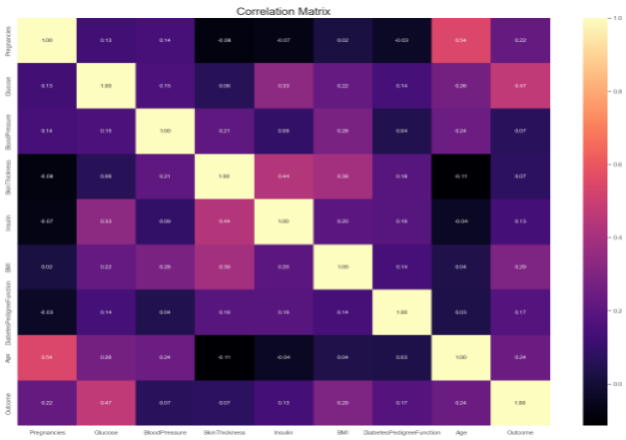

Figure 5: Correlation Matrix of Diabetes

E. Experimental results for the Liver Disease Prediction

SL NO.	MODEL	SCORE
1.	Logistic Regression	69.41
2.	Support Vector Machine	71.18
3.	Decision Tree Classifier	62.94
4.	Gradient Boosting Classifier	71.18

Table 5: Results for liver disease prediction

ISSN No:-2456-2165

As shown above, Support Vector Machine and Gradient Boosting Classifier showed the highest accuracy.

Fig. 6: Correlation Matrix of Liver disease

F. Comparing the results Disease Prediction

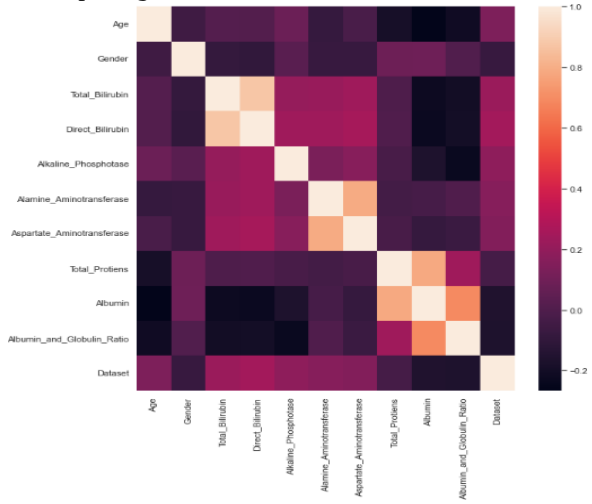

Table 5: Comparitive results for disease prediction

DISEASES	Liver	Diabetes	Breast Cancer	Heart	Kidney
MODELS					
Logistic Regression	69.41	81.14	95.91	80.22	90.00
Support Vector Machine	71.18	83.33	97.66	51.65	72.50
Decision Tree Classifier	62.94	82.89	91.81	76.92	97.50
Gradient Boosting Classifier	71.18	86.84	97.66	79.12	97.50

V. CONCLUSION

If the healthcare related diseases are diagnosed at their early stages, this will help thousands of people to save themselves and take the right medication at an early stage to stop the disease from becoming worst. We may be capable of classifying and are expecting the most disease to be classified as either a one at an early stage or the ones at an final stages by making use of the machine learning algorithms. Machine Learning algorithms may be used for scientific orientated studies, it speed up's the system, reduces the human mistakes & decreases the guide mistakes.

VI. FUTURE ENHANCEMENT

Many more Machine Learning algorithms can be used topredict a much larger number of diseases. A variety of algorithms can be combined to get betterpredictions.

Deep Learning techniques can be made use of in the healthcare sector to help in predicting and classifying diseases.

REFERENCES

[1.] DEEPTI, S. AND DILIP S. (2018). PREDICTION OF DIABETES USING CLASSIFICATION ALGORITHMS. INTERNATIONAL CONFERENCE ON COMPUTATIONAL INTELLIGENCE AND DATA SCIENCE.

[2.] FANIQUL ISLAM M.M ET AL. (2020) 'LIKELIHOOD PREDICTION OF DIABETES AT EARLY STAGE USING DATA MINING TECHNIQUES.' COMPUTER VISION AND MACHINE INTELLIGENCE IN MEDICAL IMAGE ANALYSIS. SPRINGER, SINGAPORE, 113-125.

[3.] HASSAN A.S ET AL. (2020). DIABETES MELLITUS PREDICTION USING CLASSIFICATION TECHNIQUES. INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND EXPLORING ENGINEERING (IJITEE) ISSN: 2278-3075, VOLUME-9 ISSUE-5.

[4.] MAHBOOB T. A ET AL. (2018). A MODEL FOR EARLY PREDICTION OF DIABETES. INFORMATICS IN MEDICINE UNLOCKED.

[5.] "PERFORMANCE EVALUATION OF MACHINE LEARNING METHODS FOR BREAST CANCER PREDICTION", BY YIXUAN LI, ZIXUAN CHEN OCTOBER 18, 2018.

[6.] "BREAST CANCER PREDICTION AND DETECTION USING DATA MINING CLASSIFICATION ALGORITHMS: A COMPARATIVE STUDY" BY MUMINE KAYA KELES, FEB 2019.

[7.] "BREAST CANCER PREDICTION USING DATA MINING METHOD " BY HAIFENG WANG AND SANG WON YOON, DEPARTMENT OF SYSTEMS SCIENCE AND INDUSTRIAL ENGINEERING STATE UNIVERSITY OF NEW YORK AT BINGHAMTON BINGHAMTON, MAY 2015.

[8.] "MACHINE LEARNING WITH APPLICATIONS IN BREAST CANCER DIAGNOSIS AND PROGNOSIS" BY WENBIN YUE, ZIDONG WANG, 9 MAY 2018.

[9.] M. CHEN, Y. HAO, K. HWANG, L. WANG, AND L. WANG, "DISEASE PREDICTION USING MACHINE LEARNING OVER BIG DATA." IEEE ACCESS (VOLUME: 5), PP. 8869 – 8879.

[10.] Z. CHEN, Z. ZHANG, R. ZHU, Y. XIANG AND P. B. HARRINGTON, "DIAGNOSIS OF PATIENTS WITH CHRONIC KIDNEY DISEASE BY USING TWO FUZZY CLASSIFIERS", *CHEMOMETRICS INTELL. LAB. SYST.*, VOL. 153, PP. 140-145, APR. 2016.