International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 03 | March - 2023 Impact Factor: 7.185 ISSN: 2582-3930

Multiple Disease Prediction Using a Machine Learning Algorithm

Abhishek Gawdae

Guide: Prajakta Chowk

Keraleeya Samajam's Model College, Dombivli East, Mumbai, Maharashtra, India

abhishekgawade.model@gmail.com

Abstract - Artificial intelligence and machine learning play important roles in today's world. From Autonomous cars to the medical field, we can find them everywhere. The medical industry generates a large amount of patient data and numerous diseases that can be processed in many ways with taking help of machine learning. So, taking use of machine learning we have developed a disease prediction system that can analyze more than one disease. Currently, many prediction systems can predict one disease at a time. Those have less accuracy and predictability which affect the patient's health. We have considered two diseases for now which are diabetes and heart disease, and, in the future, more diseases will be added. The user will select the disease he/ she wants to diagnose and enter the parameter of the disease the system will provide a prediction of the disease whether he or she has a disease or not. This project will help to monitor patient health and take necessary precautions to improve their health.

Keywords: Multiple disease prediction, Diabetes, Heart, Random Forest, KNN, Logistic Regression, Decision Tree, XGBoost, Gradient-based, SVM.

Introduction:

In recent days data is an asset and enormous data generated by all the fields. Data in the healthcare industry consists of information about the patients along with their diseases. here we are general architecture has a purpose for prediction of the diseases. Many existing models are available for one disease per analysis, one for cancer, one for diabetes, and one for skin disease at the time. No common system that can analyze more than one disease at a time. Thus, we are concentrating on developing multiple disease predictions which predict disease accurately and help doctors to give precise predictions so they will start further treatment as soon as possible. So, we are developing such a system which used to predict multiple diseases by using streamlit. In this project, we will analyze diabetes and heart disease, and later on, we include more disease prediction models. For the development of the disease prediction model, we consider algorithms such as Logistic Regression, SVN, KNN, Decision Tree, Random Forest, XGBoost, and Gradient-based algorithm out of all these algorithms Random Forest gives more accuracy as well as precise prediction. Python pickle library is used to save the behavior of the model. For analyzing disease using all the parameters that affect the disease will be included so it's possible to detect disease efficiently and accurately. final behavior of the model will be saved as a python pickle file.

Problem definition:

As we studied there are numerous machine learning models available for healthcare analysis focused on one disease. For example, first, for liver disease analysis, one for diabetes, one for liver, etc. if the user wants to detect multiple diseases he will go through with all other sites and do analysis, there is no common system where more than one disease can predict. Some of the models have low accuracy which affects patients' health.

Purposed system:

In multiple disease prediction, we are developing a system where we can predict more than one disease at a time. so the user does not need to traverse multiple sites for disease prediction. We are taking diseases that are diabetes and heart disease. To implement multiple disease analysis, we are using a Random Forest machine learning algorithm with the help of streamlit and pickle. On streamlit there are multiple disease model is available, the user accessing the model user send the parameter of disease. stremlit will invoke that corresponding model and provide the status of the patient.

Research objective :

- Point of review which offers what exactly we are doing in this research:
- Collecting data sets of corresponding diseases and processing the data using a machine learning algorithm.
- To study test and train the accuracy of each algorithm and found precise and accurate algorithms.
- Implementation of and deployment of the precise algorithm using streamlit.

- Testing the final predictability of the implemented model.

Literature review:

There is an existing exploration that has been conducted in this field which helps us to improve the project. In this section, we elaborate on recent studies and research on new technology. They emphasize disease prediction using machine learning in the field of medical diagnosis.

[1] disease prediction from various symptoms using machine learning, 27 July 2020

This paper mainly focuses on the development of a system that is capable of medical diagnosis using machine learning algorithms. This system will predict more than 230 diseases based on the symptoms and parameters of the diseases such as headache, chest pain, and many more will cause of the disease. For designing the disease prediction app researcher will study algorithms such as KNN, Gaussian, kernel naive Bayes, Weighed KNN, and Decision tree. All models have good accuracy among all this algorithm weighted KNN model has 93.5% accuracy. This is a preliminary disease prediction app that predicts disease based on symptoms after getting positive result patient may consult with doctors.

[2] disease prediction using machine learning, December 2020

The researcher will analyze different algorithms used in different disease predictions such as heart, kidney, breast, brain, etc. researcher found that SVM, Random Forest, and Linear regression, algorithms were mostly used in prediction, and accuracy was the most used performance metric. CNN model will be most adequate for common diseases furthermore SVM has superiority in the classification of the disease it will scale the large data set and avoids overfitting. Linear regression is reliable for heart disease prediction. Researcher says that we need to create a more complex algorithm to increase the efficiency of disease prediction. They suggest dataset should be expanded on multiple dimensions to avoid overfitting and increase the accuracy of the model and relevant feature selection will enhance the performance of the model.

[3] Cancer Prediction using Machine Learning Algorithm, August 2020

This research mainly focuses on developing the model for predicting cancer using a support vector machine algorithm and comparing the accuracy of different algorithms by using data imported from the sci-kit-learn library. Cancer is a heterogeneous disease consisting of many subtypes, that's why it will be predicted in an earlier stage as the most important task which helps patients to get appropriate treatment on time. In this research, they discuss various supervised learning algorithms such as K-NN, SVM, LR, and NB and evaluate accuracy, precision, and recall using machine learning. Researchers will observe that the K-NN algorithm gives the best result with maximum accuracy.

[4] Prediction of Heart Disease Using Machine Learning algorithm, 13 March 2022

This study represents the methodology of prediction of heart disease with the highest accuracy because a minor error will result in death. It is important to detect heart disease at an earlier stage so that the patient gets the appropriate treatment that can save his life. That's why researchers will develop a heart disease prediction system that can efficiently and accurately predict heart disease. This study will compare K-NN, Logistic regression, and Random Forest for heart disease prediction using machine learning. In this study, the Logistic Regression algorithm is the most efficient algorithm with 89% accuracy for heart disease prediction.

[5] Disease Prediction Application Using Machine Learning, March 2022

This research mainly focuses on the development of a disease prediction system that can predict disease based on machine learning algorithms that use collected patient data from hospital databases. In this research, the researcher will design a disease prediction system that uses different types of machine learning algorithms for prediction. The system will use Logistic regression and Random Forest algorithm for breast cancer, heart disease, and diabetes disease prediction. Whether the patient has a disease, then the proposed system will suggest the hospital and doctor for his or her treatment for the desired disease.

Functional requirement:

In our multiple disease application have multiple diseases, the patient will select the disease which he wants to diagnose. The user will provide a parameter of the symptoms based on this parameter model will predict the disease and suggest the doctor for further treatment. Disease prediction application is extensive and significant. The project functionality is divided into the following layers:

Data collection

In this stage, we gather information from dependable sources. Thus, we collect data on each disease from Kaggle.com. The dataset was selected based on its attributes.

Data preparation or exploratory data analysis

This is an important step before the data process. To check the data format sometimes we need to reformat it, make corrections in the data, and combine the data to get more data for testing, and training as well.

Data cleaning

In the data cleaning process of removing irrelevant, corrupted, duplicate, and incomplete data from the data set. If the data is irrelevant, the predicted disease will be unreliable.

Loading, splitting training, and testing data.

This is a process where we can split the data set into testing and training, random state will separate the data randomly train and test data.

Feature selection

Feature selection is the final step of data processing, in this stage, we can standardize the independent variable of the dataset in the specified range. In feature scaling, we put the variable in the same scale so that no other variable cannot influence the other variable.

Model building

In this stage preparing the model by a random forest algorithm on training data that has been trained to find a hidden pattern from the dataset.

Model Training.

In this phase, we train the model as a dataset that is used to train the machine learning algorithm. Here we map the output data to get the result from the corresponding input data.

Made prediction.

Prediction refers to the output generated from the model, it has been trained on the historical dataset and the model will be applied to new or untrained data sets for predicting the possibilities of the disease.

Model evaluation.

In this phase, we test our model based on currently predicted values with the actual values that use performance matrices. The accuracy score and confusion matrix is used to survey the execution. When new or test datasets are given into the model the model for each new data will predict accurate disease.

Model deployment

When the desired output is generated then the model will be deployed with the help of a web application.

Methodology

In this research, we analyze the accuracy and predictability of different machine learning algorithms used for the development of multiple disease algorithms individually. The goal of this paper is to find a model that can accurately predict the disease. Data is analyzed in Kaggle notebook, which is an open-source platform where a large number of datasets are available, we can apply various libraries in python for data preparation. Applying a machine learning algorithm and testing the accuracy of all models select the most accurate and precise algorithm, prepare the model deployed it using streamlit.

Data collection:

We are collecting datasets through Kaggle, a variety of data available on this platform. Here we are collecting two data sets first is of diabetes disease and another is heart disease.

Feature selection:

For diabetes

Sr.no	feature
1	Pregnancies
2	Glucose
3	Blood Pressure
4	Skin Thickness
5	Insulin
6	BMI
7	Age

For heart disease

Sr.no	feature
1	Age
2	Sex
3	Chest Pain(cp)
4	Blood Pressure(trtbps)
5	Cholesterol
6	Fasting blood sugar (fbs)
7	Resting electrocardiograph (restecg)
8	Thalachh (Maximum heart rate achieved)
9	Exng (exercise include angina)

10	oldpeak (ST depression induced by exercise realt ive rest)
11	Slp (slope of exercise peack)
12	Caa(number of vessels)
13	Thall(reversible defect)

Implementation

1. **Support vector machine**

Support vector machine is one of the supervised learning algorithms which is used for classification as well as regression. The goal of SVM is to create a decision boundary that can segregate n dimension space into classes so that we can easily put the new data point incorrect category. The decision boundary is called a hyperplane. SVM chooses extreme points that help in creating hyperplanes these extremes are called support vectors.

2. **KNN Algorithm**

Working of the KNN algorithm as followed:

Step 1: Select the number K of neighbors.

Step 2: calculate Euclidian distance of K number of neighbors.

Step 3: among these K neighbors, count the number of the data point in each of the categories.

Step 4: assign a new data point to the category which has a maximum of k neighbors.

Step 5: the model is ready.

3. **Decision tree**

Working of decision tree algorithm as followed:

Step 1: root node X contains the complete dataset.

Step 2: Find the best attribute using the Attribute selection measure.

Step 3: divide S into subsets containing possible values for best attributes.

Step 4: generate the decision tree node containing the best attribute.

Step 5: recursively make a new decision tree using a subset of the dataset created in step 3 continue the process until the stage reach where you can not classify the node called the final node.

4. **Random forest algorithm**

Step 1: select K random data point from the training set.

Step 1: build a decision tree along with selected data points.

Step 3: choose the n number of a decision tree that you want to build.

Step 4: repeat steps 1 and 2.

5. **XGBoost classification algorithm:**

Step 1: Make an first prediction and calculate residuals.

Step 2: Build an XGBoost tree.

Step 3: prune the tree.

Step 4: calculate the output values of leaves.

Step 5: Make a new prediction.

Step 6: calculate residuals using the new prediction.

Step 7: Repeat steps 2-6.

6. **Gradient Boosting algorithm:**

Step 1: build a base model to predict the observation within the training set.

Step 2: calculate pseudo residuals.

Step 3: Build a model on pseudo residual and makes predictions.

Step 4: find values for each leaf of the decision tree.

Step 5: update the prediction of the previous model.

Result

In a multiple disease prediction system, we used a random forest algorithm for diabetes prediction as well as heart disease prediction. This model will deploy using streamlit, the user will add a parameter according to the disease it will show whether the patient has the disease or not according to the disease selected.

Accuracy of models for diabetes:

	Model	Accuracy Score
4	Random Forest	80.52
2	SVM	80.09
0	Logestic Regression	79.65
6	Gradient Boosting Clasifier	78.79
5	XGBoost	75.32
1	KNN	72.73
3	Decision Tree	71.86

Model Accuracy for diabetes

Accuracy of models for heart disease:

	Model	Accuracy Score
4	Random Forest	0.802198
5	Gradient Boosting Classifier	0.791209
1	KNN	0.780220
6	XgBoost	0.780220
0	Logistic Regression	0.758242
2	SVM	0.758242
3	Decision Tree	0.758242

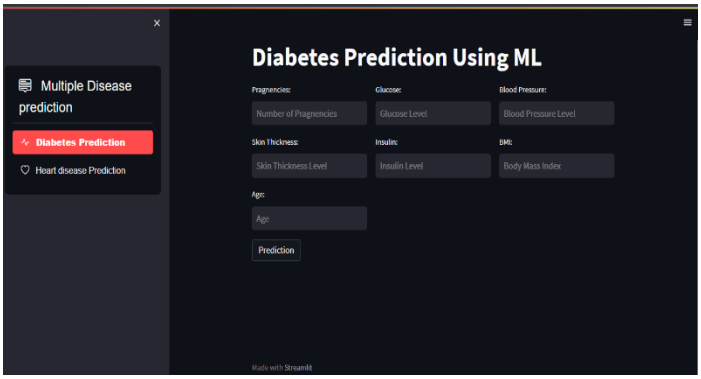
Confusion matrix and classification report of diabetes:

```
The accuracy of Randomforest train : 0.8752327746741154

The accuracy of Randomforest test : 0.8051948051948052

The Confusion Matrix: [[132  18]
[ 27  54]]

The Classification Report:              precision    recall  f1-score   support

           0       0.83      0.88      0.85       150
           1       0.75      0.67      0.71        81

    accuracy                           0.81       231
   macro avg       0.79      0.77      0.78       231
weighted avg       0.80      0.81      0.80       231
```

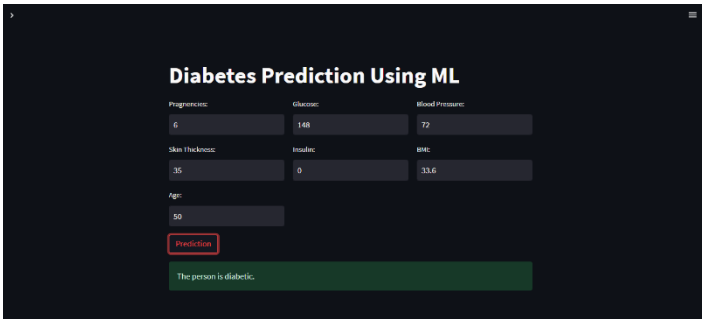
Confusion matrix and classification report of heart disease:

```
Train RF accuracy: 0.9669811320754716
test RF accuracy: 0.8021978021978022

 The confusion matric : [[ 93   4]
 [  3 112]]

 The classsification report:              precision    recall  f1-score   support

           0       0.83      0.71      0.76        41
           1       0.79      0.88      0.83        50

    accuracy                           0.80        91
   macro avg       0.81      0.79      0.80        91
weighted avg       0.81      0.80      0.80        91
```

Main UI interface:



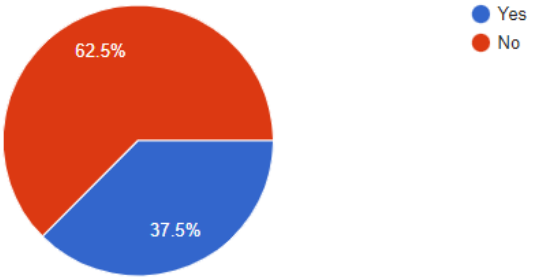
Diabetes Prediction result:



Heart disease prediction:

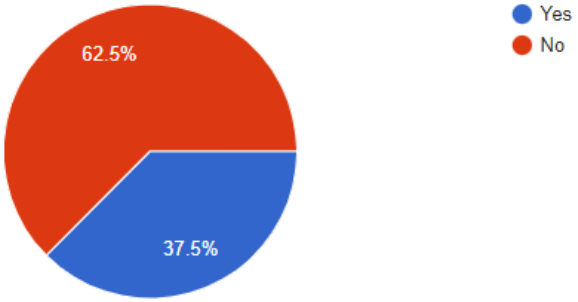
Questionary for user review and result:

1. Did you use a disease prediction web application for the prediction of disease?



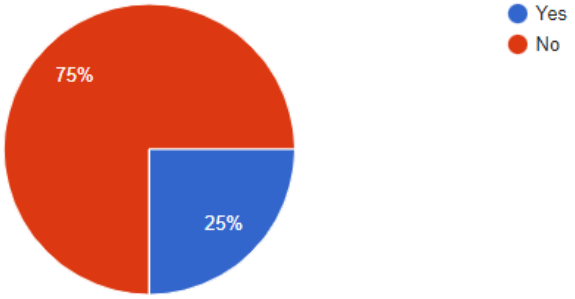
37.5 % of respondents use disease prediction applications for prediction of disease and 62.5% are not use such applications.

2. will those applications predict diseases accurately?



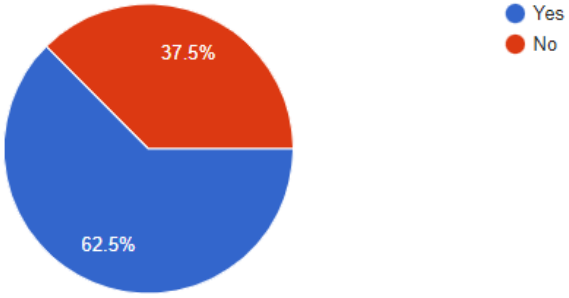
37.5 % of respondents agreed that that application will predict disease accurately and 37.5% will not agree to it.

3. Do you know any program that automates this process?



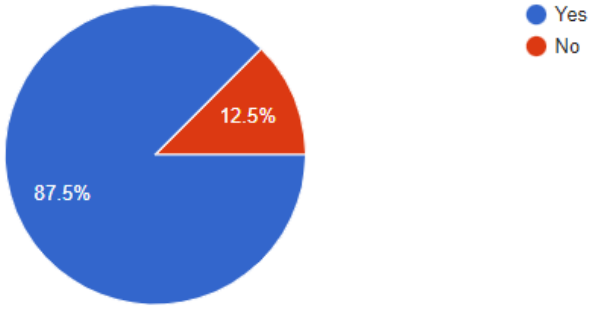
75% of respondents don't know any program that automates this process and 25% of respondents know some program that automates the process of disease prediction.

4. well, is such automation being implemented in healthcare?



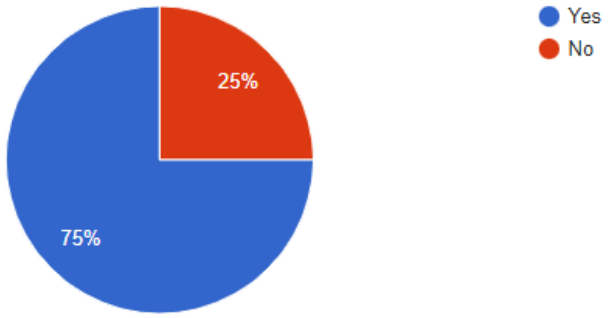
62.5% of respondents said that such automation will be implemented in healthcare and 37.5% don't.

5. Do you believe, these disease prediction applications would be useful in day-to-day life?



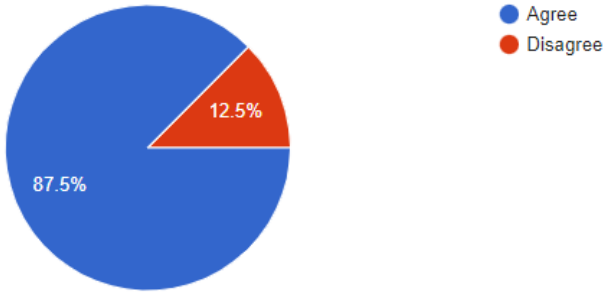
87.5% of respondents agreed that such an application will be helpful in day-to-day life and 12.5% of the respondent will not be agreed.

6. Does, it functions well for multiple diseases?



75% of respondents agree multiple algorithms will function well for multiple diseases and 25% of respondents disagreed.

7. Will these apps help doctors improve patient health?



87.5% of respondents agreed that it will help doctors to improve patient health. 12.5% of the respondent not.

Conclusion:

The main objective of this research is to create a system that would predict more than one disease along with accuracy and precision in the prediction of disease. By using this application user would not need to traverse multiple sites he can find multiple diseases for prediction on a single site. The disease will predict early so treatment will start in the earlier stages of the disease. For this project, we have studied and applied multiple machine learning algorithms and finally, we used the random forest algorithm which achieved max accuracy among them.

References:

[1] Diabetes Prediction Using Different Machine Learning Approaches, 29 March 2019

Priyanka sonar, K. Jaya Malini

Mumbai University, Mumbai, India.

https://ieeexplore.ieee.org/document/8819841/authors

[2] Disease prediction from various symptoms using machine learning, 27 July 2020

Rinkal Keniya, Aman Khakharia, Vruddhi Shah, Vrushabh Gada, Ruchi Manjalkar, Tirth Thaker, Mahesh Warag, Ninad Mehendale.

K.J. Somaiya College of Engineering

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

[3] Cancer Prediction using Machine Learning Algorithm, August 2020

Mohit Agrawal

Dept. of C.S.E & Eng at IMS Eng College Ghaziabad, UP, India

https://www.ijsr.net/archive/v9i8/SR20801174131.pdf

[4] Prediction of Heart Disease Using Machine Learning algorithm, 13 March 2022

Shriniket Dixit, Pilla Vaishno Mohan, Shrishail Ravi Terni

Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India https://doi.org/10.22214/ijraset.2022.40768

[5] Disease Prediction Application Using Machine Learning, March 2022

Arnab Das, A. Udith Sai, P. Asha

Department of Computer Science, Sathyabama Institute of Science and Technology, Chennai, India

https://www.ijraset.com/best-journal/disease-prediction-application-using-machine-learning