MULTIPLE DISEASE PREDICTION SYSTEM

A Project report submitted in partial fulfillment

Of the requirements for the awards of Degree of

Bachelor Of Technology

In

Computer Science and Engineering

By

J. SHYAMKUMAR(N180061)

M.SUBBARAO(N180505)

L.HARSHARAJESH(N180854)

Under the supervision of

Mrs. M.BABY ANUSHA

Department of Computer Science and Engineering

Rajiv Gandhi University of Knowledge Technologies -Nuzvid,Krishna district-521202

May 2023

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE AND TECHNOLOGIES
(A.P Government Act 18 of 2008)
RGUKT-NUZVID,KRISHNA Dist-521202
Tele-Fax:08656-235557/235150

CERTIFICATE OF COMPLETION

This is to certify that the work entitled, **"MULTIPLE DISEASE PREDICITION SYSTEM"** is the bonafied work of**J.Shyam Kumar** with ID No: **N180061** and **M.Subbarao** with ID No: **N180505** and **L.Harsha Rajesh** with IDNo: **N180854** carried out under my guidance and supervision for the partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in the department of Computer Science and Engineering in RGUKT Nuzvid. This work is done during the academicsession February 2023 – June 2023, under my guidance.

Mrs.M.Baby Anusha

Project Supervisor
Faculty, Dept. of CSE
RGUKT , Nuzvid

Mr. Chiranjeevi Sadu

Head of CSE Department
Assistant Professor, Dept. of CSE
RGUKT , Nuzvid

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE AND TECHNOLOGIES
(A.P Government Act 18 of 2008)
RGUKT-NUZVID,KRISHNA Dist-521202
Tele-Fax:08656-235557/235150

CERTIFICATE OF EXAMINATION

This is to certify that the work entitled, **"MULTIPLE DISEASE PREDICITION SYSTEM"** is the bonafied work of **J.Shyam Kumar** with ID No: **N180061** and **M.Subbarao** with ID No: **N180505** and **L.Harsha Rajesh** with IDNo: **N180854** here by accord our approval of it as a study carried out and presented in a mannerrequired for its acceptance in the partial fulfillment of the requirement for the award of the degree of Bachelor of Technology for which it has been submitted. This approval does not necessarily endorse or accept every statement made, opinion expressed or conclusion drawn, as a recorded in this thesis. It only signifies the acceptance of this thesis for the purpose for which it has been submitted.

Mrs.M.Baby Anusha

Project Supervisor

Faculty Dept. of CSE

RGUKT , Nuzvid

Examiner

Project Examiner

Faculty Dept. of CSE

RGUKT , Nuzvid

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE AND TECHNOLOGIES
(A.P Government Act 18 of 2008)
RGUKT-NUZVID,KRISHNA Dist-521202
Tele-Fax:08656-235557/235150

DECLARATION

We, J.Shyam Kumar with ID No: N180061 and M.Subbarao with ID No: N180505 and L.Harsha Rajesh with ID No: N180854 hereby declare that the project report entitle "**MULTIPLE DISEASE PREDICTION SYSTEM**" done by us under the guidance of Mrs.M.Baby Anusha is submitted for the partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering in the academic session February 2023 – June 2023 at RGUKT – Nuzvid. We also declare that this project is a result of our own effort and has not been copied or imitated from any source. Citations from any websites are mentioned in the references. The results embodied in this project report have not been submitted to any other university or institute for the awardof any degree or diploma.

Date : 14-08-2023

Place:Nuzvid

J. ShyamKumar(N180061)

M.Subbarao(N180505)

L.Harsha Rajesh(N180854)

ACKNOWLEDGEMENT

We would like to expose our profound gratitude and deep regards to my guide
Mrs.M.Baby Anusha for his exemplary guidance, monitoring and constant encouragement throughout the course of this thesis.
We are extremely grateful for the confidence bestowed in us and entrusting our project entitled "**MULTIPLE DISEASE PREDICTION**".
At this juncture we feel deeply honoured in expressing our sincere thanks to him for making the resources available at right time and providing valuable insights leading to the successful completion of our project.
We would like to thank RGUKT Nuzvid Director, HOD, faculty and staff for their valuable suggestions and discussions.
Last but not least I thank almighty, and I place a deep sense of gratitude to my family members and my friends who have been constant source of information during the preparation of this project work.

J.Shyam Kumar(N180061)
M.Subbarao(N180505)
L.HarshaRajesh(N180854)

Table Of Contents — Pages

ABSTRACT

This project aims to develop a comprehensive disease prediction system that focuses on predicting three prevalent medical conditions: Diabetes, Heart disease, and Parkinson's disease. By using various machine learning algorithms, the system will analyze various patient health parameters and medical histories to provide accurate predictions of potential disease risks. The system will be integrated with a user-friendly interface using Streamlit, a popular Python library for creating interactive web applications.

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. Our project is combination of two machine learning models to predict disease at early stage, we split dataset intro training and testing set and we calculate accuracy for the each algorithm. we dumped the model using Python's Pickle library and created web application using Python's Streamlit for easy user interaction.

REQUIREMENTS

1. Software Requirements

1) Python idle(2.7 or more)
2) anaconda for jupyter notebook(version 1.7.2 or above)

2. Hardware Requirements

1) Processer: I3 or more
2) Os : Windows
3) Ram : 4GB or more
4) Hard drive: 50gb

3. Technologies

As our project is a machine learning project we need many python libraries for data collection, preprocessing and modelling the data and for applying machine learning algorithms

The libraries that are needed for our project:

I. Spyder Python 3.8
II. Packages and Libraries
III. Numpy
IV. Pandas
V. Sklearn
VI. Streamlit and Streamlit option-menu
VII. Pickle

DATA COLLECTION

For the development of a comprehensive disease prediction system that includes diabetes, heart disease, and Parkinson's disease, data plays a crucial role. Kaggle, a popular platform for data science competitions, offers a wealth of datasets that can be utilized for this purpose. The first step in the data collection process is to search for relevant datasets related to each disease on Kaggle. Kaggle's search functionality allows us to explore various datasets by keywords, such as “Diabetes," “Heart disease," and "Parkinson's disease."

Few rows of diabetes dataset is shown in the figure

Pregnanci	Glucose	BloodPres	SkinThickr	Insulin	BMI	DiabetesP	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1

Fig1: Few rows of diabetes dataset

Few rows of heart dataset is shown in the figure

A	B	C	D	E	F	G	H	I	J	K	L	M	N
age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

Fig2: Few rows of heart dataset

Few rows of parkinsons dataset is shown in the figure

MDVP:Fo	MDVP:Fh	MDVP:Flc	MDVP:Jitt	MDVP:Jitt	MDVP:RA	MDVP:PPI	Jitter:DDI	MDVP:Sh	MDVP:Sh	Shimmer	Shimmer	MDVP:AP	Shimmer	NHR	HNR	status	RPDE	DFA	spread1	spread2	D2	PPE
119.992	157.302	74.997	0.00784	0.00007	0.0037	0.00554	0.01109	0.04374	0.426	0.02182	0.0313	0.02971	0.06545	0.02211	21.033	1	0.41478	0.81529	-4.81303	0.26648	2.30144	0.28465
122.4	148.65	113.819	0.00968	0.00008	0.00465	0.00696	0.01394	0.06134	0.626	0.03134	0.04518	0.04368	0.09403	0.01929	19.085	1	0.45836	0.81952	-4.07519	0.33559	2.48686	0.36867
116.682	131.111	111.555	0.0105	0.00009	0.00544	0.00781	0.01633	0.05233	0.482	0.02757	0.03858	0.0359	0.0827	0.01309	20.651	1	0.4299	0.82529	-4.44318	0.31117	2.34226	0.33263
116.676	137.871	111.366	0.00997	0.00009	0.00502	0.00698	0.01505	0.05492	0.517	0.02924	0.04005	0.03772	0.08771	0.01353	20.644	1	0.43497	0.81924	-4.1175	0.33415	2.40555	0.36898
116.014	141.781	110.655	0.01284	0.00011	0.00655	0.00908	0.01966	0.06425	0.584	0.0349	0.04825	0.04465	0.1047	0.01767	19.649	1	0.41736	0.82348	-3.74779	0.23451	2.33218	0.41034
120.552	131.162	113.787	0.00968	0.00008	0.00463	0.0075	0.01388	0.04701	0.456	0.02328	0.03526	0.03243	0.06985	0.01222	21.378	1	0.41556	0.82507	-4.24287	0.29911	2.18756	0.35778
120.267	137.244	114.82	0.00333	0.00003	0.00155	0.00202	0.00466	0.01608	0.14	0.00779	0.00937	0.01351	0.02337	0.00607	24.886	1	0.59604	0.76411	-5.63432	0.25768	1.85479	0.21176
107.332	113.84	104.315	0.0029	0.00003	0.00144	0.00182	0.00431	0.01567	0.134	0.00829	0.00946	0.01256	0.02487	0.00344	26.892	1	0.63742	0.76326	-6.1676	0.18372	2.06469	0.16376
95.73	132.068	91.754	0.00551	0.00006	0.00293	0.00332	0.0088	0.02093	0.191	0.01073	0.01277	0.01717	0.03218	0.0107	21.812	1	0.61555	0.77359	-5.49868	0.32777	2.32251	0.23157
95.056	120.103	91.226	0.00532	0.00006	0.00268	0.00332	0.00803	0.02838	0.255	0.01441	0.01725	0.02444	0.04324	0.01022	21.862	1	0.54704	0.79846	-5.01188	0.326	2.43279	0.27136
88.333	112.24	84.072	0.00505	0.00006	0.00254	0.0033	0.00763	0.02143	0.197	0.01079	0.01342	0.01892	0.03237	0.01166	21.118	1	0.61114	0.77616	-5.24977	0.391	2.40731	0.24974

Fig3: Few rows of parkinsons dataset

DIABETES PREDICTION SYSTEM

- **Machine Learning Algorithm.**

  - Support Vector Machine
    Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well it's best suited for classification. The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space

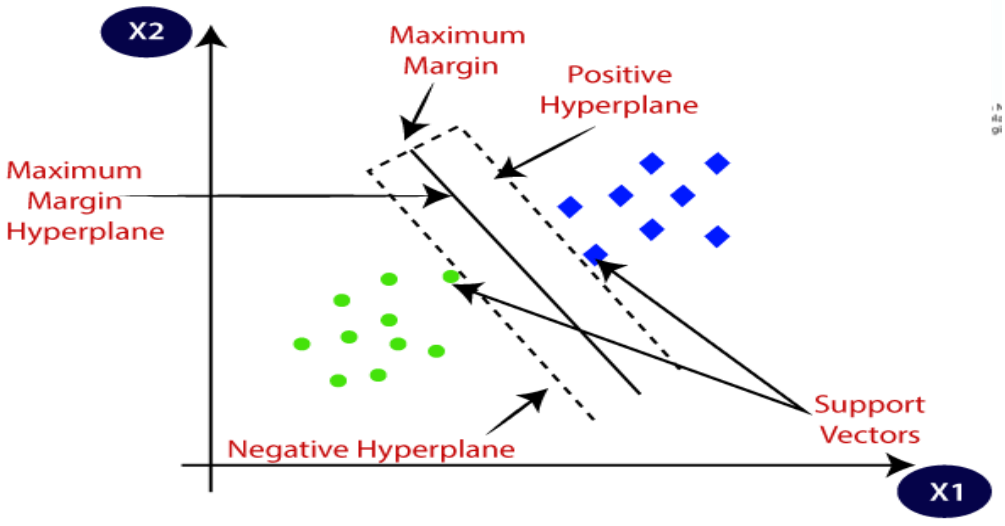

Fig: Support Vector Machine Procedure

- **Keywords In Algorithm**

  - **Hyperplane:** Hyperplane is the decision boundary that is used to separate the data points of different classes in a feature space. In the case of linear classifications, it will be a linear equation i.e. wx+b = 0.

  - **Support Vectors:** Support vectors are the closest data points to the hyperplane, which makes a critical role in deciding the hyperplane and margin.

  - **Margin**: Margin is the distance between the support vector and hyperplane. The main objective of the support vector machine algorithm is to maximize the margin. The wider margin indicates better classification performance.

  - **Kernel:** Kernel is the mathematical function, which is used in SVM to map the original input data points into high-dimensional feature spaces, so, that the hyperplane can be easily found out even if the data points are not linearly separable in the original input space. Some of the common kernel functions are linear, polynomial, radial basis function(RBF), and sigmoid.

- **Work Flow For Prediction**

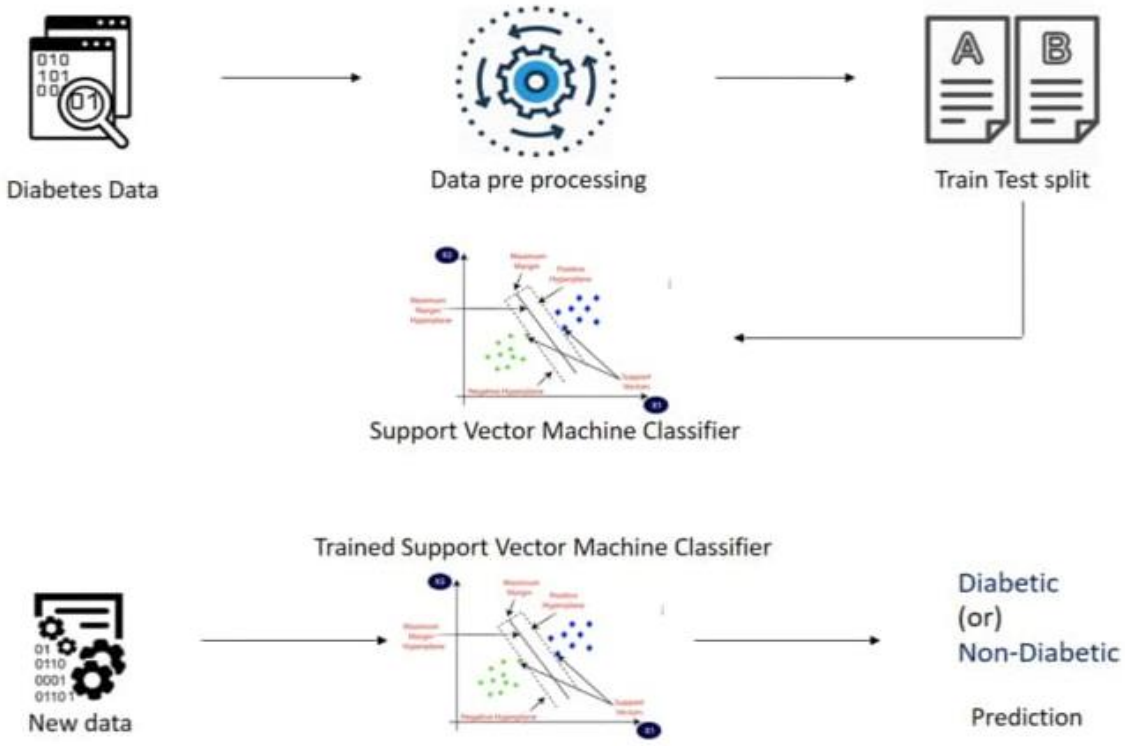

Fig: Work Flow For Diabetes Prediction System

- **Accuracy Of System**
  - Accuracy score of the training data : 0.7833876221498371
  - Accuracy score of the test data : 0.7727272727272727
- **Behaviour Of The Model**
  - Using Dump function we had saved our model behaviour in "diabetes_model.sav" file.

HEART DISEASE PREDICTION SYSTEM

- **Machine Learning Algorithm**
  - Logistic Regression

    Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class.

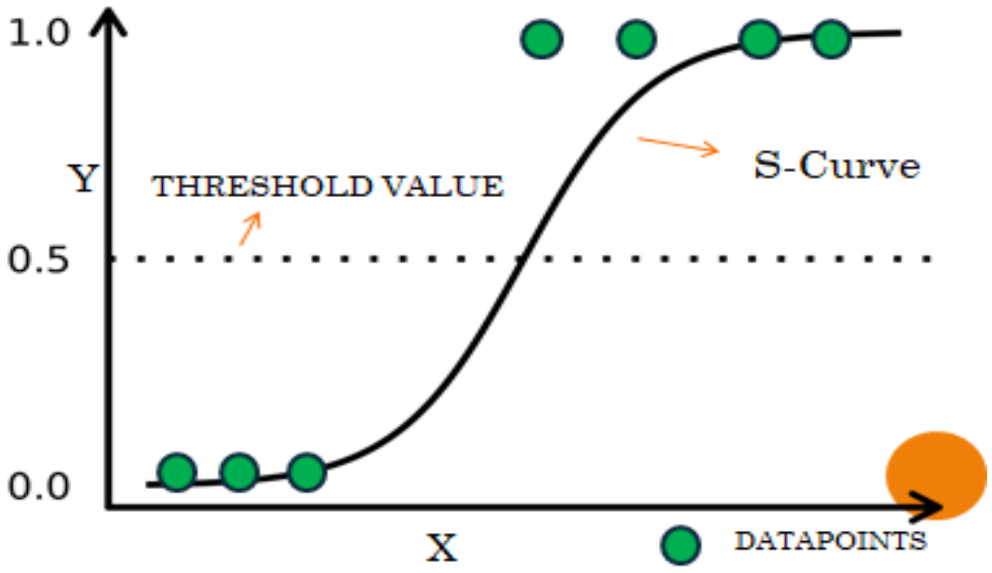

Fig: Logistic Regression Procedure

- **Sigmoid Function**

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

- **Work Flow For Prediction**

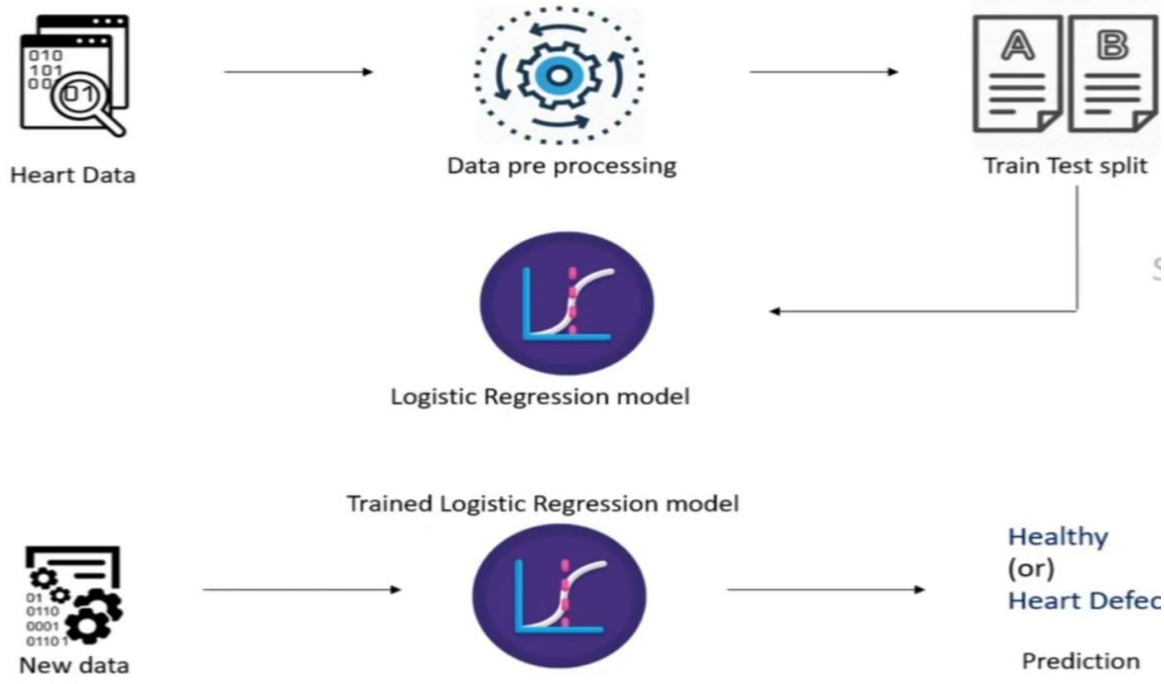

Fig: Work Flow for Heart Disease Prediction System

- **Accuracy Of System**
  - Accuracy score of the training data : 0.8512396694214877
  - Accuracy score of the test data : 0.8512396694214877

- **Behaviour Of The Model**
  - Using Dump function we had saved our model behaviour in "heart_model.sav" file.

PARKINSONS PREDICTION SYSTEM

- **Machine Learning Algorithm.**

  - Support Vector Machine
    Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well it's best suited for classification. The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space

- **Work Flow For Prediction**

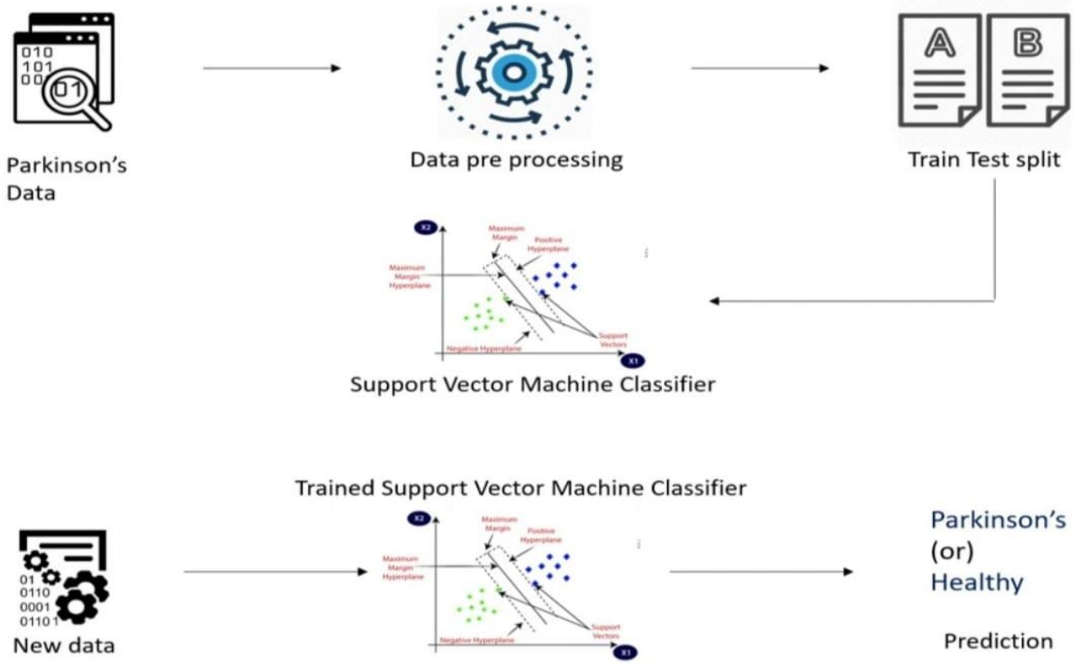

Fig: Work Flow for Parkinsons Disease Prediction System

- **Accuracy Of System**
  - Accuracy score of the training data : 0.8717948717948718
  - Accuracy score of the test data : 0.8717948717948718

- **Behaviour Of The Model**
  - Using Dump function we had saved our model behaviour in "heart_model.sav" file.

INTERFACE FOR THE SYSTEM

- Streamlit Package:
  - Streamlit is a powerful and user-friendly Python library that facilitates the creation of interactive web applications for data science and machine learning projects. It was developed to simplify the process of converting data scripts and models into interactive web applications without the need for complex web development skills.
  - Streamlit has gained significant popularity in the data science and machine learning communities due to its simplicity, versatility, and focus on data exploration and visualization. Whether you want to showcase your latest machine learning model, create data dashboards, or explore datasets

interactively, Streamlit provides a powerful and accessible tool to accomplish these tasks with ease.

User Interface Of Our System

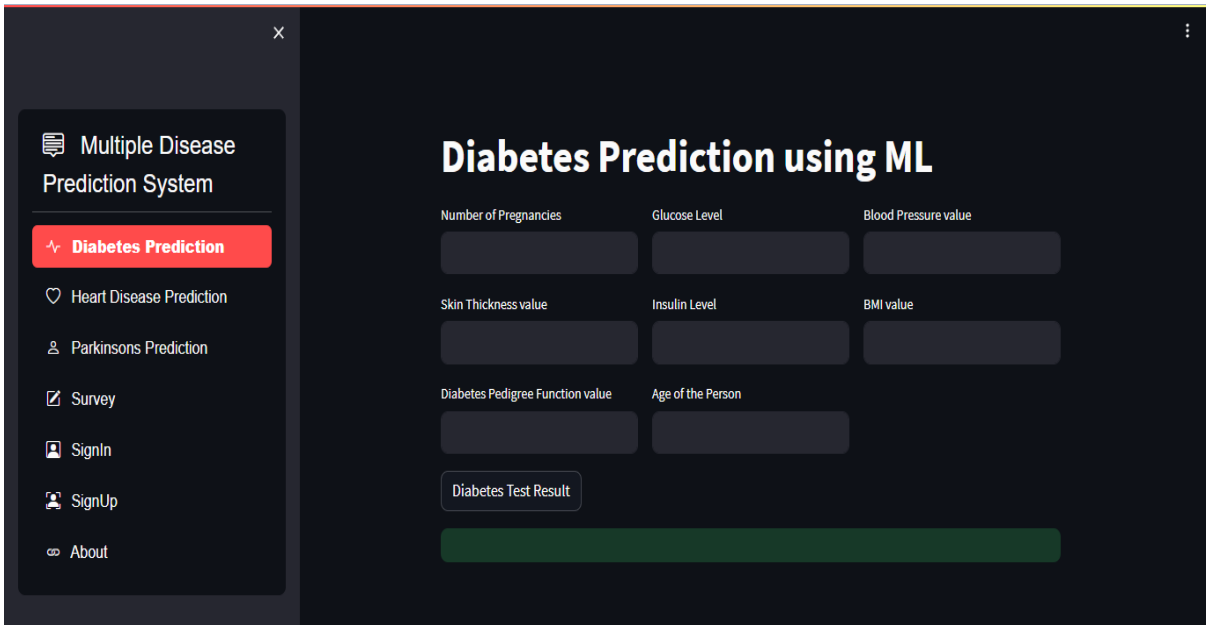

Fig: Web Interface image of our system using Streamlit

OUTPUTS

❖ Diabetes Prediction System

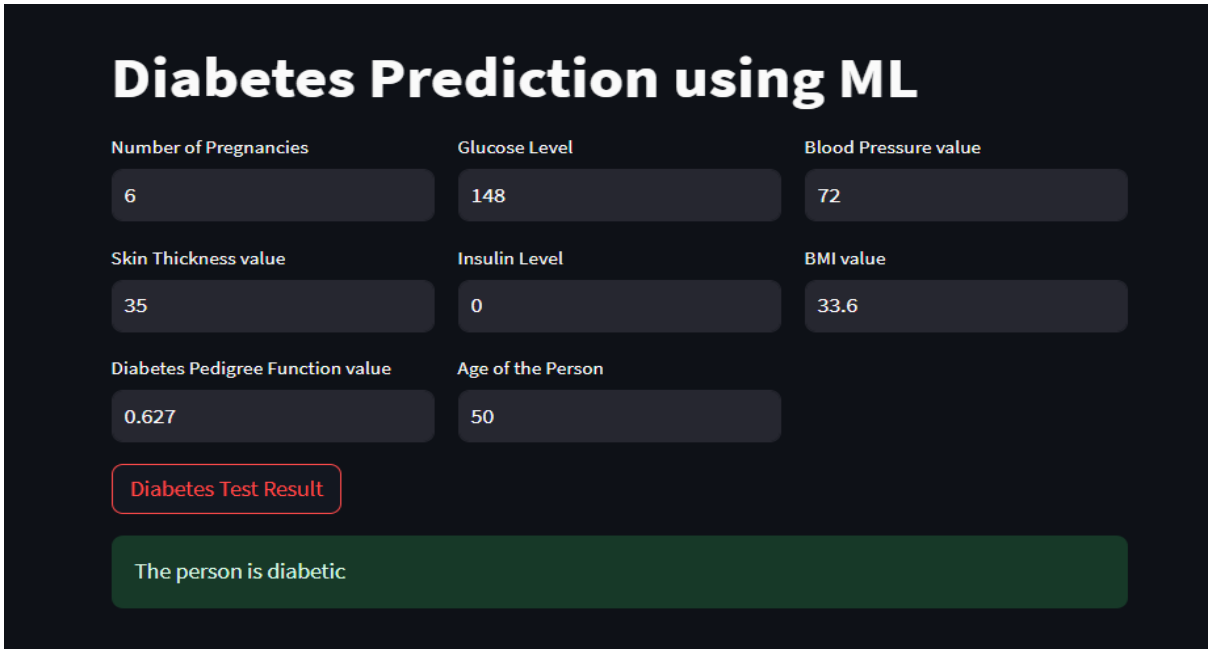

Fig: Output For Diabetes Prediction System

❖ Heart Disease Prediction System

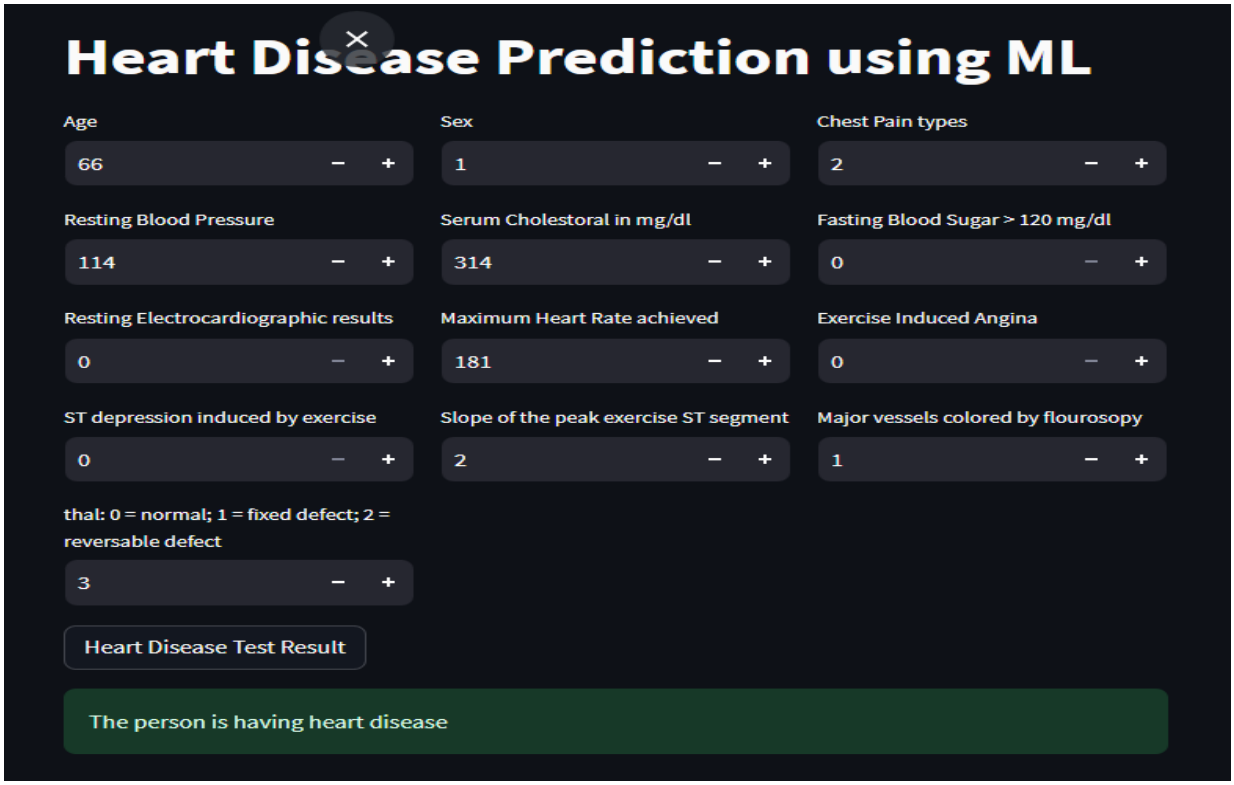

Fig: Output For Heart Disease Prediction System

❖ Parkinson's Disease Prediction System

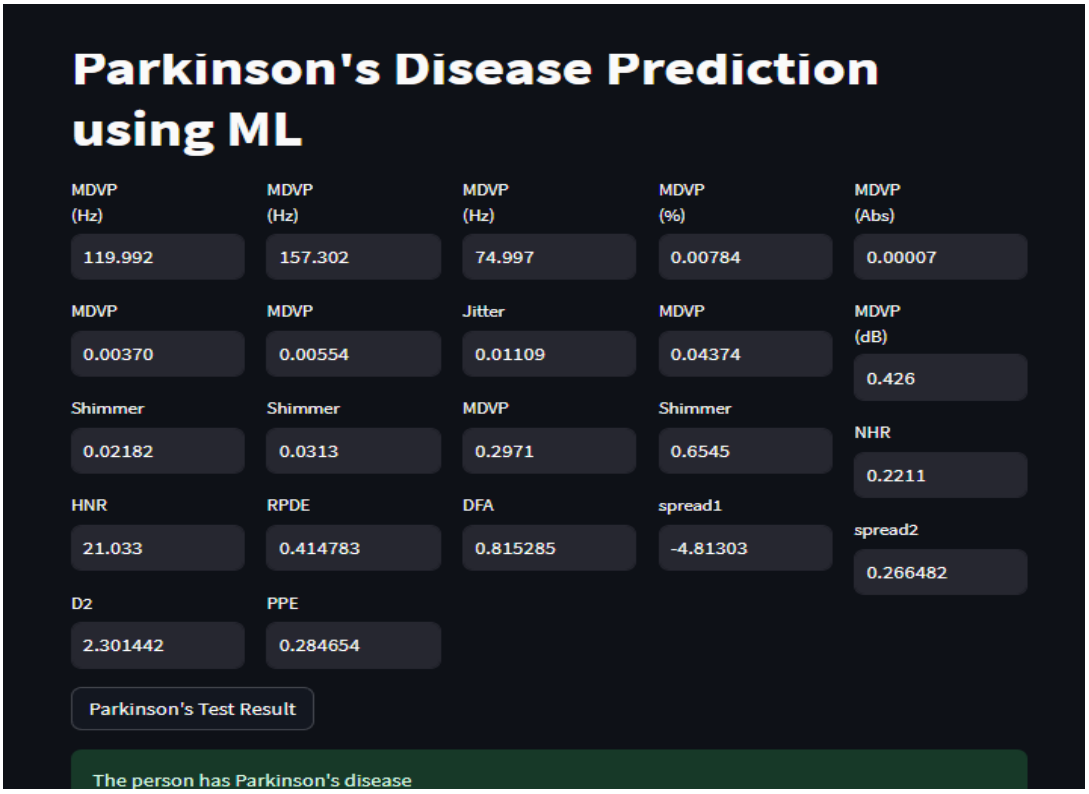

Fig: Output For Parkinsons Prediction System

CONCLUSION

- Through extensive data collection and preprocessing, we created a comprehensive dataset encompassing relevant health parameters for diabetes, heart disease, and Parkinson's disease. The chosen machine learning algorithms, SVM and Logistic Regression, have proven to be effective in capturing patterns within the data and making accurate disease predictions.
- By leveraging the power of the pickle Python library, we successfully saved the trained machine learning models, ensuring their reusability and easy integration into the Streamlit web application.
- In conclusion, the combination of machine learning algorithms, pickle for model saving, and Streamlit for the interface has resulted in accessible disease prediction system

REFERENCES

- https://www.kaggle.com/datasets/mathchi/diabetes-data-set
- https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset
- https://www.kaggle.com/datasets/vikasukani/parkinsons-disease-data-set

❖ IEEE Links For Algorithm

- https://ieeexplore.ieee.org/document/8962457
- https://ieeexplore.ieee.org/document/708428