TABLE OF CONTENTS

ABSTRACT

Machine learning and Artificial Intelligence are playing a huge role in today's world. From self-driving cars to medical fields, we can find them everywhere. The medical industry generates a huge amount of patient data which can be processed in a lot of ways. So, with the help of machine learning, we have created a Prediction System that can detect more than one disease at a time. Many of the existing systems can predict only one disease at a time and that too with lower accuracy. Lower accuracy can seriously put a patient's health in danger. We have considered three diseases for now that are Heart Disease, Parkinson's and Diabetes and in the future, many more diseases can be added. The user has to enter various parameters of the disease and the system would display the output whether he/she has the disease or not. This project can help a lot of people as one can monitor the persons' condition and take the necessary precautions thus increasing the life expectancy.

Aim of the Project:

Our point is to anticipate the various sorts of illness in a single stage by utilizing the inbuilt python module Streamlit. In this task, we are utilizing Naïve Bayes, Random Forest, Decision Tree and SVM for prediction of a particular disease. The calculation which gives more accuracy is used to train the data set before implementation.

Objective of the Project:

To implement multiple disease analysis using machine learning algorithms, Streamlit and python pickling is utilized to save the model behavior. We analyze Diabetes, Heart disease and Parkinson's disease by using some of the basic parameters such as Pulse Rate, Cholesterol, Blood Pressure, Heart Rate, etc., and also the risk factors associated with it. The disease can be found using prediction model with good accuracy and Precision. The significance of this analysis is to analyze the maximum diseases to screen the patient's condition and caution the patients ahead of time to diminish mortality proportion.

1. INTRODUCTION

In this digital world, data is an asset, and enormous data was generated in all the fields. Data in the healthcare industry consists of all the information related to patients. Here, a general architecture has been proposed for predicting the disease in the healthcare industry. Many of the existing models are concentrating on one disease per analysis. Like, one analysis for diabetes analysis, one for cancer analysis, one for skin diseases like that. There is no common system present that can analyze more than one disease at a time. Thus, we are concentrating on providing immediate and accurate disease predictions to the users about the symptoms they enter along with the disease predicted. So, we are proposing a system which used to predict multiple diseases by using streamlit. In this system, we are going to analyze Diabetes, Heart, and Parkinson's disease analysis. Later, many more diseases can be included. Python pickling is used to save the behavior of the model. The importance of this system analysis is that while analyzing the diseases all the parameters which cause the disease is included so it is possible to detect the disease efficiently and more accurately.

1.1 Machine Learning

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers.

"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E." Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, and to uncover key insights in projects.

1.2 Deep Learning

Deep learning is a class of machine learning algorithms that utilizes a hierarchical level of artificial neural networks to carry out the process of machine learning. The artificial neural networks are built like the human brain, with neuron nodes connected together like a web. While traditional programs build analysis with data in a linear way, the hierarchical function of deep learning systems enables machines to process data with a nonlinear approach.

The word "deep" in "deep learning" refers to the number of layers through which the data is transformed. More precisely, deep learning systems have a substantial credit assignment path (CAP) depth. The CAP is the chain of transformations from input to output. CAPs describe potentially causal connections between input and output.

For a feedforward neural network, the depth of the CAPs is that of the network and is the number of hidden layers plus one (as the output layer is also parameterized). For recurrent neural networks, in which a signal may propagate through a layer more than once, the CAP depth is potentially unlimited.

1.3 Project Scope

In multiple disease prediction, it is possible to predict more than one disease at a time. So, the user doesn't need to traverse different sites in order to predict the diseases. We are taking three diseases that are Brain, Diabetes, and Heart. As all the three diseases are correlated to each other. To implement multiple disease analyses we are going to use machine learning algorithms and Streamlit. When the user is accessing this API, the user has to send the parameters of the disease along with the disease name. Streamlit will invoke the corresponding model and returns the status of the patient.

1.4 Project Features

The project is successfully deployed using streamlit and user can predict 3 types of diseases in the webapp, namely:

i. Diabetes Prediction
ii. Heart Disease Prediction
iii. Parkinson's Prediction

- Our Machine Learning Model works with 87%,85%, 89% accuracy in training data and 92%,86%, 85% accuracy with testing data in diabetes, heart and Brain disease respectively.
- We have used Logistic Regression algorithm, Naïve Bayes, Random Forest, KNN and SVM algorithm to train our model.
- We have also deployed the ML model in webapp using streamlit library in python.
- Our data set consist of 5000+ samples, with various parameters.
- The project can be more improved further with more variety of training data and using other ml algorithms.

1.4.1 Functional requirements:

- The system allows the patient to predict the disease.
- The user adds the input for the particular disease and based on the trained model of the user input the output will be displayed.

1.4.2 Non-functional requirements:

- The website will provide range of the values during the prediction of the disease.
- The website should be reliable and consistent.
- The system shall be able to display the prediction results in a clear and understandable format.

2. SYSTEM REQUIREMENTS

This project can run on commodity hardware. We ran entire project on an Intel I5 processor with 8 GB Ram, 2 GB Nvidia Graphic Processor, it also has 2 cores which runs at 1.7 GHz, **2.1** GHz respectively. First part of the is training phase which takes 10-15 mins of time and the second part is testing part which only takes few seconds to make predictions and calculate accuracy.

2.1 HARDWARE REQUIREMENTS:

1.RAM: 4 GB

2.Storage: 500 GB

3.CPU: 2 GHz or faster

4.Architecture: 32-bit or 64-bit

2.2 SOFTWARE REQUIREMENTS:

1. Python Language and Juypter Notebook for data preprocessing, model training and prediction.
2. VS Code for deploying the model using Streamlit.
3. Operating System : Windows 10 and above, Mac OS, Linux.

3. METHODOLOGY

We have experimented on three diseases that are heart diabetes and parkinsons as these are correlated to each other. The first step is to the dataset for heart disease, diabetes disease and Brain disease we have imported the UCI dataset, PIMA dataset and Indian Parkinson's dataset respectively. Once we have imported the dataset then visualization of each input data takes place. After visualization pre-processing of data takes place where we check for outliers, missing values and also scale the dataset then on the updated dataset we split the data into training and testing. Next is on the training dataset we had applied KNN, Naïve Bayes, Logistic Regression and random forest algorithm and applied knowledge on the classified algorithm using testing dataset. After applying knowledge, we chose the algorithm with the best accuracy for each of the disease. Then we built a pickle file for all the diseases and then integrated the pickle file with the streamlit framework for the output of the model on the webpage.

The Machine Learning Algorithms used in Project are as follows:

- **Support Vector Machine:** Support Vector Machine (SVM) is a machine learning algorithm used for classification and regression analysis. It works by finding the optimal hyperplane that separates the data into different classes. SVM is effective in handling high-dimensional datasets, and it can also handle nonlinear datasets by using a kernel function to transform the data into a higher-dimensional space.
- **Naïve Bayes:** Naive Bayes is a probabilistic machine learning algorithm used for classification tasks. It works by calculating the probability of each class based on the input features and then selecting the class with the highest probability. It assumes that the input features are independent of each other, which is why it is called "naive".
- **Logistic Regression:** Logistic Regression is a machine learning algorithm used for binary classification tasks. It works by modeling the relationship between the input features and the probability of a particular outcome. The output is a probability score between 0 and 1, which can be interpreted as the likelihood of the input belonging to a particular class.
- **Random Forests:** In this classifier, there are different random forests that give a value and a value with more votes is the actual result of this classifier. researchers have used different machine learning classifiers to detect diseases.

- **k-Nearest Neighbors Algorithm:** It is a type of supervised machine learning algorithm used for classification and regression. It is a non-parametric algorithm, meaning it doesn't make any assumptions about data distribution. The basic idea behind the KNN algorithm is to find the k data points in the training dataset that are closest to the point to be predicted, and then classify or predict the point based on the majority class or mean value of the k nearest neighbors.

3.1 WORK-FLOW CHART:

It is a visual layout of a process, project or job in the form of a flow chart. It's a highly effective way to impart the steps more easily in a process, how each one will be completed, by whom and in what sequence.

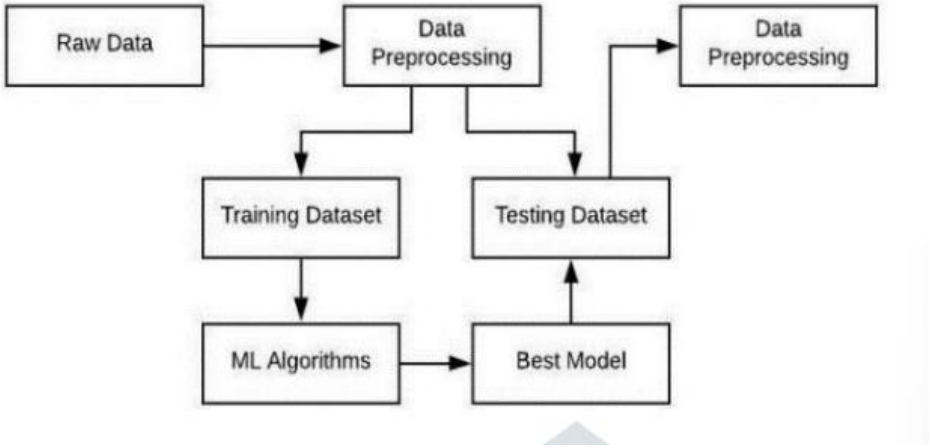

Figure 2 Flow Chart of the ML Model

3.2 Deployment of the model

Deployment of an ML-model simply means the integration of the model into an existing production environment which can take in an input and return an output that can be used in making practical business decisions.

Streamlit is an open-source Python library that makes it easy to create and share beautiful, custom web apps for machine learning and data science. In just a few minutes you can build and deploy powerful data apps. Using streamlit we have deployed the ML model in a responsive webpage.

3.3 DataSet

A machine learning dataset is a collection of data that is used to train the model. A dataset acts as an example to teach the machine learning algorithm how to make predictions.
Our data set contain following attributes.

Diabetes.csv
Attributes:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

Figure 1: Diabetes Dataset

Heart.csv
Attributes:

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Figure 2: Heart Dataset

Parkinsons.csv
Attributes:

	name	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP	MDVP:Shimmer
0	phon_R01_S01_1	119.992	157.302	74.997	0.00784	0.00007	0.00370	0.00554	0.01109	0.04374
1	phon_R01_S01_2	122.400	148.650	113.819	0.00968	0.00008	0.00465	0.00696	0.01394	0.06134
2	phon_R01_S01_3	116.682	131.111	111.555	0.01050	0.00009	0.00544	0.00781	0.01633	0.05233
3	phon_R01_S01_4	116.676	137.871	111.366	0.00997	0.00009	0.00502	0.00698	0.01505	0.05492
4	phon_R01_S01_5	116.014	141.781	110.655	0.01284	0.00011	0.00655	0.00908	0.01966	0.06425
...	...	...	...	...	...	...	...	...	...	...
190	phon_R01_S50_2	174.188	230.978	94.261	0.00459	0.00003	0.00263	0.00259	0.00790	0.04087
191	phon_R01_S50_3	209.516	253.017	89.488	0.00564	0.00003	0.00331	0.00292	0.00994	0.02751
192	phon_R01_S50_4	174.688	240.005	74.287	0.01360	0.00008	0.00624	0.00564	0.01873	0.02308
193	phon_R01_S50_5	198.764	396.961	74.904	0.00740	0.00004	0.00370	0.00390	0.01109	0.02296
194	phon_R01_S50_6	214.289	260.277	77.973	0.00567	0.00003	0.00295	0.00317	0.00885	0.01884

Figure 3.1: Parkinson's Dataset

Shimmer:DDA	NHR	HNR	status	RPDE	DFA	spread1	spread2	D2	PPE
0.06545	0.02211	21.033	1	0.414783	0.815285	-4.813031	0.266482	2.301442	0.284654
0.09403	0.01929	19.085	1	0.458359	0.819521	-4.075192	0.335590	2.486855	0.368674
0.08270	0.01309	20.651	1	0.429895	0.825288	-4.443179	0.311173	2.342259	0.332634
0.08771	0.01353	20.644	1	0.434969	0.819235	-4.117501	0.334147	2.405554	0.368975
0.10470	0.01767	19.649	1	0.417356	0.823484	-3.747787	0.234513	2.332180	0.410335
...	...	...	...	...	...	...	...	...	...
0.07008	0.02764	19.517	0	0.448439	0.657899	-6.538586	0.121952	2.657476	0.133050
0.04812	0.01810	19.147	0	0.431674	0.683244	-6.195325	0.129303	2.784312	0.168895
0.03804	0.10715	17.883	0	0.407567	0.655683	-6.787197	0.158453	2.679772	0.131728
0.03794	0.07223	19.020	0	0.451221	0.643956	-6.744577	0.207454	2.138608	0.123306
0.03078	0.04398	21.209	0	0.462803	0.664357	-5.724056	0.190667	2.555477	0.148569

Figure 3.2: Parkinson's Dataset (contd..)

4.OUTPUT:

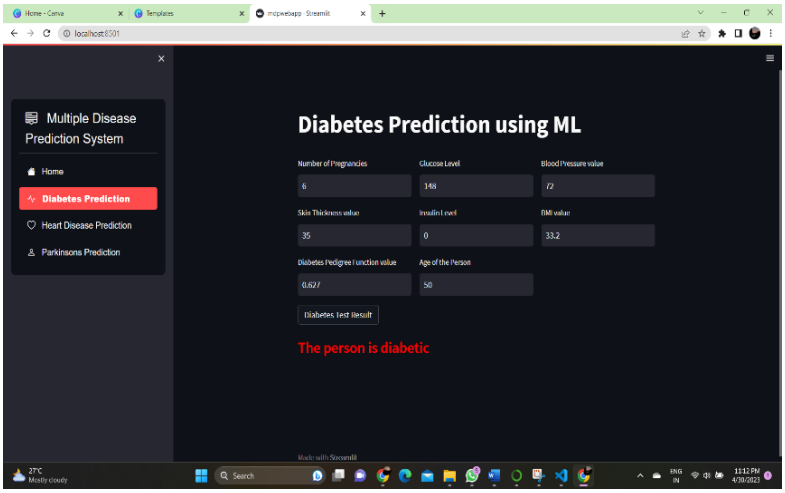

Figure 4: Diabetes Prediction Output

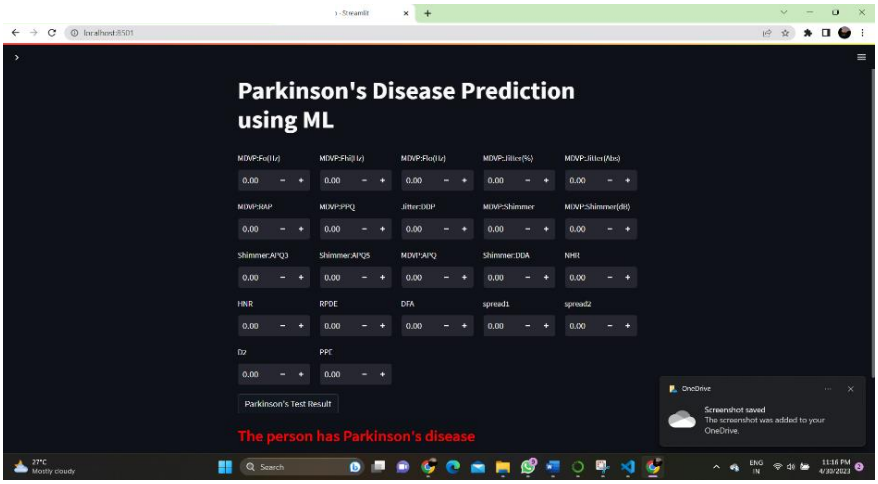

Figure 5: Heart Disease Prediction Output

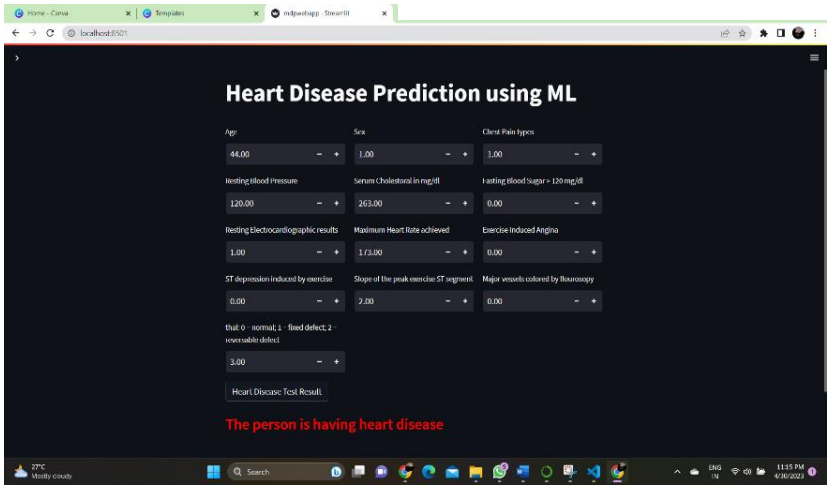

Figure 6: Parkinson's Disease Prediction Output

5. CONCLUSION & LIMITATIONS

The main objective of this project was to create a system that would predict more than one disease and with high accuracy. Because of this project the user need not to traverse different websites which saves time as well, and also gets to know the disease he's suffering from. Diseases if predicted early can increase the life expectancy as well as save you from financial troubles. For this purpose, we have used various machine learning algorithms like Random Forest, Naïve Bayes, Logistic Regression, SVM, and K nearest neighbor (KNN) to achieve maximum accuracy. In the future, we shall add more diseases in the existing API and try to improve the accuracy of prediction in order to decrease the mortality rate.

6. BIBLIOGRAPHY

1. https://www.geeksforgeeks.org/multiple-disease-prediction-using-machine-learning/
2. www.kaggle.com
3. https://www.pantechsolutions.net/multiple-disease-prediction -using-machine-learning
4. https://docs.streamlit.io/
5. https://www.youtube.com/watch?v=nacLBdyG6jE
6. CHATGPT