Binay P, 2025, 13:5
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Multiple Disease Prediction using Machine Learning Algorithm

Binay P, Anil H, Ankith G, Tanya B, Prof. Arathi H L
Department of Computer Science and Engineering, East West College of Engineering

Abstract - Machine learning techniques like Logistic Regression, The Support Vector Machine i.e. (SVM) classifiers, The Random Forest classifiers i.e. (RFC), The Decision Tree classifiers i.e. (DTC), and K-Nearest Neighbor (KNN), as well as basic metrics like heart rate, blood pressure, cholesterol, and pulse rate, the goal of this project is to forecast the occurrence of various diseases like diabetes, heart disease, and Parkinson's disease. The most accurate calculation is used to train the dataset, while Python pickling and streamlit are used to record the model behavior. By entering pertinent disease- related information, the initiative seeks to determine the risk factors for the diseases and provide users a prognosis of whether they have the condition or not. This program can assist people in keeping an eye on their health and taking the necessary actions to prolong.

Keyword - Parkinson's, diabetes, heart disease, pickle, streamlit, logistic regression, random forest, decision tree, KNN, and SVM.

I. INTRODUCTION

Early identification, prevention, and effective care of illnesses including diabetes, heart disease, and Parkinson's disease are urgently needed on a global scale. Machine learning models may be trained to predict a person's likelihood of contracting a certain disease based on their demographics and symptoms. But a user-friendly user interface is required if these models are to be used as web apps. In this situation, Streamlit is useful.

In this paper, we detail the entire process of developing a web app that can predict multiple diseases using machine learning models trained on real-world datasets. We used the pickle module to save our trained models and Streamlit to create the user interface for the web app.

Background:
ML methods including LR, SVM, RFC, KNN and DTC have been used to construct disease prediction models. These models forecast the risk that a person would acquire a certain illness using several input features, including symptoms and measurements of the heart rate, blood pressure, cholesterol, and other body functions. But in order to implement these models as a web application, it is necessary to have a user- friendly interface that can manage user inputs, process them, and deliver useful results. We used Streamlit, a potent tool that makes it simple to construct interactive web apps using Python, to create our web app for illness prediction

II. METHODOLOGY

We developed our web app using Python, the Streamlit framework, and the pickle module. We collected real- world datasets like diabetes.csv [15], heart.csv [16], parkinsons.csv [17] for diabetes, heart disease, and Parkinson's disease, which were used to train our machine learning models.

The best-performing models were chosen for deployment in the web app after being trained using a variety of ML methods, including LR, SVM, RFC, KNN and DTC. The pickle module was used to store the trained models, making it simple to import them into the web interface.

Overall, our study shows how machine learning, pickle, and Streamlit can be used to create a user-friendly online tool that can predict a variety of illnesses. In order to improve individual health outcomes, we expect that this software will be

helpful in the early diagnosis and management of various conditions.

III. LITERATURE SURVEY

The development of a model for predicting diabetes mellitus sickness using the LGBM algorithm is described in this paper by B. Shamreen Ahamed. The dataset might be further enhanced by using cutting-edge methods like transformer-based learning, according to the authors, who also suggest that alternative attribute combinations might be utilized for identification. Additionally, they suggest that the classifiers might be modified to better precisely predict the condition and that the likelihood that the disease will manifest itself could be calculated to further improve the model's accuracy. Overall, the research raises the possibility that these developments might lead to a model that is better able to forecast the onset of diabetes mellitus in people who already have it. [1]

In this paper, Aishwarya Mujumdar applies a range of machine learning approaches to a dataset for diabetes prediction and discovers that Logistic Regression has the highest accuracy (96%). The accuracy is raised to 98.8% by employing a pipeline, with the AdaBoost classifier acting as the ideal model. By comparing the performance of machine learning approaches with two independent datasets, the study shows that the model improves the accuracy and precision of diabetes prediction compared to current datasets. The authors suggest that further study be done to determine the likelihood that non-diabetic people may develop diabetes in the next years. [2]

In this paper, M.K. Hassan presents a novel technique to predict type-II diabetes mellitus in this research. They collected precise lifestyle data from 1939 patients from a range of demographic groups. Make advantage of exploratory data analysis to improve the assessment of the dataset's quality. They used ensemble learning techniques (Bagging, Boosting, and Voting) for the prediction. We were able to get the highest accuracy rate possible using a bagged decision tree, which was 99.14%. [3]

The algorithms KNN and Random Forest are used in this study by Apurva Garg et al. to predict heart illness. Using the UCI dataset from Kaggle, the data is investigated for relationships between qualities and the goal value. It has been demonstrated that the desired trait positively correlates with maximum heart rate and chest pain. The dataset was divided in half 80:20 for training and testing, and the accuracy of the KNN and Random Forest models, respectively, was 86.885% and 81.967%. [4]

The accuracy of a classification model may be increased by carefully managing null values and feature selection, as A.B. Nassif shows in this work. According to the authors' evaluation of classification models, the Random Forest classifier performs best overall with a 95.63% accuracy rate and favorable recall, F1-score, and ROC values. They suggest adopting new classification algorithms with improved feature selection approaches in future studies and predict that the null values and low contribution features in the dataset may be the reason for the poor classification performance. [5]
Author Rindhe's goal in this research is to shed light on machine learning techniques for categorizing cardiac diseases, which is important for deciding how to treat patients. The study's objective is to develop accurate and reliable techniques for assessing cardiovascular risk and benefiting from preemptive treatment. The authors conclude that while each method performs well in certain circumstances, they may not perform well in others, and that machine learning algorithms have a great deal of potential for predicting heart-related illnesses. [6]

In this paper, Author Avi used a dataset from the UCI machine learning library to examine the precision of the DTC, LR, RF, and Naive Bayes algorithms for predicting heart disease. With an accuracy score of 90.16%, the Random Forest algorithm was deemed to be the most effective. According to the authors, future research may develop a web application based on the Random Forest algorithm and utilize a larger dataset to obtain even better results and assist medical practitioners in properly forecasting cardiac problems. [7]

Nayab focused on using data mining techniques to discover heart disease in this study's healthcare system. The authors evaluated the performance of different algorithms, including KNN, Neural Networks, DTC, Naive Bayes, and RFC, using measures such as accuracy, true negative (TN), false positive (FP), true positive (TP), and false positive (TP) rates. The study increased the detection of cardiac illness by using data mining techniques. [8]
In this paper, Harshavardhan Tiwari analyses the use of six classification algorithms to the acquired data set in this research. An algorithm like LR, SVM, DTC, KNN, and XGBOOST (Extreme gradient boosting) is used to assess whether the person is healthy or suffering from Parkinson disease based on the features of the voice input. The results were compared after that. In overall, this study is in favour of using these models to identify Parkinson's disease. There are several ways to diagnose a condition, thus there is room for technical advancement. [9]

In this paper, author Basil K. Varghese's goal in this research is to examine several machine learning models for estimating the severity of Parkinson's disease and creating a precise model to identify the condition sooner. The Parkinson's Telemonitoring dataset from the UCIML repository will be used in the investigation. By making an accurate diagnosis of Parkinson's disease earlier on, researchers hope to help doctors treat and rehabilitate people with the condition. [10]

In this paper, Shivchitra M. used a newly created network called the Mc-FCRBF (Meta- cognitive Fully Complex- valued Radial Basis Function) to forecast Parkinson's disease. In order to predict Parkinson's disease, the performance of the Mc-FCRBF network was compared to that of an FC-RBF network and a real-valued extreme learning machine. The Mc-FCRBF network beat the competition, and the increase in effectiveness was attributable to the metacognitive component's self-regulatory learning process. [11]

In this paper, author Y.R.K. Reddy A feature selection approach is employed in this study info- gain to improve the classification model's accuracy. Logistic regression gave the best accuracy, which is 92.76%. MLP and SVM were utilized to further improve the accuracy., Nahiduzzaman et al. [12]

In this paper, Shikha Singh used a neural network with Long-Short Term Memory (LSTM) architecture to suggest a method for the early identification of Parkinson's disease. The system focuses on extracting temporal patterns from the gait cycle and analyses changing vertical ground reaction force (vGRF) using information from 16 sensors placed in the soles of each subject's shoes. The method successfully detects patients with Parkinson's disease with a detection accuracy of 95.2% by utilizing LSTM networks. This degree of precision is equivalent to current techniques and shows the promise of gait analysis and LSTM-based neural networks for early Parkinson's disease identification. [14]

motivation

A single analysis cannot forecast more than one disease using a same system. Some of the models' lesser accuracy can have a significant negative impact on patients' health. An organization must deploy numerous models in order to analyze the health reports of its patients, which in turn increases the time and cost involved. Some of the systems now in use take very little into account, which can lead to inaccurate findings.

The goal of advancing machine learning might be another driver for this study. The understanding of the potential applications of machine learning in healthcare can be advanced by creating precise illness prediction models utilizing methods like Streamlit and Pickle.

Proposed System

Since diabetes, Parkinson's, and heart disease are all interrelated, we are concentrating on these three conditions. In order to provide multiple disease prediction, which saves the user from having to move between several sites, we need to be able to anticipate more than one ailment at a time.

System Design

Fig. 1. Architecture Diagram

Our research has focused on the three diseases Parkinson's, diabetes, and heart disease because of their connections. The beginning of each of these illnesses' unique datasets was signaled by the import of the datasets for diabetes, heart disease, and Parkinson's disease, in that order. Each piece of input data is visualized once the dataset has been loaded. After pre-processing, which entails looking for outliers, missing values, and scaling the dataset, data is divided into training and testing on the revised dataset. We next used the SVM, LR, RFC, DTC, and KNN algorithm to the training dataset in order to put what we had learned about the classification approach into practice. The approach with the best degree of accuracy for each of the inputs will be selected after knowledge application.

Implementation of Algorithm Used

Support Vector Machine (SVM) Algorithm:

For classification and regression analysis, the supervised learning technique Support Vector Machine (SVM) is used. Due to its robustness and adaptability, SVM is a machine learning model that excels in high- dimensional domains. The fundamental aim of SVM is to find a hyperplane that partitions the input data into discrete classes with the biggest practicable margin. The margin is the separation between each class's closest data points or support vectors and the hyperplane. In a binary classification problem, SVM looks.

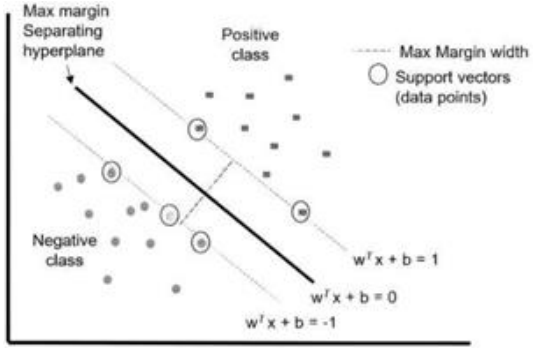

Fig. 2. Support Vector Machine Algorithm (SVM)

for the hyperplane with the largest difference between the two classes. SVM employs the kernel technique to transfer the input data into a more complicated space if it cannot be broken up linearly.

Logistic Regression:

Both linear and nonlinear connections between the input variables and the output variable may be handled using logistic regression. It could also address situations when there are unaccounted for numbers or a data imbalance. However, since logistic regression assumes a linear relationship between the input and output variables, it might not work well if the relationship is significantly nonlinear.

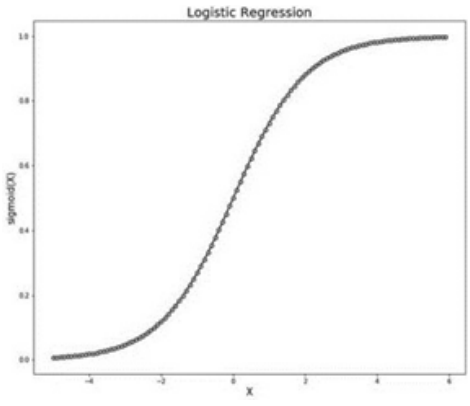

Fig. 3. Logistic Regression

In LR, a complex cost function called the sigmoid function or logistic function is employed. The sigmoid function is used in LR to model the data.

The function's equation is

$f(x) = 1/(1+ e^x)$ (1)

F(x) returns a number between 0 and 1. x = input of the function
Base of a natural logarithm equals e.

Random Forest Classifier:
A ML system called an RFC mixes different decision trees to provide predictions. A randomly chosen sample of the data and features is used to train each decision tree, which helps to decrease overfitting and boost generalization performance. Based on the results of numerous decision trees, a choice is reached. It is an effective method that can address several machine learning issues. Dealing with issue involving numerous characteristics, noisy data, or outliers benefits greatly from it.

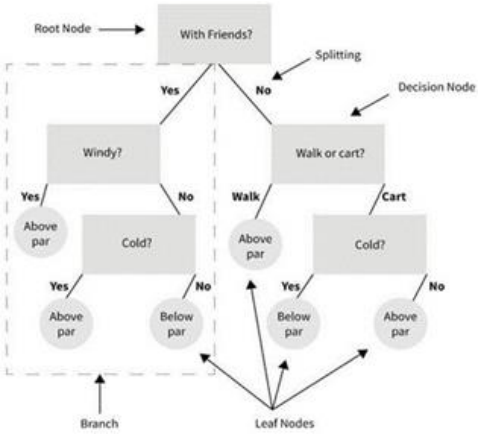

Fig. 4. Random Forest Classifier

One of the best supervised learning algorithms for both classification and regression applications is RF. Multiple decision trees are combined in the random forest regression ensemble learning technique, which predicts the outcome by averaging the output from each tree. The foundation models, which contain the formal representation, are created using the connected decision trees.

$$g(x)=f\,0(x)+f\,1(x)+f\,2(x)+\ldots \quad (2)$$

Decision Tree Classifier:
The DTC is a popular ML technique for classification issues. It functions by displaying options and their results in a tree-like layout. The decision-making process is represented by nodes in the tree structure, while potential outcomes are represented by branches. Because they can handle both category and numerical data and are simple to understand and visualize, decision tree classifiers are commonly used. They can manage missing data and outliers as well. They might, however, underperform on imbalanced datasets and overfit the training data if wrongly pruned or regularized.

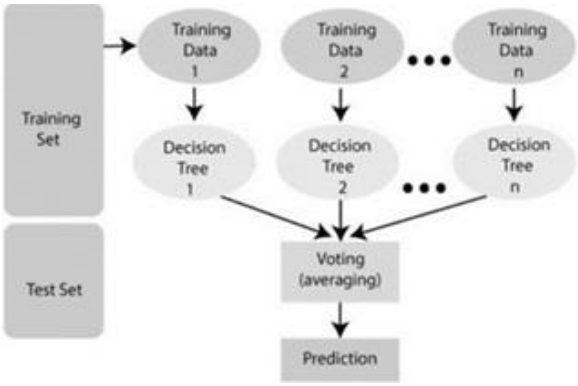

Fig 5. Decision Tree Classifier

K – Nearest Neighbor (KNN) Algorithm:
Machine learning techniques for classification and regression problems include K-nearest neighbor (K-NN).

Then, a prediction is created using the class or value of the KNN of a certain input datapoint given in the training dataset.

The K-NN technique is computationally expensive for large datasets because it must determine the distance between each input data point and each other data point in the training dataset. Because it is simple, non- parametric, and capable of handling classification tasks involving multiple classes and labels, the K-NN technique is well-liked. It might not perform as well on high-dimensional data, and it could require careful feature selection and preprocessing to perform better.

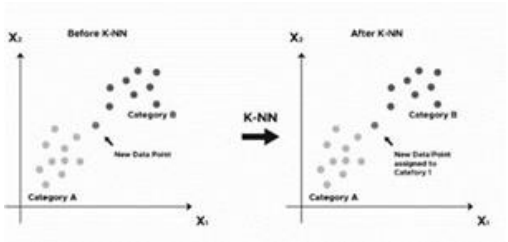

Fig. 5. K-Nearest Neighbor Algorithm

Classification and regression problems can be resolved using the supervised machine learning method KNN. KNN employs the formula 2 and observation of the K that are most like the test observation X0 to calculate the conditional probability that an observation belongs to class j:

$$Pr (Y= j | X = x0) = 1/k * \sum i E N0 I(yi = j) \quad (3)$$

Result

Therefore, the diabetes, heart disease, and Parkinson's disease prediction models in the system simultaneously employed the SVM algorithm, LR, RFC, DTC and KNN Algorithm. It will become clear if the patient has the condition in issue when the patient provides the illness- specific parameter. If any of the values are invalid, out of range, or empty, a warning message demanding a valid value will be shown. The parameters will display the range of the necessary values.

Decision trees or K-Nearest Neighbors may perform well if the data set is minimal. For bigger data sets, RF and SVM are the better options.

If there are complex relationships between the variables in the data set, RF and SVM may perform better.

DTC and KNN perform well for smaller data sets.

Accuracy for Each Disease

Diabetes

Table I. Accuracy for Diabetes

Algorithm	Training accuracy	Testing accuracy
SVM	78%	77%
RFC	100%	71%
DTC	100%	71%
KNN	100%	71%

Heart Disease:

Table Ii. Accuracy for Heart Disease

Algorithm	Training accuracy	Testing accuracy
LR	85%	81%
RFC	100%	75%
DTC	78%	62%
KNN	100%	78%

Parkinson's Disease:

Table III. Accuracy for Parkinson's Disease

Algorithm	Training accuracy	Testing accuracy
SVM	87%	87%
RFC	100%	82%
DTC	100%	74%
KNN	87%	74%

Therefore, we can conclude that Support Vector Machine (SVM) and Random Forest Classifier provide results that are more categorical and accurate than those produced by any other approach, and they do so while maintaining a good fit and avoiding overfitting.

Evaluation Parameters

Table IV. Evaluation Parameters for Each Disease

	Diabetes(%)	Heart Disease(%)	Parkinsonn disease (%)
Precision	75	84	88
Recall	51	81	96
F1 score	61	83	92
MSE	0.008136	0.249200	0.227624
RMSE	0.0902	0.4992	0.4771
Confusion matrix	[[91 9] [26 28]]	[[23 5] [6 27]]	[[4 4] [1 30]]

Diabetes Disease:

Fig. 6. Diabetes disease prediction page interface

Heart Disease:

Fig 8. Heart Disease prediction page interface

Parkinson's Disease:

Fig. 7. Parkinson's disease prediction page interface

IV. CONCLUSION

This paper demonstrates how Pickle, Streamlit, and machine learning may be used to develop a user-friendly web tool that can forecast various ailments. To improve patient outcomes and the management of disorders, the app can help with the early identification and prevention of diseases including diabetes, cardiac arrest, and Parkinson's disease. More study might increase the machine learning models' precision and broaden the types of ailments that the web application can forecast. All things considered, using our web program to manage one's health and prevent the beginning of serious illnesses may be helpful for both patients and healthcare providers.

REFERENCES

1. Diabetes Mellitus Disease Prediction Using Machine Learning Classifiers and Techniques Using the Concept of Data Augmentation and Sampling 2023, Lecture Notes in Networks and Systems
2. Aishwarya Mujumdar, V Vaidehi, Diabetes Prediction using Machine Learning Algorithms, Procedia Computer Science,Volume 165, 2019, ISSN 1877-0509
3. M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers," in IEEE Access, vol. 8, pp. 76516-76531,2020, doi:10.1109/ACCESS.2020.2989857.
4. Apurv Garg et al 2021 IOP Conf. Ser.:Mater. Sci. Eng. 1022 012046
5. C. Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," 2022 Advances in Science and Engineering Technology International Conferences(ASET), Dubai, United Arab Emirates, 2022, pp. 1-6 doi:10.1109/ASET53988.2022.9734880.
6. Rindhe, Baban & Ahire, Nikita & Patil, Rupali & Gagare, Shweta & Darade, Manisha. (2021). Heart Disease Prediction Using Machine Learning. International Journal of Advanced Research in Science, Communication and Technology. 267-276.10.48175/IJARSCT- 1131.
7. Rajdhan, Apurb & Agarwal, Avi & Sai, Milan & Ghuli, Poonam. (2020). Heart Disease Prediction using Machine Learning. International Journal of Engineering Research and. V9 10.17577/IJERTV9IS040614.
8. Akhtar, Nayab. (2021). Heart Disease Prediction.
9. Early Prediction of Parkinson Disease Using Machine Learning and Deep Learning Approaches EasyChair January 12, 2021
10. Harshvardhan Tiwari, Shiji K Shridhar, Preeti V Patil, K R Sinchana and G Aishwarya
11. Basil K Varghese, Geraldine Bessie Amali D, Uma Devi K S. Prediction of Parkinson'sDisease using Machine Learning Techniques on Speech dataset. Research J. Pharm. and Tech 2019; 12(2):644-648.
12. Gokul S., Sivachitra M. and Vijayachitra S., "Parkinson's disease prediction using machine learning approaches," 2013 Fifth International Conference on Advanced Computing (ICoAC), Chennai, India, 2013, pp. 246-252, doi: 10.1109/ICoAC.2013.6921958.

13. M. S. Roobini, Y. R. K. Reddy, U. S.G. Royal, A. K. Singh and K. Babu, "Parkinson's Disease Detection Using Machine Learning," 2022 International Conference on Communication, Computing and Internet of Things (IC3IoT), Chennai, India, 2022, pp. 1-6
14. S. Kamoji, D. Koshti, V. V. Dmello, A.A. Kudel and N. R. Vaz, "Prediction of Parkinson's Disease using Machine Learning and Deep Transfer Learning from different Feature Sets," 2021 6th International Conference on Communication and Electronics
15. Systems (ICCES), Coimbatre, India, 2021, pp. 1715-1720.
16. Singh, Shikha & Sarote, Priti & Shingade, Nikita & Yelale, Deepti & Ranjan, Nihar. (2022). Detection of Parkinson's Disease using Machine Learning Algorithm. International Journal of Computer Applications. 184. 24-29. 10.5120/ijca2022922016.
17. https://www.kaggle.com/datasets/saura bh00 007/diabetescsv
18. https://www.kaggle.com/datasets/zhaoy ingz hu/heartcsv
19. https://www.kaggle.com/datasets/vikas ukani/parkinsons-disease- data-set