A

Project Report

On

MULTIPLE DISEASE PREDICTION USING MACHINE LEARNING

Submitted to

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE AND TECHNOLOGIES, RK VALLEY, KADAPA

in partial fulfillment of the requirements for the award of the Degree of

BACHELOR OF TECHNOLOGY

IN

ELECTRONICS AND COMMUNICATION ENGINEERING

Submitted by

C.Jaishnavi	R170367
T.Dharani	R170413
K.Sreelekha	R171031

Under the Guidance of

B.MADAN MOHAN, Assistant Professor

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES

(Catering the Educational Needs of Gifted Rural Youth of AP)

R.K. Valley, Vempalli (M), Kadapa(dist)– 516330

2019-2023

RAJIV GANDHI UNIVERSITY OF KNOWLEDGE TECHNOLOGIES

(Catering the Educational Needs of Gifted Rural Youth of AP)
R.K. Valley, Vempalli(M), Kadapa(dist)– 516330

2019-2023

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

CERTIFICATE

This is to certify that the project report entitled **"Multiple Disease Prediction Using Machine Learning"** a bonafide record of the project work done and submitted by

C. JAISHNAVI	**(R170367)**
T. DHARANI	**(R170413)**
K. SREELEKHA	**(R171031)**

for the partial fulfillment of the requirements for the award of B.Tech. Degree in **ELECTRONICS AND COMMUNICATION ENGINEERING,RGUKT, RK Valley, Kadapa.**

GUIDE	**Head of the Department**
B.MADAN MOHAN	B.MADAN MOHAN

INTERNAL EXAMINER **EXTERNAL EXAMINER**

DECLARATION

We here by declare that the project report entitled **“ MULTIPLE DISEASE PREDICTION USING MACHINE LEARNING”** submitted to the Department of **ELECTRONICS AND COMMUNICATION ENGINEERING** in partial fulfillment of requirements for the award of the degree of **BACHELOR OF TECHNOLOGY**. This project is the result of our own effort and that it has not been submitted to any other University or Institution for the award of any degree or diploma other than specified above.

C. Jaishnavi	**(R170367)**
T. Dharani	**(R170413)**
K. Sreelekha	**(R171031)**

ACKNOWLEDGEMENT

We are thankful to our guide and Head of the Department of ELECTRONICS AND COMMUNICATION ENGINEERING, **Mr. B. MADAN MOHAN** for his valuable guidance and encouragement. His helping attitude and suggestions have helped us in the successful completion of the project.

We have great pleasure in expressing our hearty thanks to our beloved Director **Mrs. Sandhya Rani** for spending her valuable time with us to complete this project.

Successful completion of any project cannot be done without proper support and encouragement. We sincerely thanks to the **Management** for providing all the necessary facilities during the course of study.

We would like to thank our parents and friends, who have the greatest contributions in all our achievements, for the great care and blessings in making us successful in all our endeavors.

C. Jaishnavi **(R170367)**

T. Dharani **(R170413)**

K. Sreelekha **(R171031)**

ABSTRACT

Many of the existing machine learning models for health care analysis are concentrating on one disease per analysis. Like one analysis if for diabetes analysis, one for brain diseases like that. There is no common system where one analysis can perform more than one disease prediction. In this project proposing a system which used to predict multiple diseases by using Streamlit in Python. In this article used to analyse Diabetes analysis, Parkinson disease analysis. To implement multiple disease analysis used machine learning algorithms, Streamlit and web designing. Python pickling is used to save the model behaviour and python unpickling is used to load the pickle file whenever required. The importance of this project analysis in while analysing the diseases all the parameters which causes the disease is included so it possible to detect the maximum effects which the disease will cause. For example for diabetes analysis in many existing systems considered few parameters like age, sex, BMI , insulin, glucose, blood pressure, diabetes pedigree function, pregnancies, considered in addition to age, sex, BMI, insulin, glucose, blood pressure, diabetes pedigree function, pregnancies included serum creatinine, potassium, Glasgow Com Scale, heart rate/pulse Rate, respiration rate, body temperature, low density lipoprotein (LDL), high density lipoprotein (HDL), TG (Triglycerides).Final models behaviour will be saved as python pickle file. After user accessing this website, the user has to send the parameters of the disease along with disease name. This website will invoke the corresponding model and returns the status of the patient. The importance of this analysis to analyse the two diseases, so that to monitor the patient's condition and warn the patients in advance to decrease mortality ratio.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

SVM	Support Vector Machine
CSV	Comma Separated Values

CHAPTER -1

INTRODUCTION

During a lot of analysis over existing systems in healthcare analysis considered only one disease at a time.When any organization wants to analyse their patient's health reports then they have to deploy many models. The approach in the existing system is useful to analyse only particular disease. Now a day's mortality got increased due to exactly not identifying exact disease. Even the patient got cured from one disease may be suffering from another disease.

Some existing systems used few parameters while analysing the disease. Due to that may be not possible to identify the diseases which will be caused due to the effect of that disease. For example, due to diabetes, there may be chance of neuropath, retinopathy, hearing loss, and dementia. In this article considered Diabetes analysis and Parkinson disease data sets. In future many other diseases like skin diseases can be included, fever related diseases and many more. This analysis is flexible that later included many diseases for analysis. While adding any new disease analysis to this existing Website, the developer has to add the model file related to the analysis of the new disease. When developing new disease the developer have to prepare python picking to save model behaviour. When using this Streamlit in Python, the developer can load pickled file to retrieve the model behaviour. When user wants to analyse the patient's health condition either then can predict a particular disease or if the report contains parameters which are used to predict other diseases then this analysis will produce maximum identification of relevant diseases. The aim of this article is used to prevent mortality ratio increasing day by day by warning the patients in advance based on their health conditions. Due to many diseases models and predictions done at one place cost of patient analysis can be reduced.

1.1 DIABETES DISEASE

1.1.1 Introduction to Diabetes Disease

Diabetes is the fast growing disease among the people even among the youngsters.

In understanding diabetes and how it develops, we need to understand what happens in the body without diabetes. Sugar (glucose) comes from the foods that we eat,specifically carbohydrate foods. Carbohydrate foods provide our body with its main energy source everyday, even those people with diabetes, needs carbohydrate. Carbohydrate foods include bread,cereal, pasta, rice, fruit, dairy products and vegetables (especially starchy vegetables). When we eat these foods,

the body breaks them down into glucose. The glucose moves around the body in the bloodstream. Some of the glucose is taken to our brain to help us think clearly and function. The remainder of the glucose is taken to the cells of our body for energy and also to our liver, where it is stored as energy that is used later by the body.In order for the body to use glucose for energy, insulin is require .Insulin is a hormone that is produced by the beta cells in the pancreas.Insulin works like a key to a door. Insulin attaches itself to doors on the cell, opening the door to allow glucose to move from the blood stream,through the door, and into the cell.If the pancreas is not able to produce enough insulin(insulin deficiency) or if the body cannot use the insulin it produces (insulin resistance), glucose builds up in the blood stream(hyperglycaemia) and diabetes develops.Diabetes Mellitus means high levels of sugar (glucose) in the blood stream and in the urine.

1.1.2 Types Of Diabetes

Type 1: Diabetes means that the immune system is compromised and the cells fail to produce insulin in sufficient amounts. There are no eloquent studies that prove the causes of type 1 diabetes and there are currently no known methods of prevention.

Type 2: Diabetes means that the cells produce a low quantity of insulin or the body can't use the insulin correctly. This is the most common type of diabetes, thus affecting 90% of persons diagnosed with diabetes. It is caused by both genetic factors and the manner of living.

Gestational diabetes appears in pregnant women who suddenly develop high blood sugar. In two thirds of the cases, it will reappear during subsequent pregnancies. There is a great chance that type 1 or type 2 diabetes will occur after a pregnancy affected by gestational diabetes.

1.1.3 Symptoms Of Diabetes

- Frequent Urination
- Increased thirst
- Tired/Sleepiness
- Weight loss
- Blurred vision
- Mood swings
- Confusion and difficulty concentrating

1.1.4 Causes of Diabetes

Genetic factors are the main cause of diabetes. It is caused by at least two mutant genes in the chromosome 6, the chromosome that affects the response of the body to various antigens. Viral infection may also influence the occurrence of type 1 and type 2 diabetes. Studies have shown that infection with viruses such as rubella, Coxsackievirus, mumps, hepatitis B virus, and cytomegalovirus increase the risk of developing diabetes.

Risk factors for diabetes

Risk factors for diabetes depend on the type of diabetes. Family history may play a part in all types.Environmental factors and geography can add to the risk of type 1 diabetes.

Sometimes family members of people with type 1 diabetes are tested for the presence of diabetes immune system cells (autoantibodies). If you have these autoantibodies, you have an increased risk of developing type 1 diabetes. But not everyone who has these autoantibodies develops diabetes.

Race or ethnicity also may raise your risk of developing type 2 diabetes. Although it's unclear why, certain people —including Black, Hispanic, American Indian and Asian American people — are at higher risk.

Prediabetes, type 2 diabetes and gestational diabetes are more common in people who are overweight or obese.

1.2 PARKINSON'S DISEASE

1.2.1 Introduction to Parkinson's Disease

Parkinson's disease is a progressive disorder that affects the nervous system and the parts of the body controlled by the nerves. Symptoms start slowly. The first symptom may be a barely noticeable tremor in just one hand. Tremors are common, but the disorder may also cause stiffness or slowing of movement.

In the early stages of Parkinson's disease, your face may show little or no expression. Your arms may not swing when you walk. Your speech may become soft or slurred. Parkinson's disease symptoms worsen as your condition progresses over time. Although Parkinson's disease can't be cured, medications might significantly improve your symptoms. Occasionally, your health care provider may

suggest surgery to regulate certain regions of your brain and improve your symptoms.

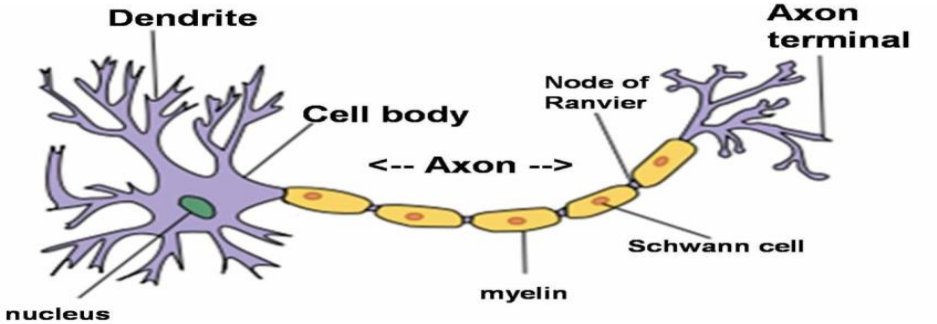


Fig:1.2.1 Structure of Neuron

1.2.2 Types of parkinsons's disease

Idiopathic Parkinsons's : Idiopathic parkinsons means the cause is unknown.Most of the people are suffering from Idiopathic parkinson.The most common symptoms of idiopathic parkinson's are tremor,rigidity and slowness of movement.

Vascular Parkinson's: Vascular parkinson also known as arteriosclerotic parkinson. It affects people with restricted blood supply to the brain.Sometimes people who have had a mild stroke may develop this form of parkinsonism.Common symptoms include problems with memory,sleep,mood and movement.

Drug-induced Parkinson's: Neuroleptic drugs (used to treat schizophrenia and other psychotic disorders), which block the action of the chemical dopamine in the brain, are thought to be the biggest cause of drug-induced parkinsonism. The symptoms of drug-induced parkinsonism tend to stay the same only in rare cases do they progress in the way that Parkinson's symptoms do. Drug-induced parkinsonism only affects a small number of people, and most will recover within months – and often within days or weeks – of stopping the drug that's causing it.

1.2.4 Parkinson's disease symptoms

- Tremor
- Slowed movement (bradykinesia)
- Rigid muscles
- Impaired posture and balance
- Loss of automatic movements

- Speech changes
- Writing changes

1.2.5 Causes of Parkinson's disease

In Parkinson's disease, certain nerve cells (neurons) in the brain gradually break down or die. Many of the symptoms are due to a loss of neurons that produce a chemical messenger in your brain called dopamine. When dopamine levels decrease, it causes atypical brain activity, leading to impaired movement and other symptoms of Parkinson's disease.

Reasons for Parkinson's disease include:

Age : Young adults rarely experience Parkinson's disease. It ordinarily begins in middle or late life, and the risk increases with age. People usually develop the disease around age 60 or older. If a young person does have Parkinson's disease, genetic counseling might be helpful in making family planning decisions. Work, social situations and medication side effects are also different from those of an older person with Parkinson's disease and require special considerations.

Heredity: Having a close relative with Parkinson's disease increases the chances that you'll develop the disease. However, your risks are still small unless you have many relatives in your family with Parkinson's disease.

Sex : Men are more likely to develop Parkinson's disease than women.

Exposure to toxins : Ongoing exposure to herbicides and pesticides may slightly increase your risk of Parkinson's disease.

1.3 INTRODUCTION TO MACHINE LEARNING

Machine Learning may be a sub-area of AI, whereby the term refers to the power of IT systems to independently find solutions to problems by recognizing patterns in databases. In other words: Machine Learning enables IT systems to acknowledge patterns in the idea of existing algorithms and data sets and to develop adequate solution concepts. Therefore, in Machine Learning, artificial knowledge is generated on the idea of experience. In order to enable the software to independently generate solutions, the prior action of people is important. For example, the required algorithms and data must be fed into the systems in advance and the respective analysis rules for the recognition of patterns in the data stock

must be defined. Once these two steps have been completed, the system can perform the following tasks by Machine Learning:

1.Finding, extracting and summarizing relevant data.

2.Making predictions based on the analysis data.

3.Calculating probabilities for specific results.

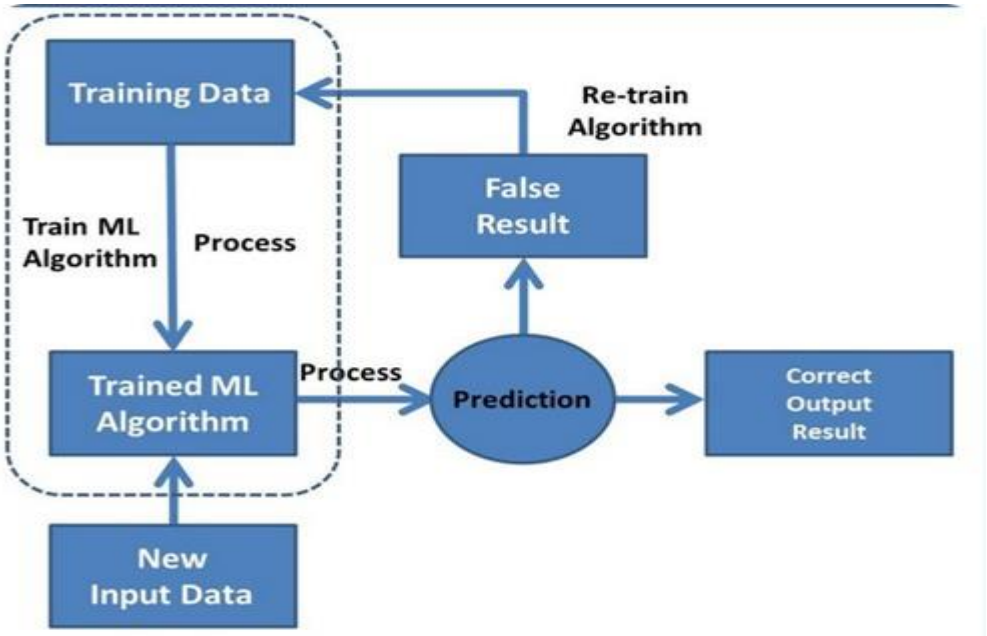


Fig 1.3 Machine learning Algorithm Processing Flow Chart

1.3.1 Supervised Learning

In the course of monitored learning, example models are defined beforehand.So as to make sure an adequate allocation of the knowledge to the respective model groups of the algorithms, these then need to be specified. In other words, the system learns on the idea of given input and output pairs. Within the course of monitored learning, a programmer, who acts as a sort of teacher, provides the acceptable values for specific input. The aim is to coach the system within the context of successive calculations with different inputs and outputs to determine connections.Supervised learning is where you've got input variables (X) and an output variable (Y) and you employ an algorithm to find out the mapping function from the input to the output. Y = f(X) The goal is to approximate the mapping function so well that once you have a new input file (X) that you simply can predict the output variables (Y) for that data. It's called supervised learning because the method of an algorithm learning from the training dataset is often thought of as an educator supervising the training process. We all know the correct answers, the

algorithm iteratively makes predictions on the training data and is corrected. Learning stops when the algorithm achieves a suitable level of performance.

Techniques of Supervised Machine Learning algorithms include linear and logistic regression, multi-class classification, Decision Tree, and Support Vector Machine.

Supervised Learning problems are a kind of machine learning technique often further grouped into Regression and Classification problems. The difference between these two is that the dependent attribute is numerical for regression and categorical for classification:

- **Regression**

Linear regression could also be a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and thus the only output variable (y). More specifically, that y is usually calculated from a linear combination of the input variables (x).When there's one input variable (x), the tactic is mentioned as simple linear regression. When there are multiple input variables, literature from statistics often refers to the tactic as multiple linear regression.

- **Classification**

Classification could also be a process of categorizing a given set of data into classes, It is often performed on both structured or unstructured data. the tactic starts with predicting the category of given data points. The classes are often mentioned as target, label, or categories.

In short, classification either predicts categorical class labels or classification data supported the training set and thus the values(class labels) in classifying attributes and uses it in classifying new data.

There is a variety of classification models. Classification models include Logistic Regression, Decision Tree, Random Forest, Gradient Boosted Tree, One-vs.-One, and Naïve Bayes,SVM.

1.3.2 Unsupervised Learning

In unsupervised learning, AI learns without predefined target values and without rewards. It's mainly used for learning segmentation (clustering). The machine tries to structure and type the info entered consistent with certain characteristics. For instance, a machine could (very simply) learn that coins of various colors are often sorted consistent with the characteristic "color" so as to structure them. Unsupervised Machine Learning algorithms are used when the knowledge used to train is neither classified nor labeled. The system doesn't

determine the right output but it explores the data and should draw inferences from datasets to elucidate hidden structures from unlabeled data. Unsupervised Learning is that the training of Machines using information that's neither classified nor labeled and allowing the algorithm to act thereon information without guidance.

Unsupervised Learning is accessed into two categories of algorithms inherent grouping in the data such as grouping customers by purchasing behaviour.

- **Clustering**

A clustering problem is where you would like to get the inherent grouping in the data such as grouping customers by purchasing behaviour.

- **Association**

An Association rule learning problem is where you would wish to get rules that describe large portions of your data such as folks that buy X also tend to shop for Y.

1.4 PROBLEM STATEMENT

The main aim is to predict the prediction efficiency that would be beneficial for the patients who are suffering from Diabetes,Parkinsons and the percentage of the disease will be reduced. Generally in the first stage, these can be cured by treatment.so it’s important to identify the PD at the early stage for the betterment of the patients.The main purpose of this project is to find the best prediction model i.e. the best machine learning technique which will distinguish the patient from the healthy person. The techniques used in this problem are SVM,Logistic Regression. The experimental study is performed on the dataset of Diabetes,voice dataset of Parkinson’s patients which is downloaded from the Kaggle. The prediction is evaluated using evaluation metrics like confusion matrix, precision, recall accuracy, and f1-score. We used where the important features are taken into consideration to detect Diabetes,Parkinsons.

1.5 OBJECTIVE

The aim of this project is to develop a system which can perform early prediction of diabetes and parkinson disease for a patient with a higher accuracy by combining the results of different macine learning techinques.This project aims to predict diabetes and parkinson disease using machine learning method i.e SVM(Support Vector Machine). This project also aims to propose an effective technique for earlier detection of the diabetes disease and parkinson disease.

CHAPTER – 2

LITERATURE SURVEY

Yasodhaet uses the classification on diverse types of datasets that can be accomplished to decide if a person is diabetic or not. The diabetic patient's data set is established by gathering data from hospital warehouse which contains two hundred instances with nine attributes. These instances of this dataset are referring to two groups i.e. blood tests and urine tests. In this study the implementation can be done by using WEKA to classify the data and the data is assessed by means of 10-fold cross validation approach, as it performs very well on small datasets, and the outcomes are compared. The naïve Bayes, J48, REP Tree and Random Tree are used. It was concluded that J48 works best showing an accuracy of 60.2% among others.

Aiswaryaet aims to discover solutions to detect the diabetes by investigating and examining the patterns originate in the data via classification analysis by using Decision Tree and Naïve Bayes algorithms. The research hopes to propose a faster and more efficient method of identifying the disease that will help in well-timed cure of the patients. that will help in well-timed cure of the patients. the study concluded that J48 algorithm gives an accuracy rate of 74.8% while the naïve Bayes gives an accuracy of 79.5% by using 70:30 split.

Gupta aims to find and calculate the accuracy, sensitivity and specificity percentage of numerous classification methods and also tried to compare and analyse the results of several classification methods in WEKA, the study compares the performance of the classifiers when implemented on some other tools which includes Rapidminer and Matlab using the same parameters (i.e. accuracy, sensitivity and specificity). They applied JRIP, Jgraft and Bayes, Net algorithms. The result shows that Jgraft shows highest accuracy i.e., 81.3%, sensitivity is 59.7% and specificity is 81.4%. It was also concluded that WEKA works best than Matlab and Rapidminner.

Ozcift, proposed a paper "SVM feature selection based rotation forest ensemble classifiers to improve computer-aided diagnosis of Parkinson disease" . In this paper, the author summarizes that improve the PD diagnosis accuracy with the use of support vector machine feature selection. To evaluate the performances the author used accuracy, kappa statistics, and area under the curve of the classification algorithms. The rotation forest ensemble of these classifiers used to increase the performance of the system.

CHAPTER- 3

PROPOSED WORK AND PROCEDURE FOR MODEL DESIGN

3.1 EXISTING SYSTEM

Many of existing analysis involved analysing particular disease. When a user wants to analyse diabetes needs to use one analysis and same user wants to analyse parkinson's disease then user has to use one more model. This is a time taking process. And also if any user having more than one disease but in existing system if it is able to predict only one disease then there is a chance of mortality rate increase due to not able to predict the other disease in advance.

3.2 PROPOSED SYSTEM

In multi disease model prediction, it is possible to predict more than one disease at a time. So user no need to traverse many models to predict the diseases. It will reduce time and also due to predicting multiple diseases at a time there is a chance of reducing mortality rate.

3.2.1 System Architecture

Machine learning has given computer systems the ability to automatically learn without being explicitly programmed. In this project, we used machine learning algorithm (SVM). The architecture diagram describes the high-level overview of major system components and important working relationships. It represents the flow of execution and it involves the following five major steps:

The architecture diagram is defined with the flow of the process which is used to refine the raw data and used for predicting the data. The next step is pre-processing the collected raw data into an understandable format. Then we have to train the data by splitting the dataset into train data and test data.

The Diabetes and Parkinson's data is evaluated with the application of a machine learning algorithm that is SVM and the classification accuracy of this model is found. After training the data with these algorithms we have to test on the same algorithms. Finally, the result of these algorithm is compared on the basis of classification accuracy.

- Speech Dataset
- Pre-processing data
- Training data
- Apply Machine Learning Algorithm

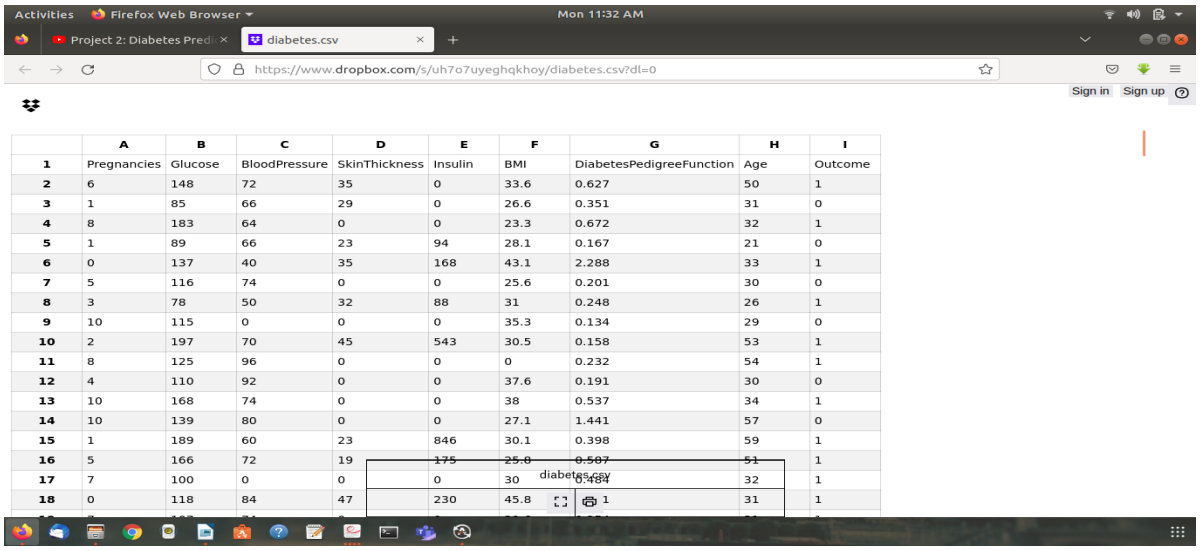
- SVM
- Testing Data

3.3 MODULES DIVISION

Let us discuss about the various modules in our proposed system and what each module contributes in achieving our output.

3.3.1 Speech Dataset

The main aim of this step is to spot and acquire all data-related problems. During this step, we'd like to spot the various data sources, as data are often collected from various sources like files and databases. The number and quality of the collected data will determine the efficiency of the output. The more are going to be the info, the more accurate are going to be the prediction. We've collected our data from the Kaggle website.



	A	B	C	D	E	F	G	H	I
1	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0
13	10	168	74	0	0	38	0.537	34	1
14	10	139	80	0	0	27.1	1.441	57	0
15	1	189	60	23	846	30.1	0.398	59	1
16	5	166	72	19	175	25.8	0.587	51	1
17	7	100	0	0	0	30	0.484	32	1
18	0	118	84	47	230	45.8	0.551	31	1

Fig 3.3.1.1 Sample of acquired dataset from kaggle for Diabetes

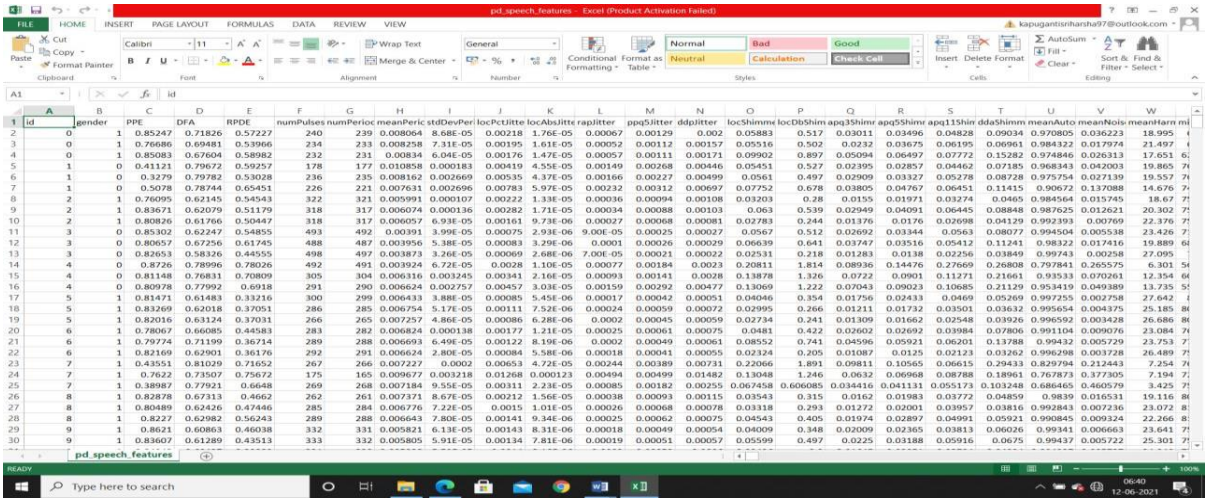


Fig 3.3.1.2 Sample of acquired speech dataset from kaggle for Parkinson disease

In the above Fig-3.3.1.2, we can see the speech dataset that has collected from kaggle website. This acquired dataset has around 755 patient's data and each row has 755 different features. But in this paper, we chosen 10 main features that required to find the prediction.

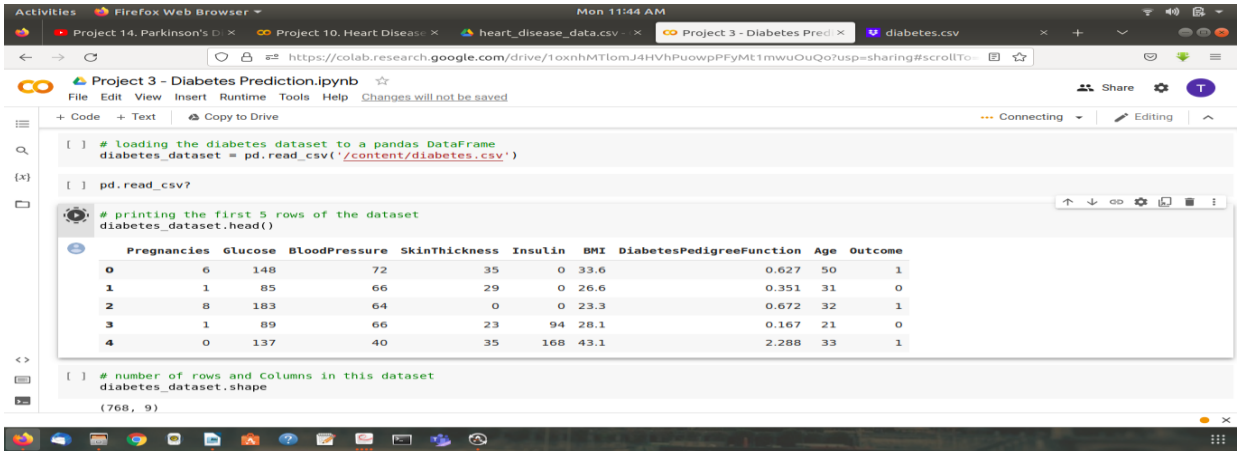


Fig 3.3.1.3 Reading the dataset from the CSV file into notebook

The dataset we chose is in the form of CSV (Comma Separated Value) file. After acquiring the data our next step is to read the data from the CSV file into the Google colab also called a Python notebook. Python notebook is used in our project for data pre-processing, features selection, and for model comparison. In the fig-3.3, we have shown how to read data from CSV files using the inbuilt python functions that are part of the pandas library.

3.3.2 Data Pre-Processing

The main aim of this step is to study and understand the nature of data that was acquired in the previous step and also to know the quality of data. A real-world data generally contains noises, missing values, and maybe in an unusable format that cannot be directly used for machine learning models. Data pre-processing is a required task for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model. Identifying duplicates in the dataset and removing them is also done in this step.

Actually, in this dataset, we have 755 features out of which some may not be useful in building our model. So, we have to leave out all those unnecessary features which are not responsible to produce the output. If we take more features in this model the accuracy we got is less. When we check the correlation of the features, some of them are the same.

3.3.3 Training data

Splitting the dataset into Training set and testing set:

In machine learning data pre-processing, we have to break our dataset into both training set and test set. This is often one among the crucial steps of knowledge pre-processing as by doing this, we will enhance the performance of our machine learning model.

Suppose, if we've given training to our machine learning model by a dataset and that we test it by a totally different dataset. Then, it'll create difficulties for our model to know the correlations between the models.

If we train our model alright and its training accuracy is additionally very high, but we offer a replacement dataset there to, then it'll decrease the performance. So we always attempt to make a machine learning model which performs well with the training set and also with the test dataset.

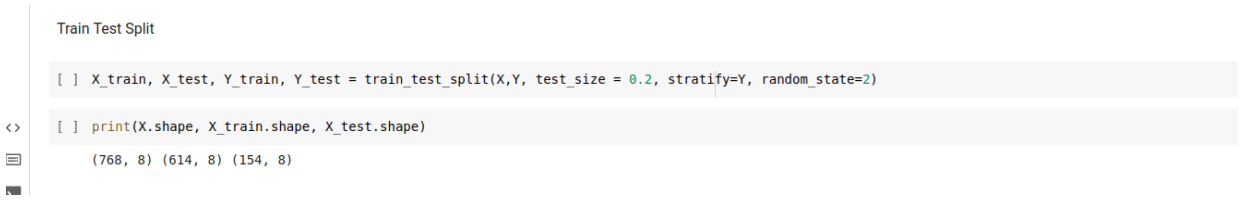


Fig 3.3.3 Splitting dataset into training data and test data

Usually, we split the dataset into train and test in the ratio of 7:3 i.e., 70 percent of data is used for training and 30 percent of data is used for testing the model. We have done it in the same way and it has been shown in the above Fig 3.3.3.

3.3.4 Apply Machine Learning Algorithm

Now, we've both the train and test data. The subsequent step is to spot the possible training methods and train our models. As this is often a classification problem, we've used classification method SVM. Each algorithm has been run over the training dataset and their performance in terms of accuracy is evaluated alongside the prediction wiped out the testing data set.

3.3.4.1 Support vector machine

Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression

problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine.

Support Vector Machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. SVMs have their unique way of implementation as compared to other machine learning algorithms. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables.

The followings are important concepts in SVM -

Support Vectors - Data Points that are closest to the hyperplane are called support vectors. Separating line will be defined with the help of these data points.

Hyperplane - As we can see in the below diagram, it is a decision plane or space which is divided between a set of objects having different classes.

Margin - It may be defined as the gap between two lines on the closest data points of different classes. It can be calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin is considered as a bad margin.

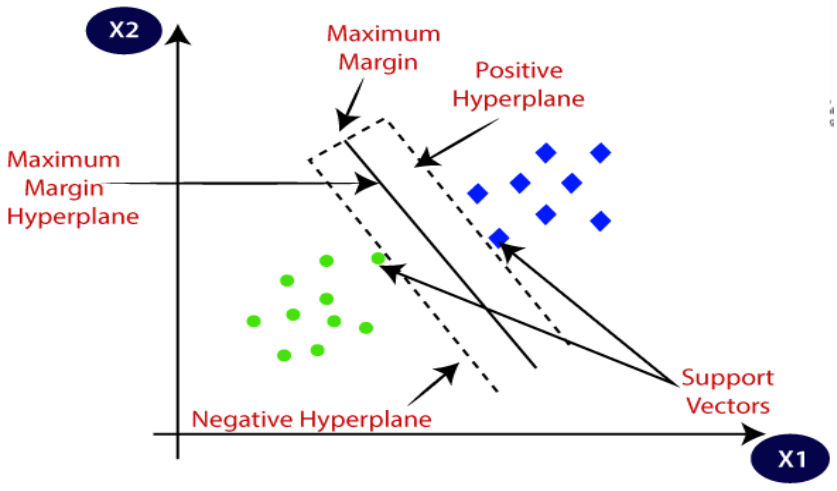


Fig 3.3.4.1.1 Support Vector Machine

The objective of the support vector machine algorithm is to find a hyperplane in an N- dimensional space (N - the number of features) that distinctly classifies the data points.

The advantages of support vector machines are:

● Effective in high dimensional spaces.

● Still effective in cases where the number of dimensions is greater than the number of samples.

● Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

● Versatile: different kernel functions can be specified for the decision function.

Common kernels are provided, but it is also possible to specify custom kernels.

Fig 3.3.4.1.2 Training the data by using SVM

3.3.5 Testing Data

Once multiple disease Prediction model has been trained on the pre-processed dataset, then the model is tested using different data points. In this testing step, the model is checked for correctness and accuracy by providing a test dataset to it. All the training methods need to be verified for finding out the best model to be used test dataset. These predicted values on testing data are used for model comparison and accurate calculation.

CHAPTER 4

EXPERIMENTAL STUDY

4.1 SYSTEM CONFIGURATION

4.1.1 Software Requirements

1. Software:
   Spyder
   Google Colab
2. Tools : Web Browser ,Google Chrome
3.Python Libraries : numpy, pandas, streamlit , sklearn , pickle**.**

4.1.2 Introduction to Python

Python is an interpreter, high-level, general-purpose programming language. Python is simple and easy to read syntax emphasizes readability and thus reduces system maintenance costs. Python supports modules and packages, which promote system layout and code reuse. It saves space but it takes a rather higher time when its code is compiled. Indentation must be taken care while coding.

Python does the following:

- ➔ Python are often used on a server to make web applications.
- ➔ It connects the database systems. It also read and modify files.
- ➔ It often able to handle big data and perform complex mathematics.
- ➔ It can be used for production-ready software development.

Python has many inbuilt library functions that can be used easily for working with machine learning algorithms. All the necessary python libraries must be pre-installed using “pip” command.

4.1.3 Introduction to Streamlit

Streamlit is meant to help Data scientists or machine learning engineers who are not web developers and they're not interested in spending weeks learning to use these frameworks to build web apps.

The best thing about Streamlit is that you don't even need to know the basics of web development to get started or to create your first web application. So

if you're somebody who's into data science and you want to deploy your models easily, quickly, and with only a few lines of code, Streamlit is a good fit.

4.1.4 Python Libraries

NumPy

NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the elemental package for scientific computing with Python. It contains various features including these important ones:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Tools for integrating C/C++ and Fortran code
- Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, NumPy also can be used as an efficient multi-dimensional container of generic data.

Pandas

Pandas is an open-source library that’s built on top of NumPy library. It is a Python package that gives various data structures and operations for manipulating numerical data and time series. It is fast and it has high-performance & productivity for users. It provides high-performance and is easy-to-use data structures and data analysis tools for the Python language. Pandas is employed during a wide range of fields including academic and commercial domains including economics, Statistics, analytics, etc.

Sklearn

Scikit-learn (Sklearn) is that the most useful and robust library for machine learning in Python. It is an open-source Python library that implements a variety of machine learning, pre-processing, cross-validation and visualization algorithms employing a unified interface. Sklearn provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via 36a consistence interface in Python. This library, which is essentially written in Python, is made upon NumPy, SciPy and Matplotlib.

Pickle

Python pickle module is employed for serializing and de-serializing a Python object structure. Pickling is a way to convert a python object (list, dict, etc.) into a character stream. The idea is that this character stream contains all the information necessary to reconstruct the thing in another python script. Pickling is beneficial for applications where you would like a point of persistency in your data. Your program's state data are often saved to disk, so you'll continue working on it later on.

4.2 CODE

```
import pickle
import streamlit as st
from streamlit option menu import option_menu

# loading the saved models

diabetes_model = pickle.load(open('/home/mrudhula/Desktop/Multiple disease prediction/savedmodels-20221213T062058Z-001/saved models/diabetes_model.sav', 'rb'))

parkinsons_model = pickle.load(open('/home/mrudhula/Desktop/Multiple disease prediction/savedmodels-20221213T062058Z-001/saved models/parkinsons_model.sav', 'rb'))

# sidebar for navigation
with st.sidebar:

    selected = option_menu('Multiple Disease Prediction System',

                           ['Diabetes Prediction',
                            'Parkinsons Prediction'],
                           icons=['activity','person'],
                           default_index=0)
# Diabetes Prediction Page
if (selected == 'Diabetes Prediction'):
    # page title
    st.title('Diabetes Prediction using ML')


    # getting the input data from the user
    col1, col2, col3 = st.columns(3)
```

```
    with col1:
        Pregnancies = st.text_input('Number of Pregnancies')

    with col2:
        Glucose = st.text_input('Glucose Level')

    with col3:
        BloodPressure = st.text_input('Blood Pressure value')

    with col1:
        SkinThickness = st.text_input('Skin Thickness value')

    with col2:
        Insulin = st.text_input('Insulin Level')

    with col3:
        BMI = st.text_input('BMI value')

    with col1:
        DiabetesPedigreeFunction = st.text_input('Diabetes Pedigree Function value')

    with col2:
        Age = st.text_input('Age of the Person')

    # code for Prediction
    diab_diagnosis = ''

    # creating a button for Prediction

    if st.button('Diabetes Test Result'):
        diab_prediction = diabetes_model.predict([[Pregnancies, Glucose, Blood
Pressure, SkinThickness, Insulin, BMI, Diabetes Pedigree Function, Age]])
         if (diab_prediction[0] == 1):
          diab_diagnosis = 'The person is diabetic'
        else:
          diab_diagnosis = 'The person is not diabetic'

    st.success(diab_diagnosis)

# Parkinson's Prediction Page
if (selected == "Parkinsons Prediction"):
     # page title
    st.title("Parkinson's Disease Prediction using ML")
    col1, col2, col3, col4, col5 = st.columns(5)
```

```
with col1:
    fo = st.text_input('MDVP:Fo(Hz)')

with col2:
    fhi = st.text_input('MDVP:Fhi(Hz)')

with col3:
    flo = st.text_input('MDVP:Flo(Hz)')

with col4:
    Jitter_percent = st.text_input('MDVP:Jitter(%)')

with col5:
    Jitter_Abs = st.text_input('MDVP:Jitter(Abs)')

with col1:
    RAP = st.text_input('MDVP:RAP')

with col2:
    PPQ = st.text_input('MDVP:PPQ')

with col3:
    DDP = st.text_input('Jitter:DDP')

with col4:
    Shimmer = st.text_input('MDVP:Shimmer')

with col5:
    Shimmer_dB = st.text_input('MDVP:Shimmer(dB)')
     with col1:
    APQ3 = st.text_input('Shimmer:APQ3')
  with col2:
    APQ5 = st.text_input('Shimmer:APQ5')

with col3:
    APQ = st.text_input('MDVP:APQ')

with col4:
    DDA = st.text_input('Shimmer:DDA')

with col5:
    NHR = st.text_input('NHR')

with col1:
```

```
        HNR = st.text_input('HNR')

    with col2:
        RPDE = st.text_input('RPDE')

    with col3:
        DFA = st.text_input('DFA')

    with col4:
        spread1 = st.text_input('spread1')

    with col5:
        spread2 = st.text_input('spread2')

    with col1:
        D2 = st.text_input('D2')

    with col2:
        PPE = st.text_input('PPE')

    # code for Prediction
    parkinsons_diagnosis = ''

      # creating a button for Prediction
    if st.button("Parkinson's Test Result"):
        parkinsons_prediction = parkinsons_model.predict([[fo, fhi, flo,
Jitter_percent, Jitter_Abs, RAP,
PPQ,DDP,Shimmer,Shimmer_dB,APQ3,APQ5,APQ,DDA,NHR,HNR,RPDE,DF
A,spread1,spread2,D2,PPE]])
         if (parkinsons_prediction[0] == 1):
          parkinsons_diagnosis = "The person has Parkinson's disease"
        else:
          parkinsons_diagnosis = "The person does not have Parkinson's disease"
         st.success(parkinsons_diagnosis)
```

4.3 ACCURACY

The below table shows the accuracy of SVM used for Diabetes, Parkinson disease.

Type of algorithm	Training Accuracy %	Testing Accuracy %
Decision Tree	80	70
Random Forest	85	71
Support Vector Machine	88	72

4.4 RESULTS

To demonstrate the results of our project, we take the remaining test data and it is tested using algorithms. After that our trained model to ready to predict the disease is present or not. The test accuracy is done in the Google colab which is our python notebook.

Below we described how the algorithms are processed for all diseases. First, SVM algorithm is trained with the training dataset and later it was tested with the remaining test data. In Fig 4.4.1, a screenshot of our notebook is showing that how the process of SVM algorithm is done for diabetes and the accuracy the model returns and it is of 78.67%.

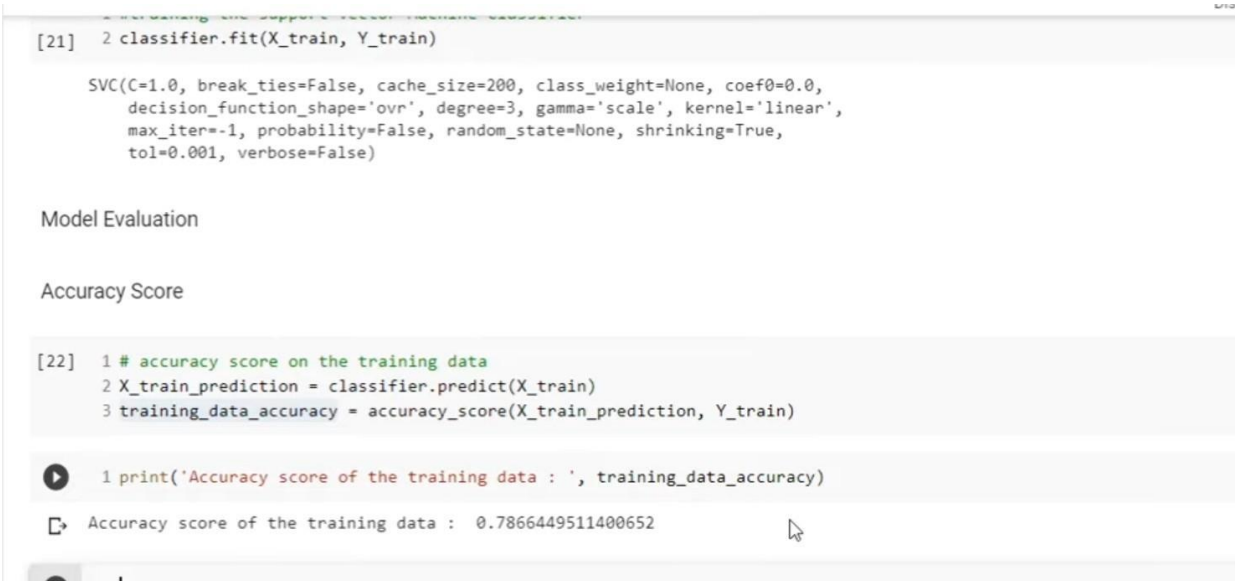


Fig 4.4.1 Accuracy score on training data of SVM algorithm (Diabetes)

Second, SVM algorithm is trained with the training dataset and later it was tested with the remaining test data. In Fig 4.4.3, a screenshot of our notebook is showing that how the process of SVM algorithm is done for Parkinson disease and the accuracy the model returns and it is of 80.46%.

Fig 4.4.2 Accuracy score on training data of SVM algorithm (Parkinson disease)

The model is loaded into the pickle file and that file is opened in the chrome as a website and compares the user input values with this corresponding model. Finally it results with a text message displaying that either the patient having diabetes disease or not.

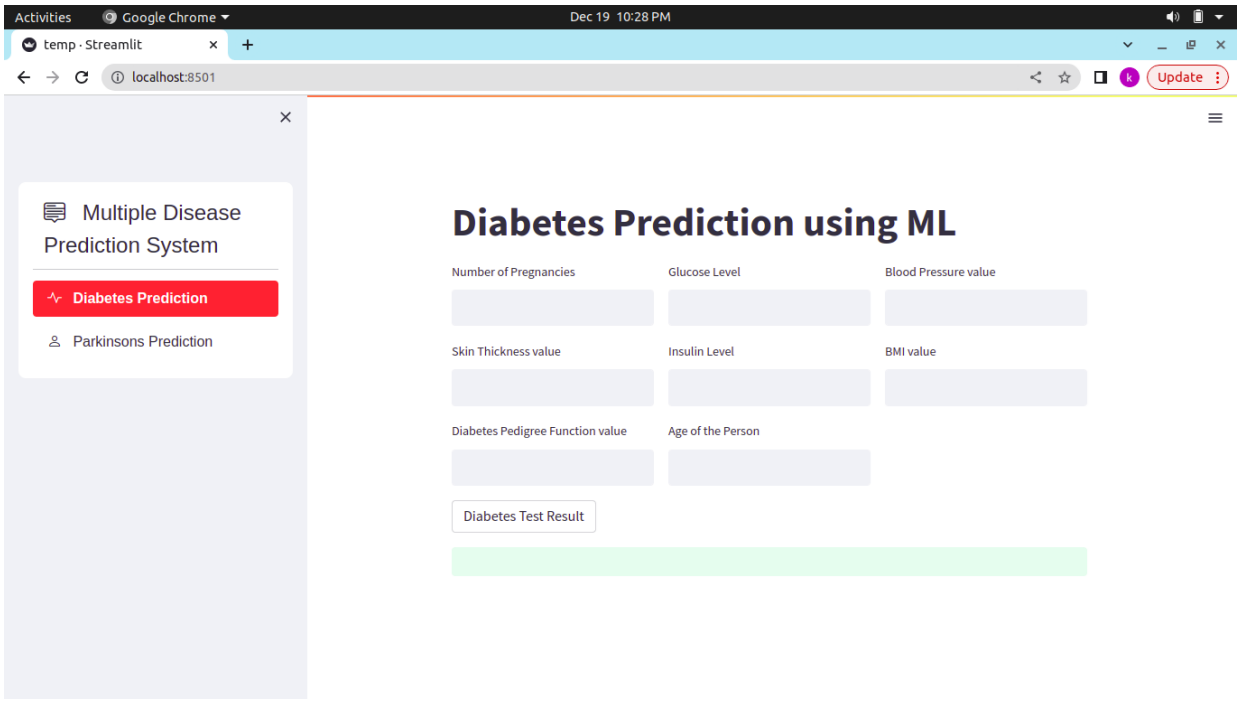


Fig 4.4.3 User Interface to enter the patient details to predict diabetes

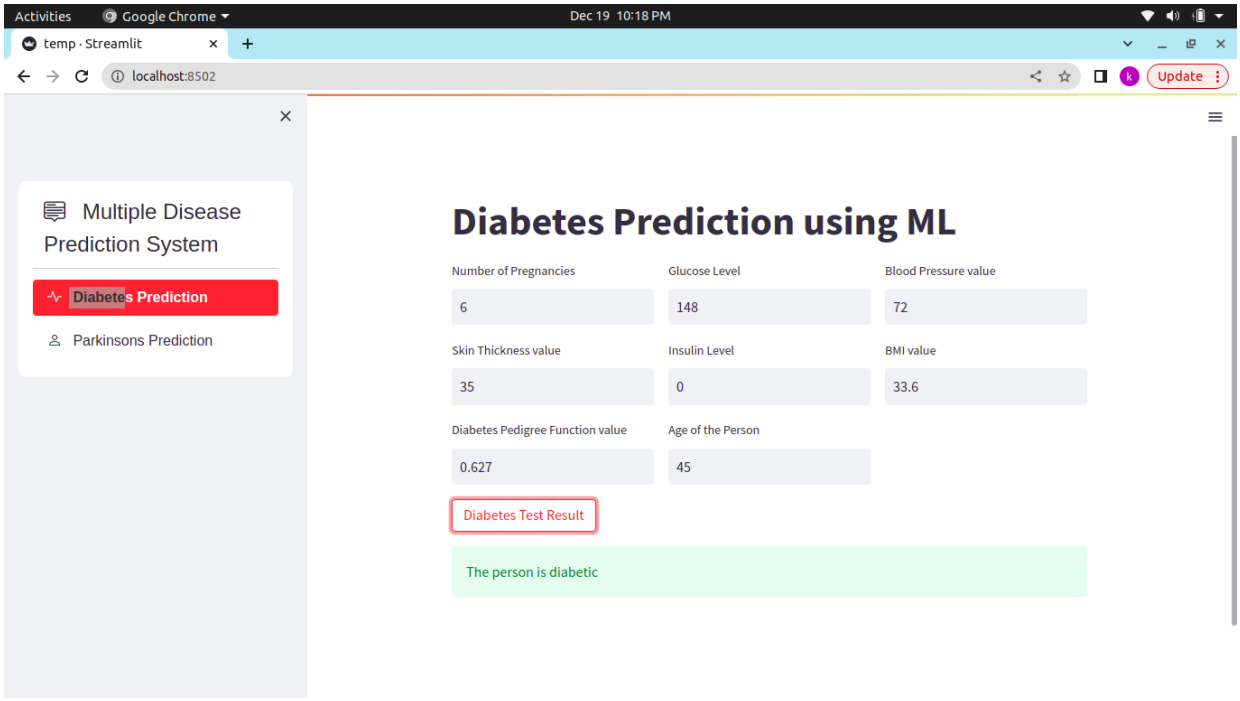


Fig 4.4.4 Predicted Result showing for fig 4.4.3 details

The model is loaded into the pickle file and that file is opened in the chrome as a website and compares the user input values with this corresponding model. Finally it results with a text message displaying that either the patient having Parkinson's disease or not.

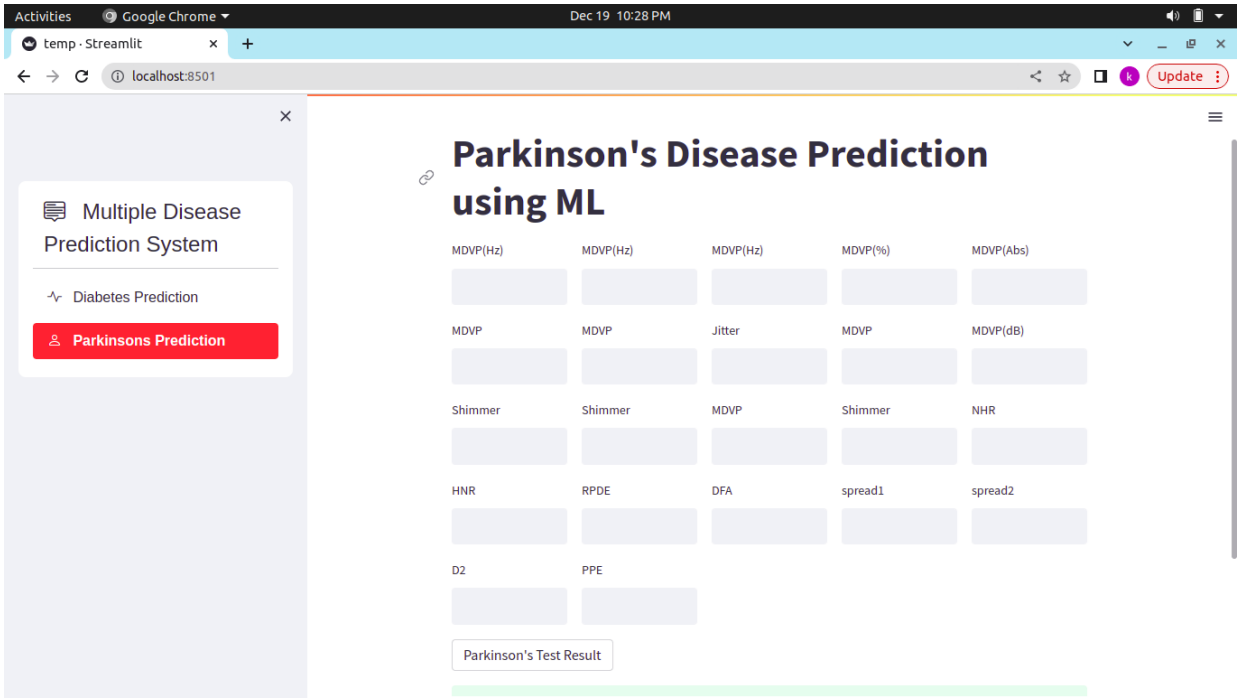


Fig 4.4.5 User Interface to enter the patient details to predict Parkinson disease

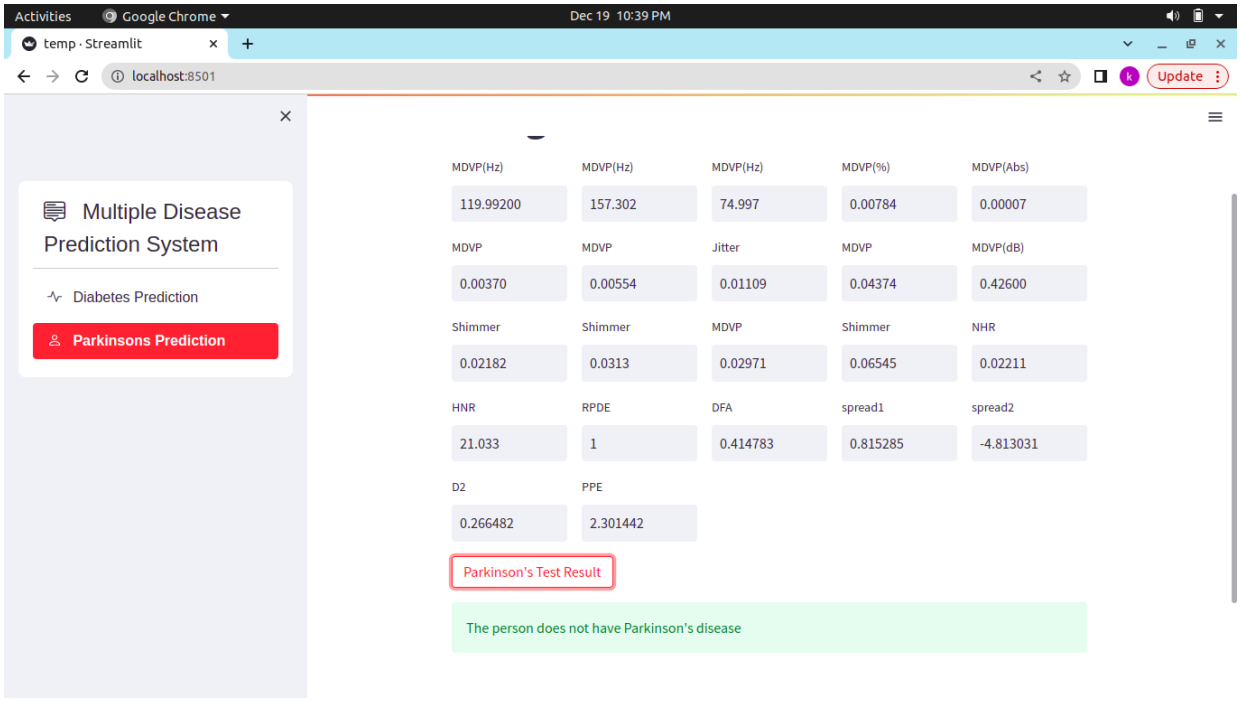


Fig 4.4.6 Predicted Result showing for fig 4.4.5 details

CHAPTER-5

CONCLUSION AND FUTURE WORK

5.1 CONCLUSION

Diabetes is a very serious disease with many life threatening consequences, but if it is taken care of properly, diabetes can live a normal life. Parkinson's disease is the second most dangerous neurodegenerative disease which has no cure till now and to make it reduce prediction is important. In this project, we have used prediction model(SVM) to predict the Diabetes, Parkinson's disease which are Machine Learning Technique.

The dataset is trained using these models and we also compared these different models built. We have used the dataset that contains some features of the patients which is available in the Kaggle website. The dataset consists of more than 700 features and 750 patient details. The models are built using the some best features which were identified by feature selection.

This system we designed can make the predictions of the Diabetes and Parkinson's disease.

5.2 FUTURE WORK

In future, these models can be trained with different datasets that have best features and can be predicted more accurately. If the accuracy rate increases, it can be used by the laboratories and hospitals so that it is easy to predict in early stages. This models can be also used with different medical and disease datasets. In future the work can be extended by building a hybrid model that can find more than three diseases with an accurate dataset and that dataset has common features of two diseases. In future the work can extended to build a model that may extract more important features among all features in the dataset so that it produce more accuracy.

CHAPTER 6

REFERENCES

[1] Md. Redone Hassan,etal, “A Knowledge Base Data Mining based on Parkinson’sDisease” International Conference on System Modelling & Advancement in ResearchTrends, 2019**.**

[2] R. P. Duncan, A. L. Leddy, J. T. Cavanaugh et al., “Detecting and predicting balance decline in Parkinson disease: a prospective cohort study” Journal of Parkinson’s Disease,vol-5, no. 1, pp. 131–139, 2015.

[3 T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins and R. H. Ghomi,"Parkinson’s Disease Diagnosis Using Machine Learning and Voice," IEEE Signal Processing in Medicine and Biology Symposium (SPMB), pp.1-7, doi:10.1109/SPMB.2018.8615607, 2018.

[4] Yashoda and M. Kannan, "Analysis of a Population of Diabetic Patients Databases in Waikato", International Journal of Scientific & Engineering Research, vol. 2, no. 5, 2011.

[5] A. Ayer, J. S and R. Sumbala, "Diagnosis of Diabetes Using Classification Mining Techniques", IJDKP, vol. 5, no. 1, pp. 01-14, 2015.

[6] NIyati Gupta, A. Rawal, and V. Narasimhan, “Accuracy, Sensitivity andSpecificity Measurement of Various Classification Techniques on Healthcare Data", IOSR Journal of Computer Engineering, vol. 11, no. 5, pp. 70-73,2013.