VISVESVARAYA TECHNOLOGICAL UNIVERSITY
"JNANA SANGAMA", BELAGAVI– 590018

PROJECT PHASE 2

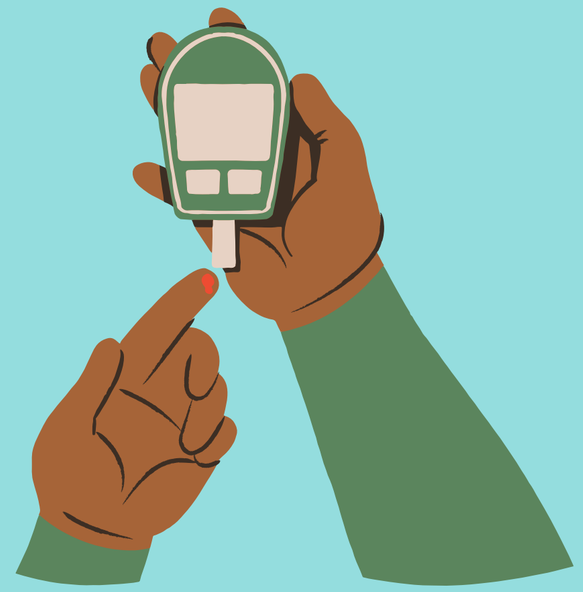
(18CSP83)

PREDICTION OF DIABETES THROUGH MEDICAL DATASET USING ML

PRESENTED BY
AYUSH KUMAR(1AT19IS021)
HARSHITHA V(1AT19IS038)
INCHARA A(1AT19IS043)

UNDER THE GUIDANCE OF
MR. OMPRAKASH B
ASSISTANT PROFESSOR
DEPT. OF ISE, ATRIA IT

ATRIA INSTITUTE OF TECHNOLOGY
BANGALORE-24, KARNATAKA
DEPARTMENT OF INFORMATION SCIENCE & ENGINEERING



Prediction of Diabetes through medical dataset using ML

Paper Publication Details

Paper Title : Prediction of Diabetes through medical dataset using ML

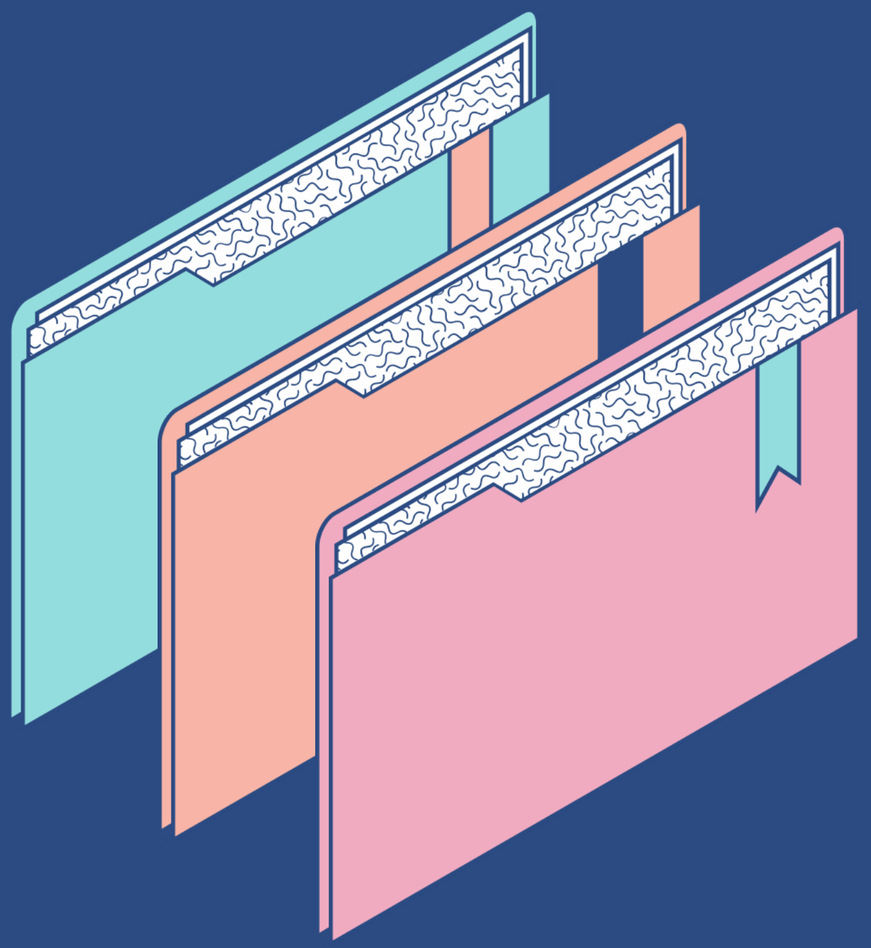
Journal Name : International Journal of Novel Research and Development

Journal ISSN : 2456-4184

Journal Vol : Volume 8 | Issue 4 | April-2023

Journal website:

http://www.ijnrd.org/viewpaperforall.php?paper=IJNRD2304172



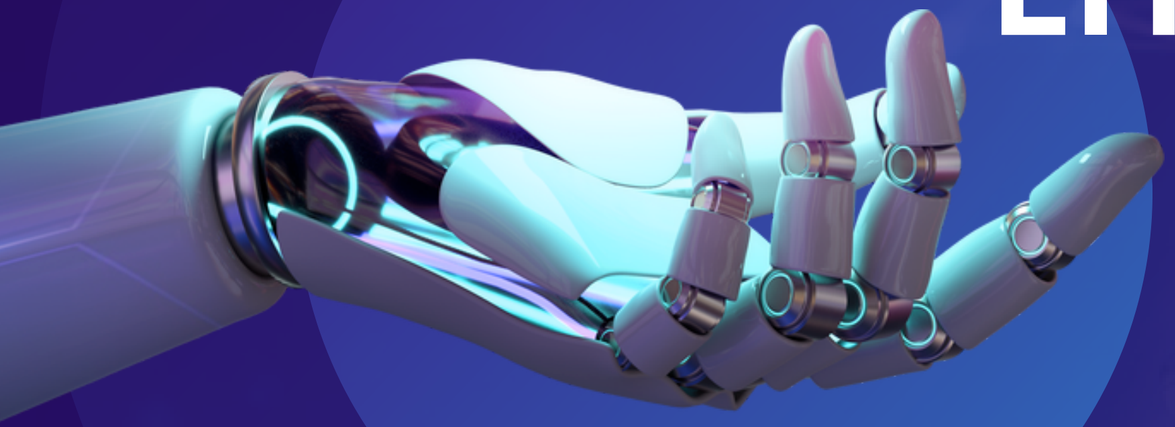
CONTENTS

INTRODUCTION

- Diagnosis of Diabetes disease at beginning stage is important for healthier treatment.
- In today's scenario equipment's like sensors are used for discovery of infections. Few algorithms (Logistic Regression , Artificial Neural Network, SVM, Random forest) were implemented for classification using python platform.
- To deal with Classification and prediction of medical datasets there are many challenges.
- Machine learning technique has been a tremendous support for making prediction of a particular system by training. **Due to recent advances in machine learning, medical analysis increases diagnostic accuracy, reduces cost and reduces human resource.**
- Data Mining methods can improve the quality of medical decisions significantly.
- We propose a model for medical predictions based on **Random Forest.**

- In this project, we will explore the use of ML algorithms to analyze medical datasets and develop models for predicting diabetes.
- We will use a variety of ML techniques, to identify patterns in the data and make accurate predictions about patient outcomes.
- Our ultimate goal is to develop an accurate and reliable predictive model that can assist healthcare professionals in making informed decisions about patient care and improve the overall management of diabetes.

LITERATURE SURVEY



Literature Survey

Sl . No	Paper Title	Authors, Publisher & Publication Year	Problem Identified	Techniques used	Outcome
1	Predication of diabetes empowered with fused machine learning	USAMA AHMED, SHABIB AFTAB,MUNIR AHMAD. January 11, **2022**.	Inefficiency of fuzzy system.	Support Vector Machine (SVM) and Artificial Neural Network (ANN) models.	The proposed fused ML model has a prediction accuracy of 80.48.
2	Machine Learning Tools for Long-Term Type 2 Diabetes Risk Prediction	NIKOS FAZAKIS , OTILIA KOCSIS , NIKOS FAKOTAKIS. July 20, **2021**.	Inefficiency to predict type 1	Sensitivity and AUC of the ML model based on a bi-objective genetic algorithm.	Ensemble methods constitute a useful tool for pre-dicting type 2 diabetes.
3	Early Prediction of Diabetes Using an Ensemble of Machine Learning Models	Aishwariya Dutta, Md. Kamrul Hasan, September **2022**	Lack of accuracy	ML techniques, including SVM, AB, Bagging, KNN algorithm	Ensuring robust and accurate prediction, enabled this research to achieve its goal of making an early prediction of diabetes.
4	Prediction and diagnosis of future diabetes risk: a machine learning approach	Roshan Birjais,Ashish Kumar Mourya,Ritu Chauhan,Harleen Kaur, **2019**	Compare various machine learning algorithms used for prediction	Different techniques like Gradient Boosting, Logistic Regression and Naive Bayes,	Boosting, 79% for Logistic Regression and 77% for Naive Bayes.

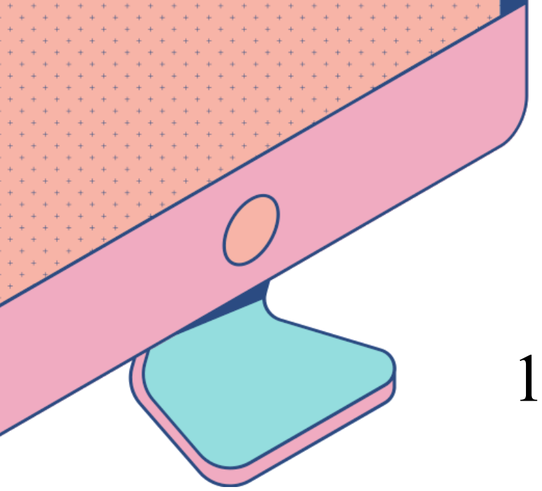
Sl . No	Paper Title	Authors, Publisher & Publication Year	Problem Identified	Techniques used	Outcome
5	Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review	Jyotismita Chaki a,S. Thillai Ganesh b, S.K Cidham b, S. Ananda Theertan, 4 July **2020**	Diabetes Mellitus (DM) is a condition induced by unregulated diabetes that may lead to multi-organ failure in patients.	pre-processing methods, feature extraction methods, machine learning-based identification, classification, and diagnosis of DM.	DM detection and self-management.
6	Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus(T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study	ASIF HASSAN SYED AND TABREJ KHAN,October 30, **2020**	high risk of developing Type 2 Diabetes mellitus (T2DM) in the western region of Saudi Arabia.	Chi-Squared test and binary logistic regression	The tuned two-class Decision Forest (DF) model showed better performance with an average F1score of 0.8453 ± 0.0268.
7	Ethical and Legal Issues for Medical Data Mining	Ashwinkumar U.M. and Dr. Anandakumar K.R. **2017**	the main problems associated with tackle square measure moral, legal and social aspects.	information sets that square measure massive, complex, heterogeneous, hierarchical , statistic and of variable quality	Issues were resolved
8	Design and Evaluation of Logistic Regression Model for Pattern Recognition Systems	Pranav Rao and Manikandan J ,**2019**	the popularity accuracy is increased on exploitation planned mapping functions	logistical regression	style pattern recognition systems exploitation logistical regression model and few mapping functions area unit planned for a similar.
9	Input Feature Extraction for Multilayered Perceptrons Using Supervised Principal Component Analysis	Santosh S, Meenakshi Y, **2020**	to handle these ill-structured medicine information	data processing techniques	square measure the classification and prediction of medical datasets
10	5G-Smart Diabetes: Toward Personalized Diabetes Diagnosis with Healthcare Big Data Clouds	Min Chen, Jun Yang, Jiehan Zhou, Yixue Hao, Jing Zhang, and Chan-Hyun Youn, April **2018**	diabetes detection model lacks a data sharing mechanism	convolutional neural network (CNN), machine learning, and big data to generate comprehensive sensing and analysis for patients suffering from diabetes.	wearable 2.0, machine learning, and big data to generate comprehensive sensing and analysis for patients suffering from diabetes.

Proposed System

1. **Data collection**: Gathering relevant medical data of patients from various sources such as electronic health records (EHRs), lab results, and imaging data.
2. **Data pre-processing**: Preparing the medical dataset for analysis by cleaning, filtering, and transforming the data.
3. **Feature selection**: Identifying the most relevant features in the dataset that can be used for predicting diabetes.
4. **Model selection**: Choosing the appropriate ML algorithm based on the dataset and the problem to be solved.
5. **Model training**: Training the ML model using the prepared dataset.
6. **Model evaluation**: Evaluating the performance of the trained model by comparing the predicted outcomes with the actual outcomes.

Advantages

1. **Early detection**: The ML model can detect diabetes at an early stage, allowing healthcare professionals to take proactive measures to prevent or delay its onset, reducing the risk of complications.
2. **Improved accuracy**: The ML model can analyze vast amounts of patient data and identify patterns that may not be evident to healthcare professionals, improving the accuracy of diagnosis and treatment.
3. **Efficient use of resources**: The ML model can reduce the need for unnecessary tests and procedures, saving time and resources.
4. **Better patient outcomes:** Early detection and personalized treatment can improve patient outcomes and quality of life.



Objectives

1. **Improve diagnosis:** The model should improve the accuracy of diagnosis and reduce the need for unnecessary tests and procedures.
2. **Improve patient outcomes:** The model should help improve patient outcomes by reducing the risk of complications and improving the quality of life.
3. **Cost-effective:** The model should be cost-effective and efficient, reducing the burden on healthcare resources.
4. **Data analysis:** The model should analyze large medical datasets and identify patterns that may not be evident to healthcare professionals, leading to better understanding of the disease and its management.

Methodology/Algorithms

In this project we have compared the dataset using 4 different ML algorithm , they are :

1.Logistic Regression

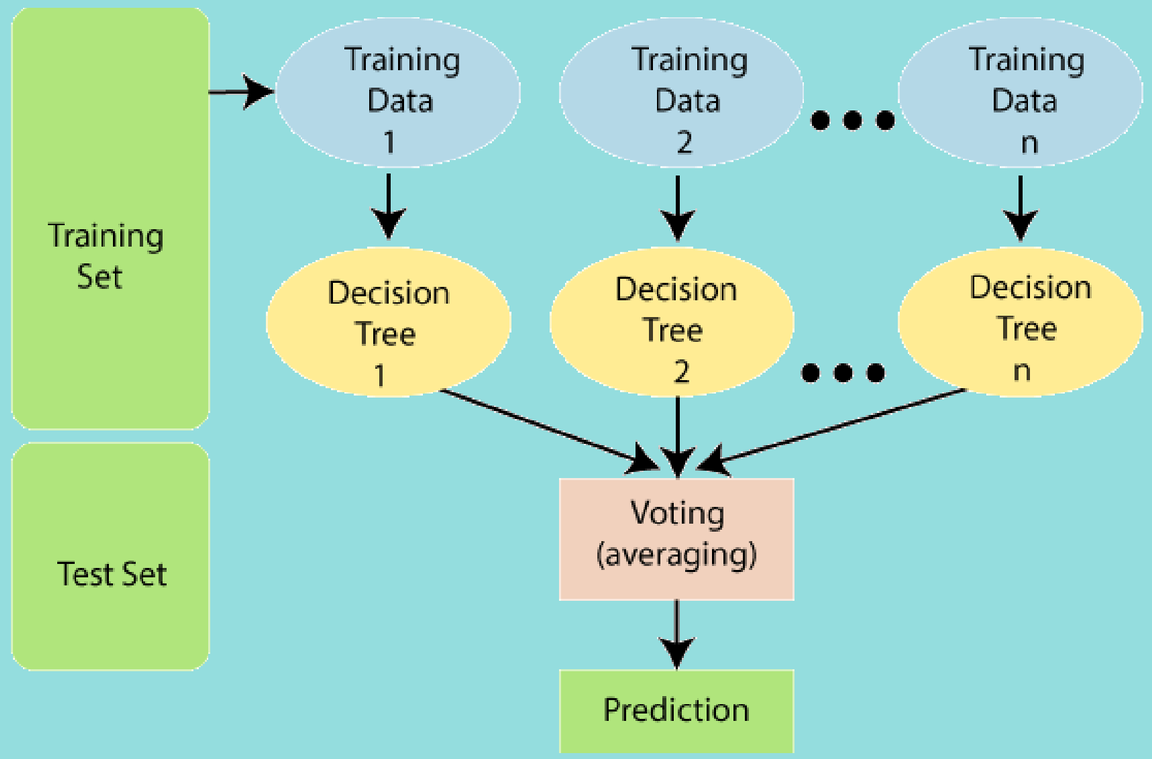
2.Support Vector Classifier

3.KNeighbor Classifier

4.Random Forest Classifier

RANDOM FOREST

Random Forest is a machine learning algorithm that is used for classification, regression, and other tasks that involve predicting a target variable based on a set of input features. It is an ensemble method that combines the predictions of multiple decision trees to improve accuracy and reduce overfitting.

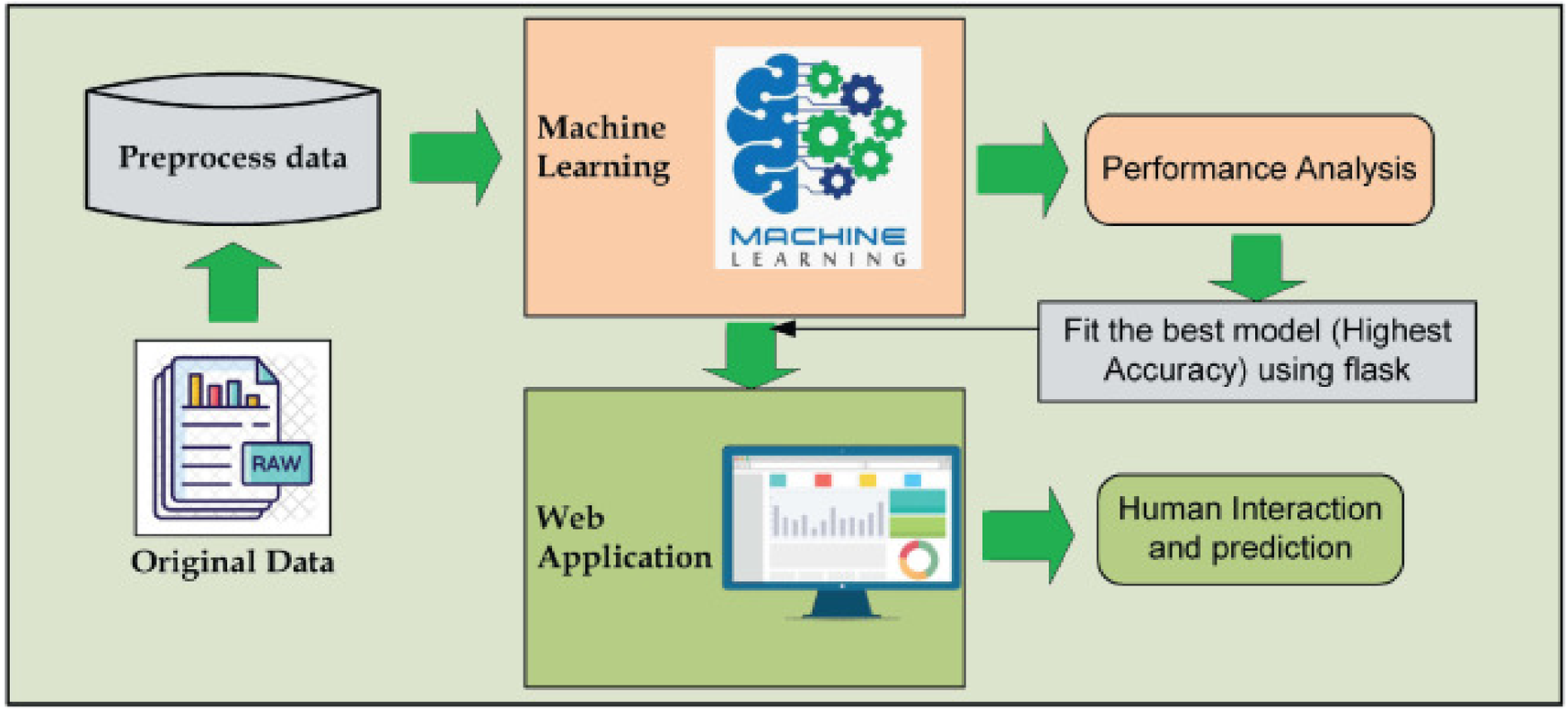


Random Forest Classifier

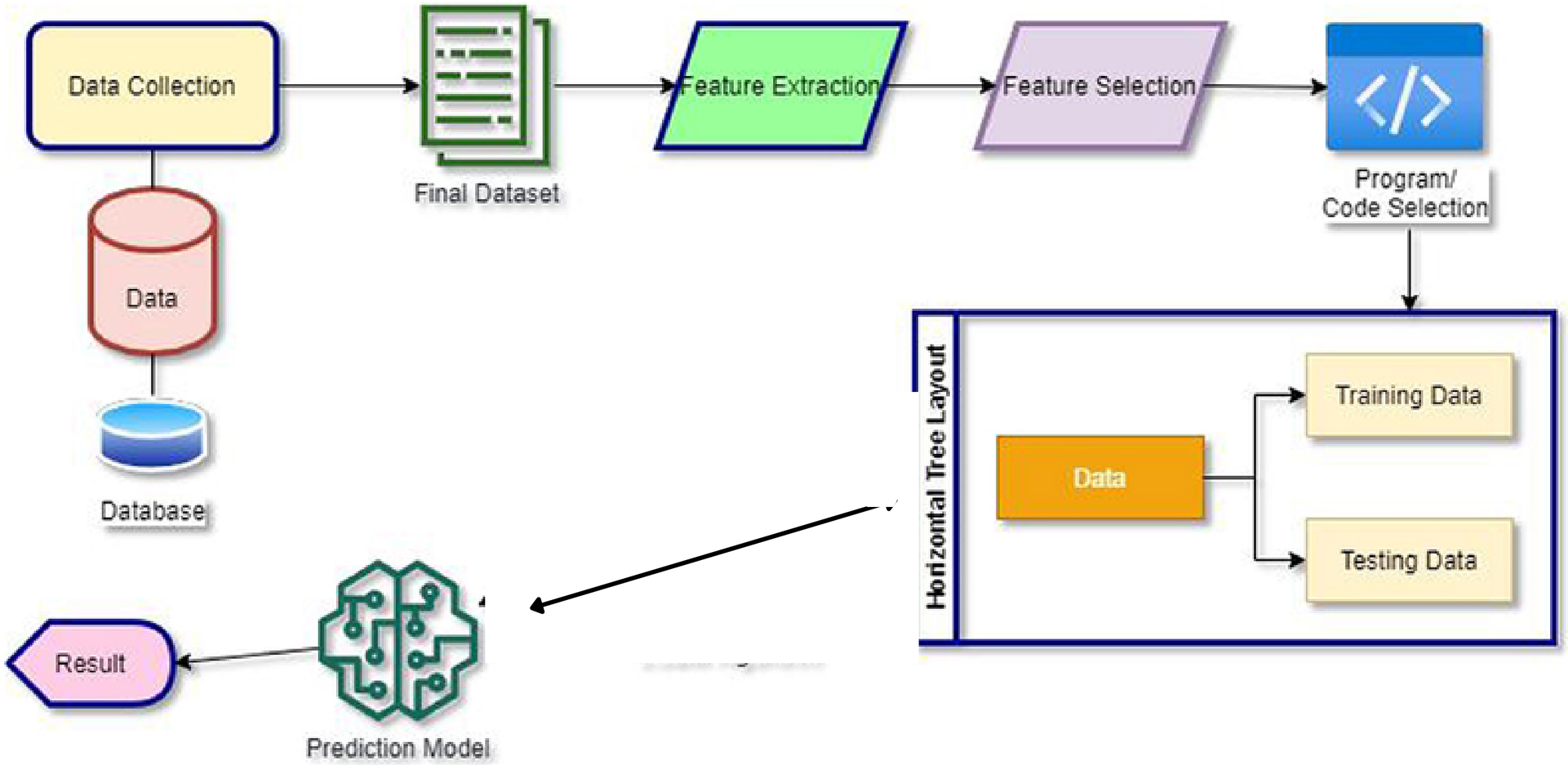
Here's how the RF algorithm can predict diabetes using the PIMA dataset:

1. **Data Preparation**
2. **Model Training**
3. **Model Evaluation**
4. **Making Predictions**

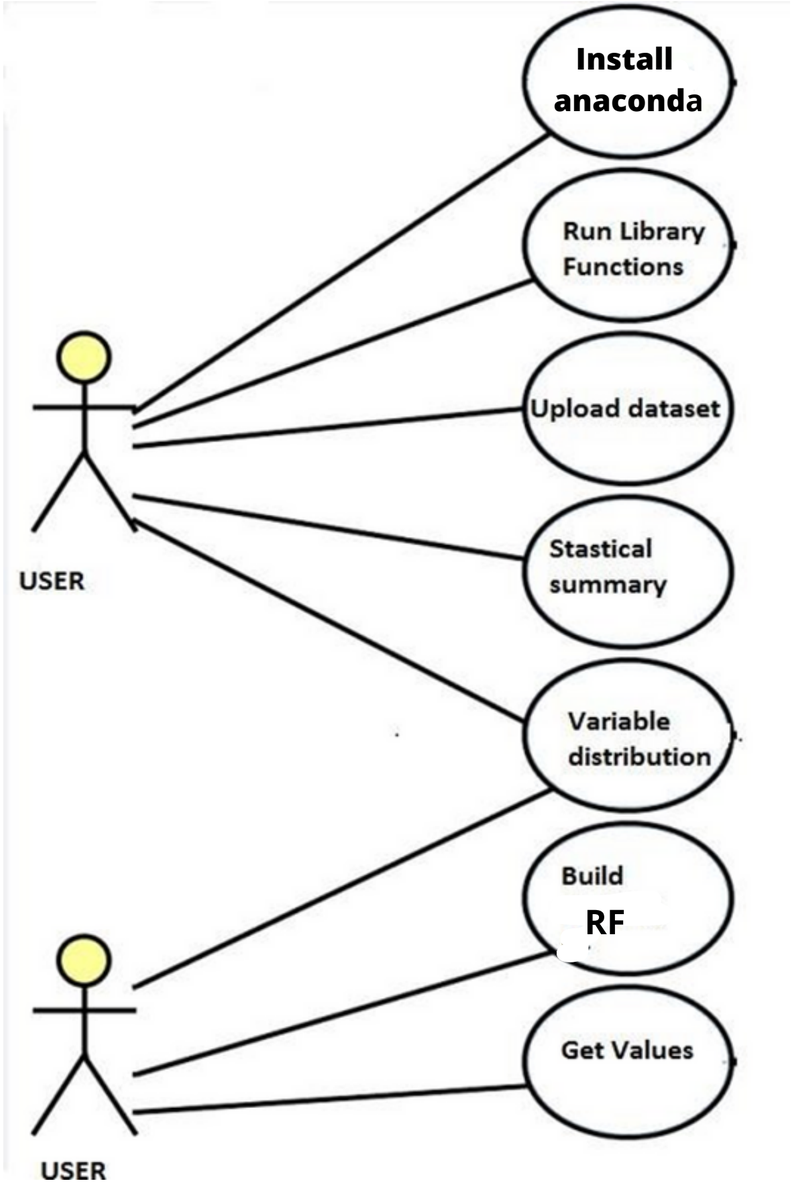
System design



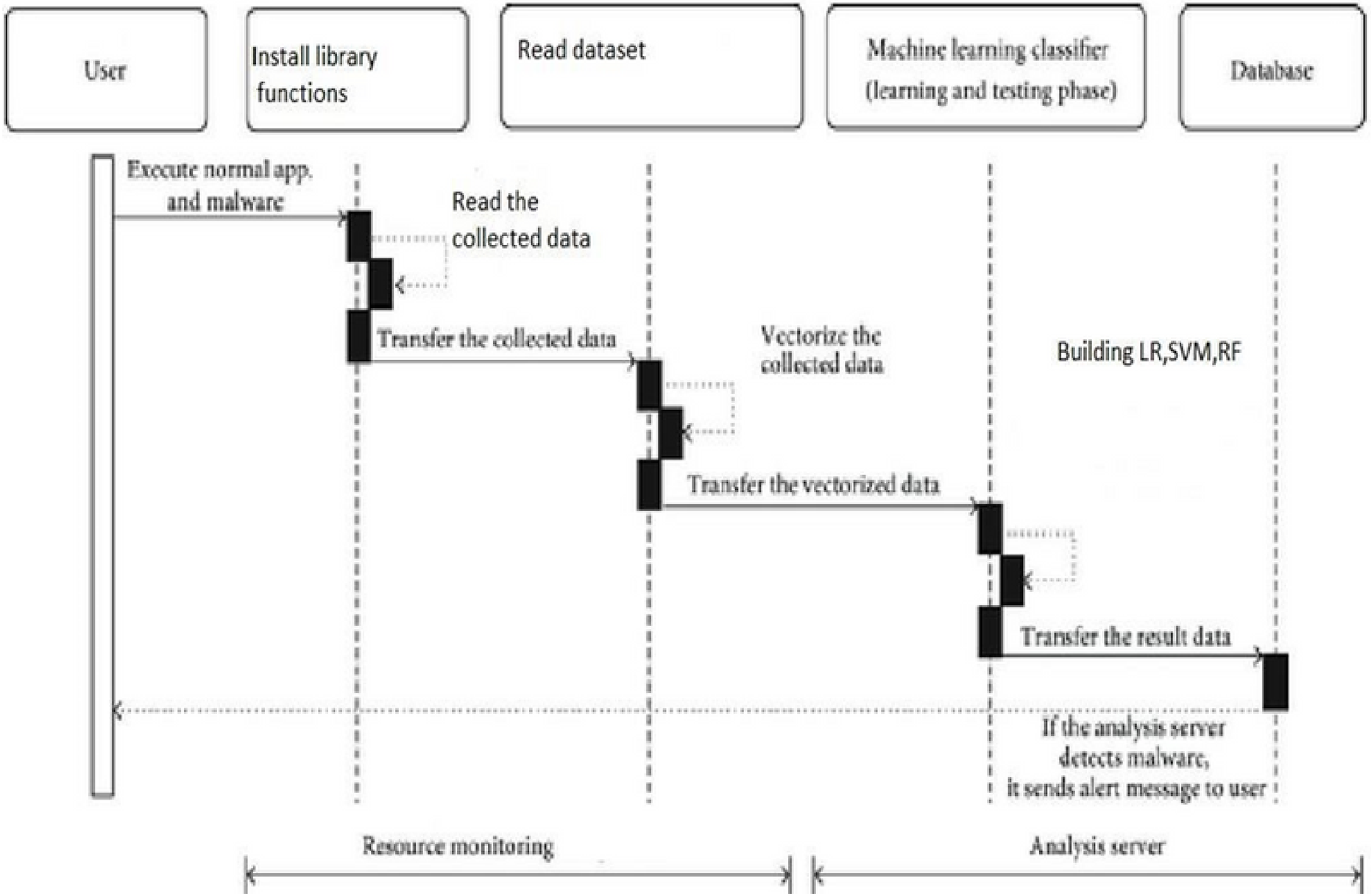
Architecture of the model



Process flowchart

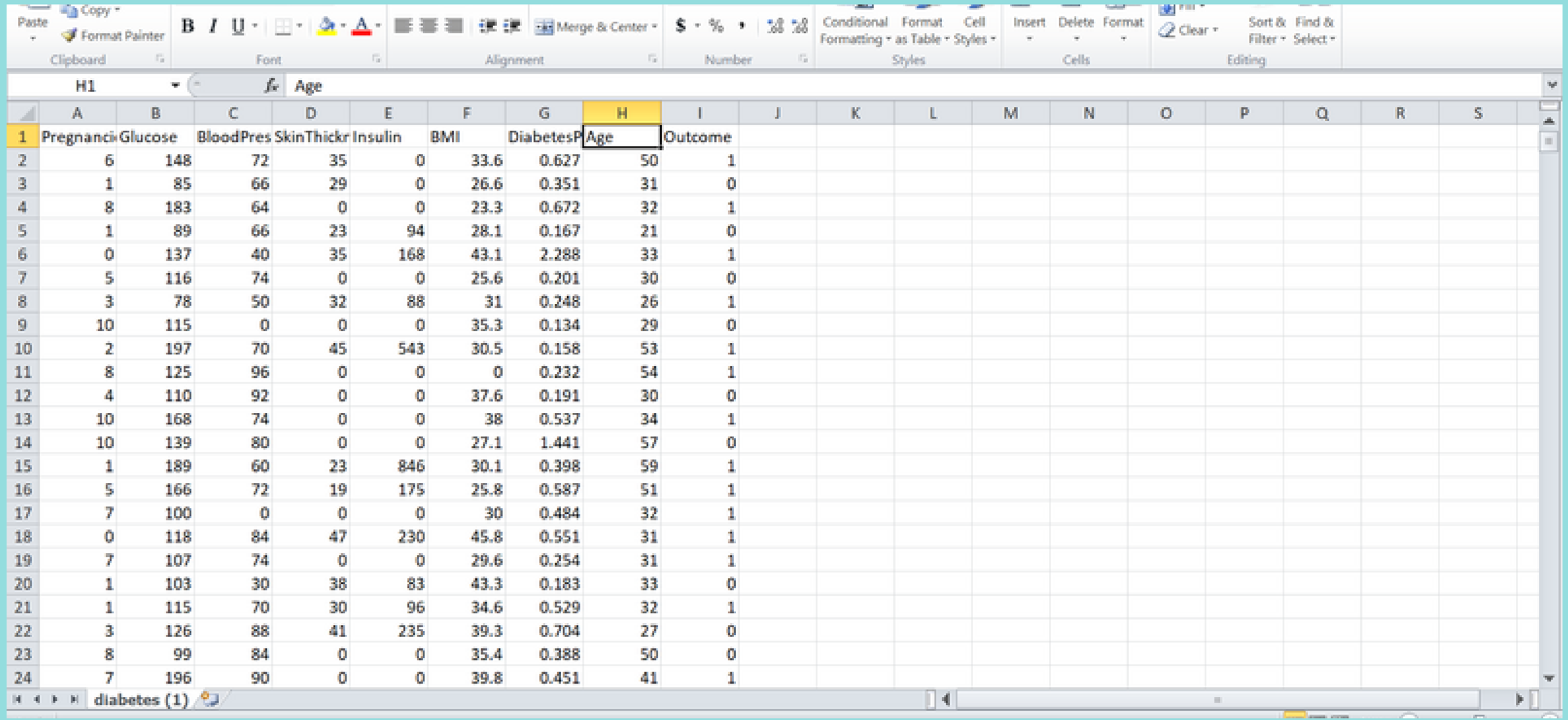


Use Case Diagram of diabetes system



Sequence Diagram of the diabetes system

Screenshot PIMA



H1 | Age

	A	B	C	D	E	F	G	H	I
1	Pregnanci	Glucose	BloodPres	SkinThickr	Insulin	BMI	DiabetesP	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
19	7	107	74	0	0	29.6	0.254	31	1
20	1	103	30	38	83	43.3	0.183	33	0
21	1	115	70	30	96	34.6	0.529	32	1
22	3	126	88	41	235	39.3	0.704	27	0
23	8	99	84	0	0	35.4	0.388	50	0
24	7	196	90	0	0	39.8	0.451	41	1

diabetes (1)

About dataset

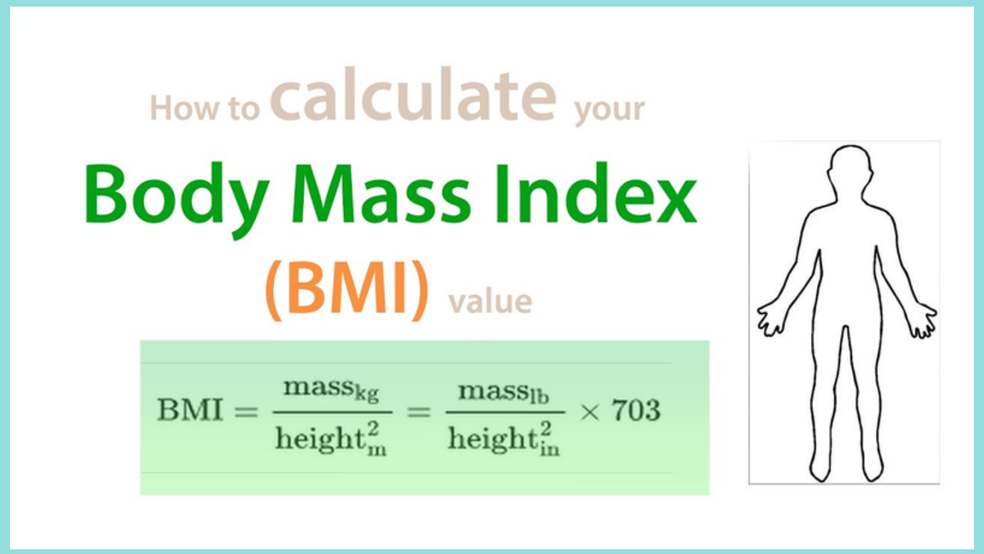
This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases.

The eight attributes used in the PIMA dataset are:

1. Number of times pregnant
2. glucose tolerance test
3. Blood pressure
4. Triceps skinfold thickness
5. insulin
6. Body mass index
7. Diabetes pedigree function
8. Age

The ranges of the attributes

Label	Feature	Mean	Min/Max
F_1	Number of times pregnant	3.8	0/17
F_2	Plasma glucose level estimated from glucose tolerance test	120.9	0/199
F_3	Diastolic blood pressure (mm Hg)	69.1	0/122
F_4	Triceps skinfold thickness (mm)	20.5	0/99
F_5	Two hour serum insulin (mu U/ml)	79.8	0/846
F_6	BMI (kg/m^2)	32	0/67.1
F_7	Function of diabetes nutrition	0.5	0.078/2.42
F_8	Age (years)	33.2	21/81



To calculate diabetes pedigree function

DPF = (0.1 x N1) + (0.01 x N2) + (0.001 x S)
where:

- N1: Number of affected individuals
- N2: Number of unaffected individuals
- S: The sum of the ages of onset of diabetes in the affected individuals

System requirements

Software requirements:

1. **Programming language**: Python
2. **ML frameworks**: Scikit-learn.
3. **Data manipulation and analysis tools**: Pandas, Numpy, seaborn and Matplotlib.
4. **Development environments**: Jupyter Notebook.
5. **OS**: Windows
6. **Front end** : tkinter

Hardware requirements

- Processor: Intel Core i5
- RAM: 4GB
- Hard Disk Space: 250 GB

Implementation with modules

Importing packages

```
#To avoid unnecessary warnings messages
import warnings
warnings.filterwarnings('ignore')

#Visualisation libraries
import seaborn as sns
import matplotlib.pyplot as plt

#To scale our data
from sklearn.preprocessing import StandardScaler

#To split the data into training and testing set
from sklearn.model_selection import train_test_split

#Classification model- Support Vector Machine
from sklearn.svm import SVC

#Classification model performance metrics
```

Implementation of LR , RandomForest and SVC to check the acccuray score

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

In [29]:
```
accuracies={}
lr=LogisticRegression()
lr.fit(x_train,y_train)
acc=lr.score(x_test,y_test)*100
accuracies['Logistic Regression']=acc
print("Test accuracy {:.2f}%".format(acc))
```

```
Test accuracy 77.33%
```

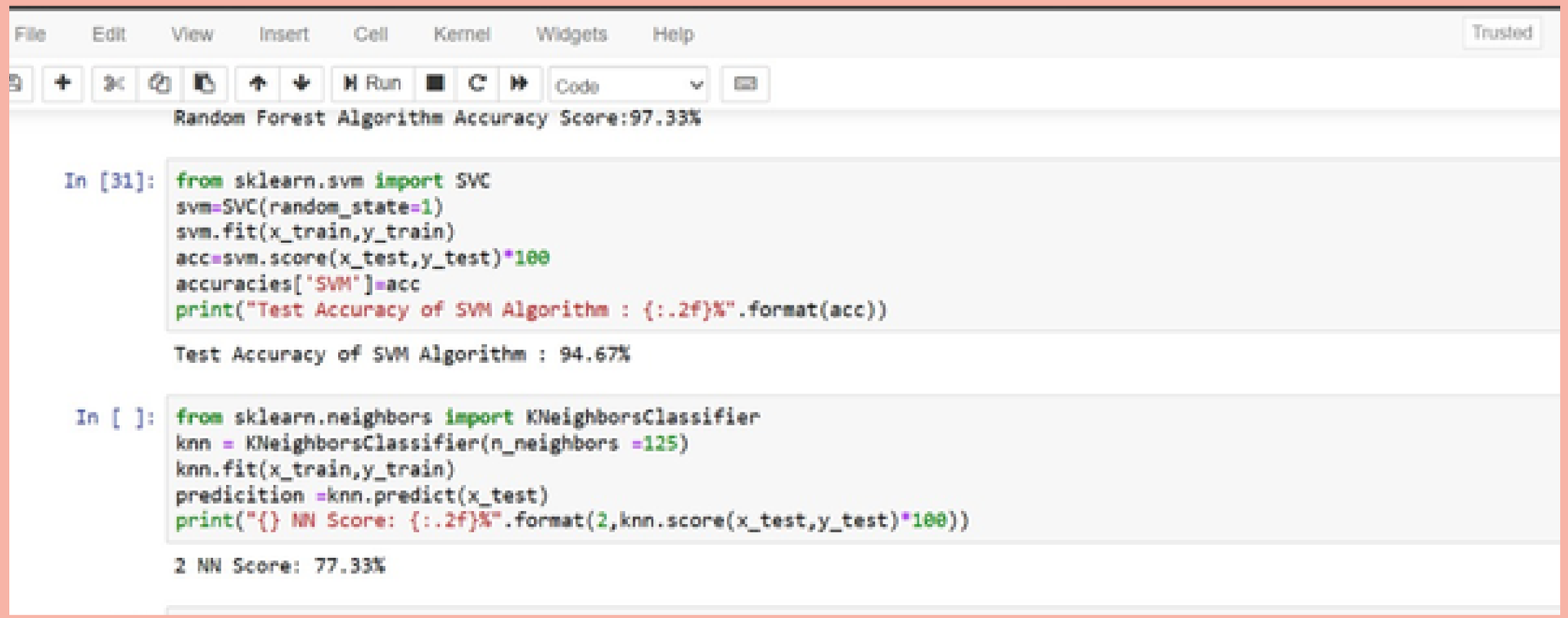
In [30]:
```
from sklearn.ensemble import RandomForestClassifier
rf= RandomForestClassifier(n_estimators=1500,random_state=12)
rf.fit(x_train,y_train)
acc=rf.score(x_test,y_test)*100
accuracies['Random Forest']=acc
print('Random Forest Algorithm Accuracy Score:{:.2f}%'.format(acc))
```

```
Random Forest Algorithm Accuracy Score:97.33%
```

In [31]:
```
from sklearn.svm import SVC
svm=SVC(random_state=1)
svm.fit(x_train,y_train)
acc=svm.score(x_test,y_test)*100
accuracies['SVM']=acc
print("Test Accuracy of SVM Algorithm : {:.2f}%".format(acc))
```

```
Test Accuracy of SVM Algorithm : 94.67%
```

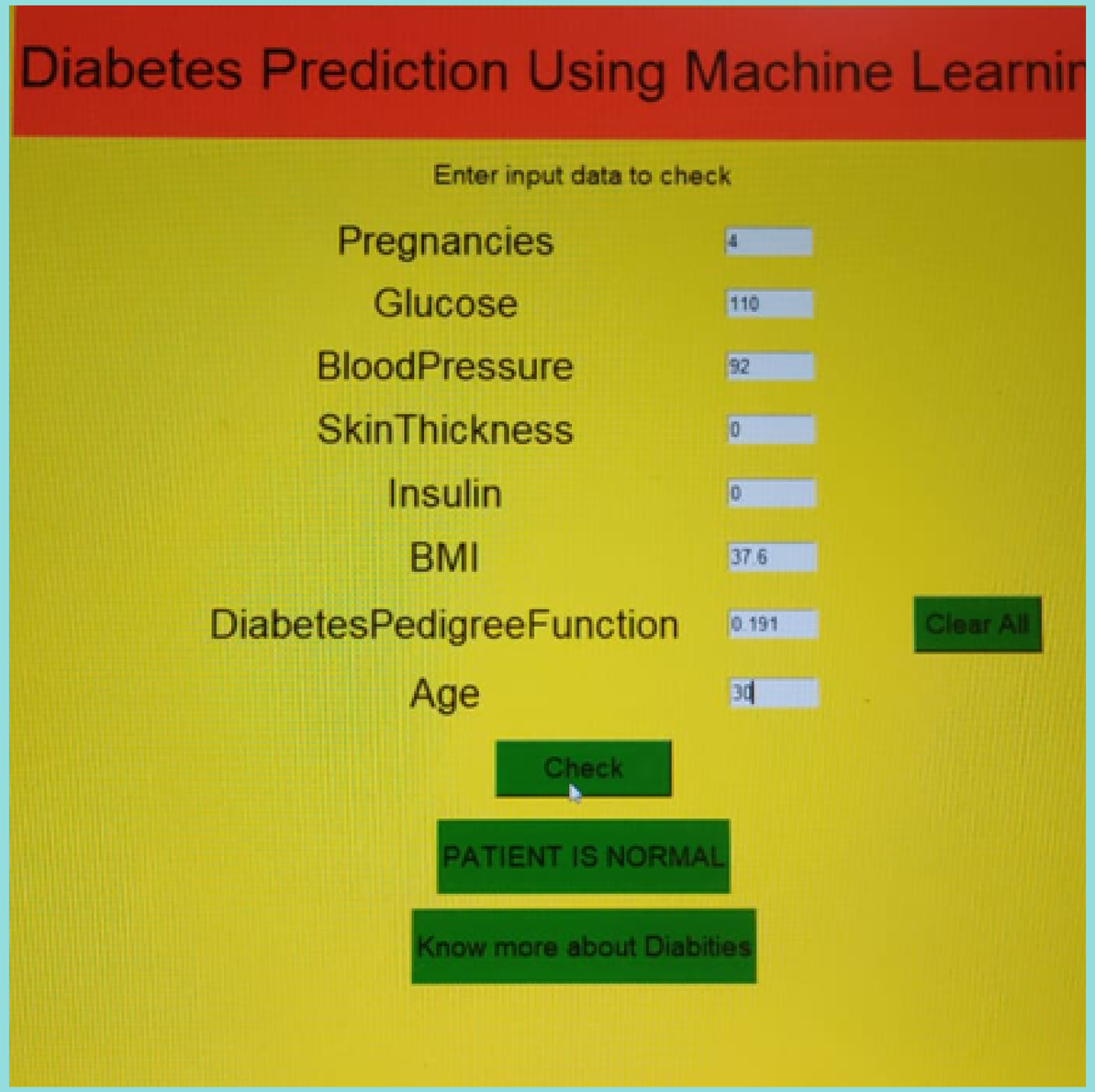
Implementation of knn model to check accuracy score



How Output should be seen in jupyter notebook?

```
In [20]: prediction([4,110,92,0,0,37.6,0.191,30])

         Patient is not Diabetic

Out[20]: array([0], dtype=int64)

In [21]: prediction([6,110,92,0,0,86.6,0.191,60])

         Patient is Diabetic

Out[21]: array([1], dtype=int64)
```

The user interface of the project



Result and discussion

As compared with SVC , LR, KNN and RF it is shown that RF algorithm has the highest accuracy rate of more than 90 % for the given dataset.

References

[1] F. ISLAM, R. FERDOUSI, S. RAHMAN, AND H. Y. BUSHRA, COMPUTER VISION AND MACHINE INTELLIGENCE IN MEDICAL IMAGE ANALYSIS. LONDON, U.K.: SPRINGER,2019.

[2] WORLD HEALTH ORGANIZATION (WHO). (2020). WHO REVEALS LEADING CAUSES OF DEATH AND DISABILITY WORLDWIDE: 2000–2019. ACCESSED: OCT. 22, 2021.[ONLINE]. AVAILABLE: HTTPS://WWW.WHO.INT/NEWS/ITEM/09-12-2020-WHO- REVEALS-LEADING-CAUSES-OF-DEATH-AND-DISABILITY-WORLDWIDE-2000-2019

[3] A. FRANK AND A. ASUNCION. (2010). UCI MACHINE LEARNING REPOSITORY.ACCESSED: OCT. 22, 2021. [ONLINE]. AVAILABLE: HTTP://ARCHIVE.ICS.UCI.EDU/ML

[4] G. PRADHAN, R. PRADHAN, AND B. KHANDELWAL, "A STUDY ON VARIOUS MACHINE LEARNING ALGORITHMS USED FOR PREDICTION OF DIABETES MELLITUS," IN SOFT COMPUTING TECHNIQUES AND APPLICATIONS (ADVANCES IN INTELLIGENT SYSTEMS AND COMPUTING), VOL. 1248. LONDON, U.K.: SPRINGER, 2021, PP. 553–561, DOI: 10.1007/978-981-15-7394-1_50.

[5] S. KUMARI, D. KUMAR, AND M. MITTAL, "AN ENSEMBLE APPROACH FOR CLASSIFICATION AND PREDICTION OF DIABETES MELLITUS USING SOFT VOTING CLASSIFIER," INT. J. COGN. COMPUT. ENG., VOL. 2, PP. 40–46, JUN. 2021, DOI: 10.1016/J.IJCCE.2021.01.001.

[6] M. A. SARWAR, N. KAMAL, W. HAMID, AND M. A. SHAH, "PREDICTION OF DIABETES USING MACHINE LEARNING ALGORITHMS IN HEALTHCARE," IN PROC.24TH INT. CONF. AUTOM. COMPUT. (ICAC), SEP. 2018, PP. 6–7, DOI: 10.23919/ICONAC.2018.8748992.

[7] S. K. DEY, A. HOSSAIN, AND M. M. RAHMAN, "IMPLEMENTATION OF A WEB APPLICATION TO PREDICT DIABETES DISEASE: AN APPROACH USING MACHINE LEARNING ALGORITHM," IN PROC. 21ST INT. CONF. COMPUT. INF. TECHNOL. (ICCIT),DEC. 2018, PP. 21–23, DOI: 10.1109/ICCITECHN.2018.8631968.

[8] A. MIR AND S. N. DHAGE, "DIABETES DISEASE PREDICTION USING MACHINE LEARNING ON BIG DATA OF HEALTHCARE," IN PROC. 4TH INT. CONF 10.1109/ICCUBEA.2018.8697439.

[9] S. SARU AND S. SUBASHREE. ANALYSIS AND PREDICTION OF DIABETES USING MACHINE LEARNING. ACCESSED: OCT. 22, 2022. [ONLINE]. AVAILABLE: HTTPS://PAPERS.SSRN.COM/SOL3/PAPERS.CFM?ABSTRACT_ID=3368308

[10] P. SONAR AND K. JAYAMALINI, "DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES," IN PROC. 3RD INT. CONF. COMPUT.METHODOLOGIES COMMUN. (ICCMC), MAR2019, PP. 367–371, DOI: 10.1109/ICCMC.2019.8819841.


Thank you