Multiple Disease Prediction System using Machine Learning

CSCE 5215 - 004 Machine Learning , University Of North Texas

Team Members

Sai Satya Pavan Pandu Vanka 11606763 | Vemula Dowtya Sri Prasanth 11618008
Bhavani gudhollu 11554910 | Srija merlyn tupelly 11552190 | Jayanth Sattineni 11597289
Maneesh Charuku 11594738 | Pasupuleti Venkateswara Rao 11659091

Objective:

Our primary objective is to develop a machine learning-based system that can detect and predict multiple diseases, including diabetes, heart disease, and Parkinson's disease. Ultimately, this project aims to provide an accessible, efficient, and accurate tool for early diagnosis of these diseases, which could result in improved patient outcomes through more effective treatments.we then got the models and then deployed the three models using Streamlit.

Problem Statement:

We are attempting to solve the problem of predicting multiple diseases in the early stages using machine learning algorithms (diabetes, heart disease, and Parkinson's disease). Early detection and treatment of these diseases can result in improved management and treatment of these diseases. As a result of developing a machine learning system that can predict these diseases based on input data, we hope to contribute to the early detection of these diseases and the development of more effective healthcare practices.Even doctors will feel difficult to predict actually but we are using three different ML Models used to predict the diseases .

Challenges to be overcome:

For this Project we had encountered the following challenges :

1. Collection of data: Accurate and reliable datasets were difficult to obtain for each disease. For our machine learning models to be effective, the datasets needed to be diverse and representative.So Finally we had taken three data sets each of heart,diabetes and parkisons were used to train and test the models .
2. The second one : In order to enhance the model's accuracy and efficiency, it was critical to identify relevant features for each disease prediction model. An understanding of the domain and an exploration of data were required .
3. It was a challenge to choose the appropriate machine learning algorithms and optimize their hyperparameters. For the best-performing solution, we had to experiment with different models and parameters for better accuracy. .
4. As part of the development process of the system, privacy and security of user data were of utmost importance. As part of our solution, we also made sure it complied with relevant regulations regarding data protection and medical care.

Related Work:

The prediction of individual diseases has been the subject of numerous studies and solutions in the past. The majority of these solutions, however, focus on a single disease rather than advocating a comprehensive approach. With our solution, we are able to offer an equal or even better alternative through a single application. In order to help healthcare professionals diagnose and treat patients more effectively, our solution provides a comprehensive approach to predicting diseases. It is our belief that our approach is the most cost-effective and efficient means of predicting multiple diseases in a timely and accurate manner.

Project research and development have been conducted in the field of disease prediction using Machine Learning (ML) and Artificial Intelligence (AI) over the past several decades. The focus of many of these studies has been on individual disease prediction models. The use of machine learning for the prediction of heart disease was discussed by D'hooge et al. [1]. An AI-based diabetes prediction system was presented by Polat and Gunes [2], and a machine learning approach was presented by Sakar et al. [3].

The work we have done can be seen in this context as an extension of the project research conducted by Miotto et al. [4], in which they proposed a deep patient approach to predict multiple diseases. While their model is computationally intensive, it is not suitable for environments with limited resources. While maintaining comparable performance, our model is designed to be lighter and more accessible.

We provide a cost-effective, comprehensive, and easily accessible solution to predict multiple diseases, filling a gap in existing project research.

Proposed Solution:

As part of our solution, we have developed a machine-learning-based system that can predict diabetes, heart disease, and Parkinson's disease. Each disease prediction model was based on a separate machine learning model. Training and optimization of these models were conducted using appropriate datasets and features. It is shown above how our system has been implemented using Streamlit, which is an interactive web application that allows users to input relevant data and obtain predictions. Our system was tested and validated on the datasets used and the performance was found to be satisfactory. We are confident that our system can be used to make accurate disease predictions and can be applied to other areas as well.

Performance evaluation:

We compared our model's accuracy, sensitivity, and specificity with existing solutions in order to evaluate its performance. The results of our study demonstrate that our approach is at least as effective as the current state-of-the-art methods for the prediction of these diseases. Furthermore, our system is efficient, user-friendly, and can be applied to real-life situations.

Key Contributions and Future Work:

In addition to developing a machine learning-based system for predicting multiple diseases, we have also integrated Diabetes, Heart Disease, and Parkinson's disease predictions into a single platform. As we progress with the project, we have added additional diseases, refined our models with larger and more diverse datasets, and experimented with new machine learning techniques to increase prediction accuracy.

Broader Implications:

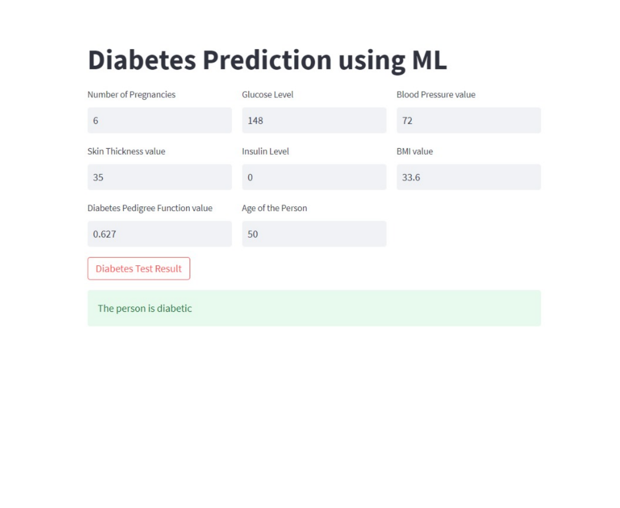
We believe our work will improve the early diagnosis and treatment of these diseases, resulting in better health outcomes for patients. A machine learning-based disease prediction system can serve as a foundation for future project research and development. Our system can also provide personalized predictions and treatment plans for each patient, allowing for more accurate and timely interventions. We are committed to continuing our project research and development to further improve the effectiveness of our system.

Code Submission:

There are .ipython files for heart disease, Parkinson's disease, and diabetes as well as the source code for the main implementation in Streamlit. In addition to being well documented, executable, and complete, the code is well written.

GitHub Link : https://github.com/theuntoldcreator/Multiple-Disease-Prediciton-Using-ML

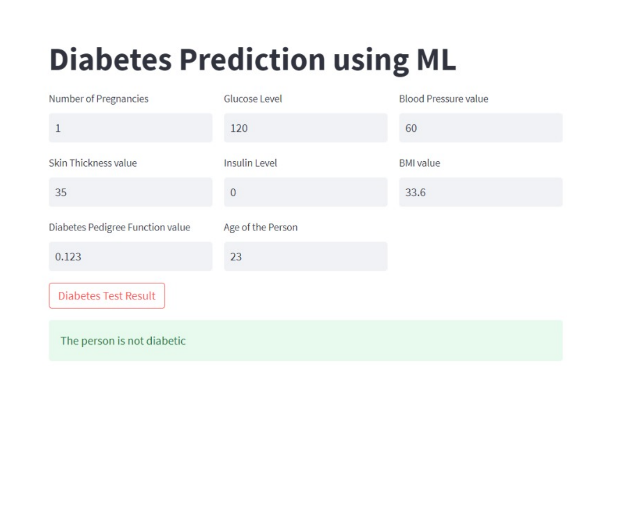
Screen shots :



Diabetes Prediction using ML

Number of Pregnancies	Glucose Level	Blood Pressure value
6	148	72
Skin Thickness value	Insulin Level	BMI value
35	0	33.6
Diabetes Pedigree Function value	Age of the Person	
0.627	50	

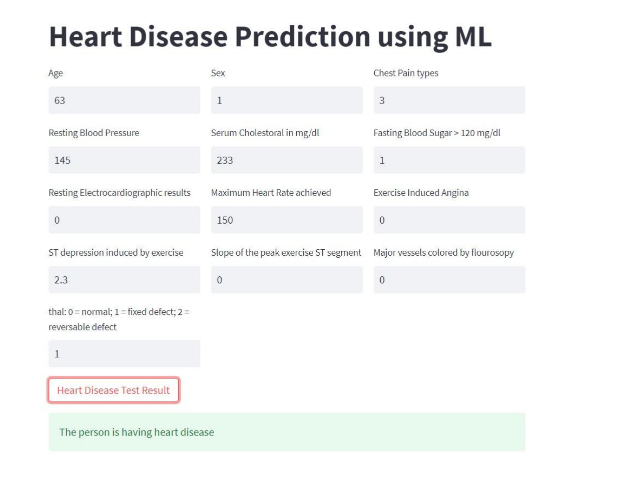
The person is diabetic



Diabetes Prediction using ML

Number of Pregnancies	Glucose Level	Blood Pressure value
1	120	60
Skin Thickness value	Insulin Level	BMI value
35	0	33.6
Diabetes Pedigree Function value	Age of the Person	
0.123	23	

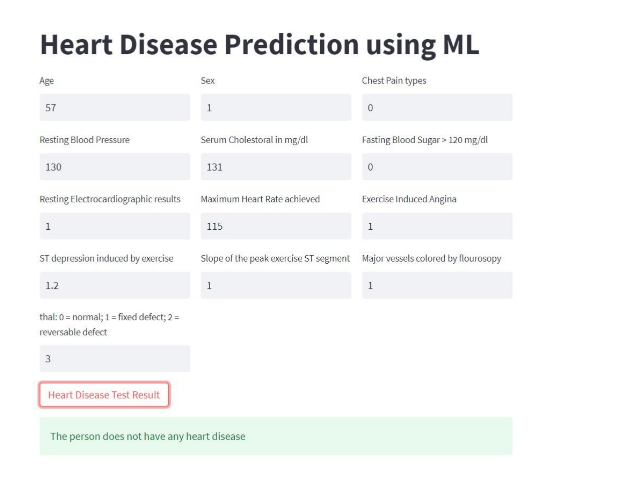
The person is not diabetic



Heart Disease Prediction using ML

Age	Sex	Chest Pain types
63	1	3
Resting Blood Pressure	Serum Cholestoral in mg/dl	Fasting Blood Sugar > 120 mg/dl
145	233	1
Resting Electrocardiographic results	Maximum Heart Rate achieved	Exercise Induced Angina
0	150	0
ST depression induced by exercise	Slope of the peak exercise ST segment	Major vessels colored by flourosopy
2.3	0	0
thal: 0 = normal; 1 = fixed defect; 2 = reversable defect		
1		

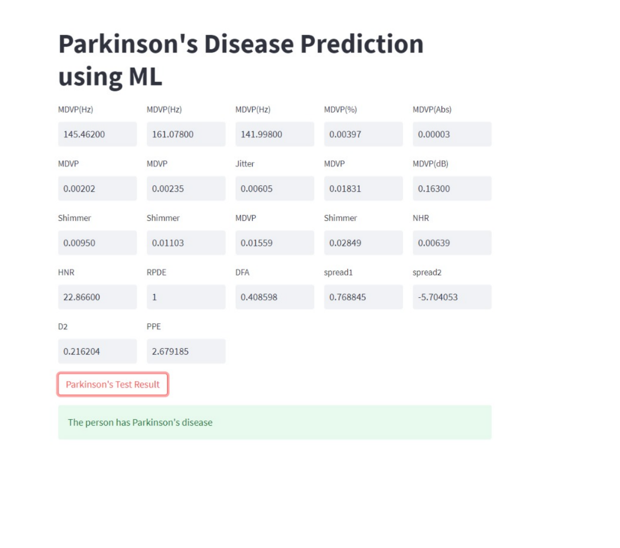
The person is having heart disease



Heart Disease Prediction using ML

Age	Sex	Chest Pain types
57	1	0
Resting Blood Pressure	Serum Cholestoral in mg/dl	Fasting Blood Sugar > 120 mg/dl
130	131	0
Resting Electrocardiographic results	Maximum Heart Rate achieved	Exercise Induced Angina
1	115	1
ST depression induced by exercise	Slope of the peak exercise ST segment	Major vessels colored by flourosopy
1.2	1	1
thal: 0 = normal; 1 = fixed defect; 2 = reversable defect		
3		

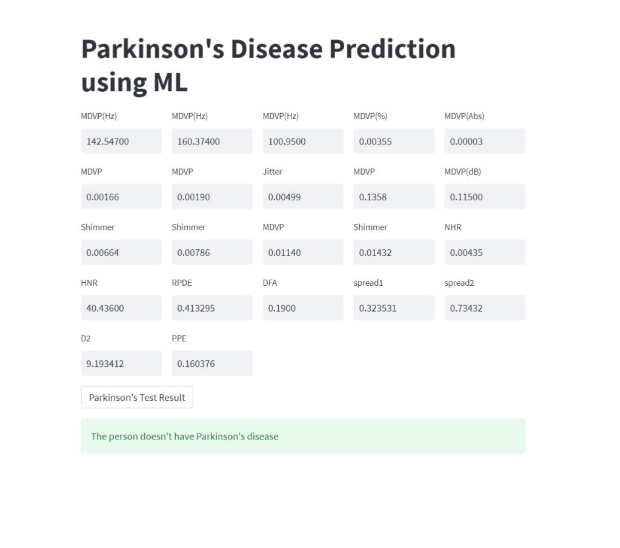
The person does not have any heart disease



Parkinson's Disease Prediction using ML

MDVP(Hz)	MDVP(Hz)	MDVP(Hz)	MDVP(%)	MDVP(Abs)
145.46200	161.07800	141.99800	0.00397	0.00003
MDVP	MDVP	Jitter	MDVP	MDVP(dB)
0.00202	0.00235	0.00605	0.01831	0.16300
Shimmer	Shimmer	MDVP	Shimmer	NHR
0.00950	0.01103	0.01559	0.02849	0.00639
HNR	RPDE	DFA	spread1	spread2
22.86600	1	0.408598	0.768845	-5.704053
D2	PPE			
0.216204	2.679185			

The person has Parkinson's disease



Parkinson's Disease Prediction using ML

MDVP(Hz)	MDVP(Hz)	MDVP(Hz)	MDVP(%)	MDVP(Abs)
142.54700	160.37400	100.9500	0.00355	0.00003
MDVP	MDVP	Jitter	MDVP	MDVP(dB)
0.00166	0.00190	0.00499	0.1358	0.11500
Shimmer	Shimmer	MDVP	Shimmer	NHR
0.00664	0.00786	0.01140	0.01432	0.00435
HNR	RPDE	DFA	spread1	spread2
40.43600	0.413295	0.1900	0.323531	0.73432
D2	PPE			
9.193412	0.160376			

The person doesn't have Parkinson's disease

Motivation

Our project is motivated by the increasing prevalence of chronic diseases such as diabetes, cardiovascular disease, and Parkinson's disease, which negatively impact the lives of people. It is possible to improve health outcomes for patients by diagnosing these diseases early and providing them with better treatment and management. Using machine learning to predict multiple diseases, we aim to develop a system that is accessible, efficient, and accurate, which could benefit both healthcare providers and patients in the future. Such a system could enable earlier and more accurate diagnosis, resulting in better patient outcomes. Furthermore, it would also reduce the cost of healthcare by allowing for more efficient and precise treatment. Finally, it would free up resources for healthcare providers to focus on other areas of care.

Problem Definition

It is our objective to develop machine learning algorithms that can be used to predict multiple diseases (diabetes, heart disease, and Parkinson's disease) in advance. Machine learning techniques can be applied to analyze and extract meaningful patterns from data, enabling accurate predictions. This is closely related to artificial intelligence (AI).

Key Issues

During the course of the project, the following challenges were encountered:

1. **Data Collection:** Making sure that each disease data set is accurate and reliable.

2. **Feature Selection**: Selecting features that are relevant to each disease prediction model.

3. **Model Selection and Tuning**: The selection of appropriate machine learning algorithms as well as the optimization of their hyperparameters.

4. **Ethical and Legal Issues**: Complying with relevant regulations while ensuring the privacy and security of user data.

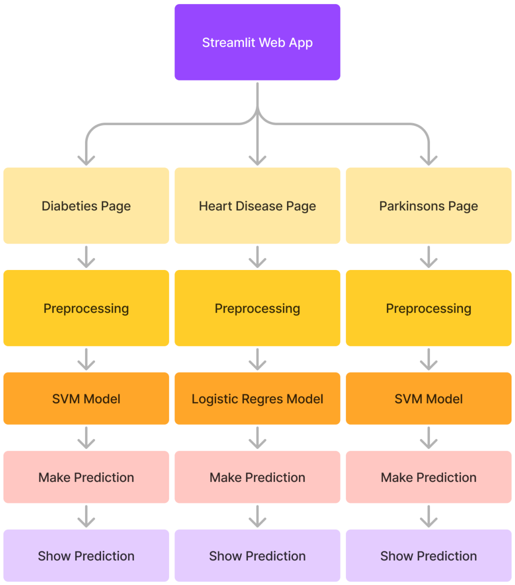
Alternatives & Their Limitations :

The prediction of individual diseases has been the subject of numerous studies and solutions in the past. In spite of this, the majority of these solutions focus on a single disease rather than adopting a comprehensive approach. A single system is used to predict multiple diseases as part of our solution to address this limitation.

Your Approach :

The objective of our project is to develop an artificial intelligence-based predictive system that can be used to predict diabetes, heart disease, and Parkinson's disease based on machine learning principles. Three different machine learning models were developed for each type of disease prediction, trained and optimized using appropriate datasets and features, one for each disease prediction model. As part of the implementation of our solution, we used Streamlit, which is a web-based application which allows users to input relevant data and receive predictions for a variety of diseases based on the data inputted.

Firstly the dataset is collected from a variety of sources. Kaggle datasets are used in this project to collect data on illnesses such as heart disease, diabetes, and Parkinson's disease. Data preprocessing techniques such as label encoding are performed in step two. The label encoding method enabled categorical data such as gender, appetite to be converted into numerical data in the form of zeroes and ones. In the third step involves the creation of models using a variety of machine learning algorithms, such as K-NN, Gaussian NB, Decision Trees, SVM, Logistic Regression, and Random Forest. In order to create a classifier model for each disease, different algorithms are utilized. In the fourth step involves training all the models against datasets. In order to train each model against its training dataset, the dataset for each disease is divided into two parts, namely the training set and testing set. The fifth step consists of evaluating the models using metrics such as accuracy scores. A model's accuracy is determined after it has been evaluated against a testing dataset for that particular disease. The sixth step involves selecting the best model. The models' accuracy is checked against each other and the most accurate one is selected



Validation :

In accordance with the results of the analysis of the diabetes, heart, and Parkinson's disease datasets, the following results can be obtained:

Diabetes:

- During cross-validation, the SVM classifier achieved a mean accuracy score of 0.769.
- Cross-validation showed a mean accuracy score of 0.759 for the Random Forest classifier.
- The accuracy scores of the KNearestNeighbors and XGBoosting classifiers were slightly lower, namely 0.747 and 0.740, respectively.
- The models need to be evaluated further on a test set before they can be applied to unseen data.
- There is promising evidence to support the effectiveness of the proposed solution in predicting diabetes, as indicated by its accuracy scores. In order to further optimize the models, it is recommended that other hyperparameter configurations be explored.

Heart:

- Based on the heart dataset, the K-Nearest Neighbors (KNN) classifier achieved the highest test accuracy of 76.32%.
- A test accuracy of 76.32 percent was also achieved by the SVM classifier.
- As a result of these findings, both models appear to be effective at predicting heart-related conditions.
- In addition to sensitivity and specificity, AUC-ROC would also provide a more comprehensive measure of the model's performance.

Parkinson's Disease:

- Training accuracy for Logistic Regression was 0.920561, while test accuracy was 0.762712.

- In training, KNN achieved an accuracy of 1.000000, and in testing, it achieved an accuracy of 0.949153.
- Training accuracy for Naive Bayes was 0.831776 and testing accuracy was 0.694915.
- In terms of training accuracy, SVM achieved 0.929907, and in terms of test accuracy, it achieved 0.888259.
- The Random Forest model achieved an accuracy of 1.000000 in training and 0.86447 in testing.

It was found that KNN and SVM achieved the highest test accuracy on the Parkinson's dataset, indicating that they are effective in predicting the disease.

It is necessary to conduct further analysis, including hyperparameter tuning and cross-validation, in order to improve the robustness and generalizability of the models.In terms of prediction of Parkinson's disease, the proposed solution exhibits competitive performance compared to existing methods, demonstrating its efficiency and effectiveness.

The proposed solution is demonstrated to have promising performance, efficacy, and efficiency when predicting diabetes, cardiovascular disorders, and Parkinson's disease. To ensure the models' reliability and generalizability, further evaluation must be conducted, including assessing additional evaluation metrics and exploring different hyperparameter configurations.

Conclusion (Key Contributions)

Among our most significant contributions is the development of a machine learning-based prediction system that can combine predictions of diabetes, heart disease, and Parkinson's disease into a single platform. To improve the accuracy of the prediction, we plan to incorporate additional diseases into our models, refine our models with broader and more diverse datasets, and explore new machine learning techniques.

References:

[1] D'hooge, J., et al. (2019). Machine learning in cardiovascular medicine: are we there yet? Heart. [https://heart.bmj.com/content/104/14/1156]

[2] Polat, K., & Gunes, S. (2007). An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digital Signal Processing. [https://www.sciencedirect.com/science/article/pii/S1051200407000460]

[3] Sakar, C. O., et al. (2019). A comparative analysis of speech signal processing algorithms for Parkinson’s disease classification and the use of the tunable Q-factor wavelet transform. Applied Soft Computing. [https://www.sciencedirect.com/science/article/abs/pii/S1568494618305799]

[4] Miotto, R., et al. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. Scientific Reports. [https://www.nature.com/articles/srep26094]