

WEB APPLICATION ON DIABETES PREDICTION

Submitted in partial fulfillment of the requirements
for the award of
Bachelor of Engineering degree in Computer Science and Engineering

by

Rimmalapudi Pavan Sai (17SCS0389)



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SCHOOL OF COMPUTING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**
Accredited with Grade "A" by NAAC

**JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI - 600 119**

MARCH – 2021



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600119

www.sathyabama.ac.in



DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **Rimmalapudi Pavan Sai (Roll No: 17SCS0389)** who carried out the project entitled **"Web Application on Diabetes Prediction"** under my supervision from November 2020 to March 2021.

Internal Guide

Dr. M. SELVI B.E., M.Tech.,

Head of the Department

**Dr. S. Vigneshwari M.E., Ph.D.,
Dr. L. Lakshmanan M.E., Ph.D.,**

Submitted for Viva voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

I **Rimmalapudi Pavan Sai (Roll No: 17SCS0389)** hereby declare that the Project Report entitled “**Web Application on Diabetes Prediction**” done by us under the guidance of **Dr. M. SELVI B.E., M.Tech** is submitted in partial fulfillment of the requirements for the award of Bachelor of Engineering degree in 2017-2021.

DATE:

PLACE:

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T.Sasikala M.E.,Ph.D., Dean**, School of Computing **Dr.S.Vigneshwari M.E., Ph.D.** and **Dr.L.Lakshmanan M.E., Ph.D.** , Heads of the Department of Computer Science and Engineering for providing us necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and a deep sense of gratitude to my Project Guide **Dr. M. SELVI B.E., M.Tech.**, for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

Diabetes is a disease caused due to lack or increase of sugar level in blood. It can also be caused when the insulin is not produced by pancreas. And also when body cells are not being affected by produced insulin. Diabetes is divided into 3 types namely Juvenile onset diabetes, Adult-onset and Gestational Diabetes. Diabetes should not be ignored since it may cause many diseases which can lead to death. It also affects our body very badly. It can also cause diseases related to kidney, heart, eye and so on. This paper we will do early prediction of Diabetes in a human body or a patient by applying various Machine Learning Techniques namely K-Nearest Neighbors, Logistic Regression, Decision Tree, Support Vector Machines, Naïve Bayes and Random Forest. The accuracy is calculated for every model. The most accurate model is considered and it is used to predict the diabetes. K-Nearest Neighbors has the highest accuracy compared to other machine learning techniques. So, it is used to predict the diabetes and it is deployed as an application which can be used using website and it is done using flask library of python. By using this web app, we can easily predict whether a person has diabetes or not by just giving the values in the website.

TABLE OF CONTENTS

Chapter No.	TITLE	Page No.
	ABSTRACT	v
	LIST OF FIGURES	viii
	LIST OF ABBREVIATIONS	ix
1	INTRODUCTION	1
	1.1. OVERVIEW	2
	1.2. MACHINE LEARNING	3
	1.3 MACHINE LEARNING METHODS	3
	1.3.1. SUPERVISED LEARNING	3
	1.3.2. UNSUPERVISED LEARNING	4
2	LITERATURE SURVEY	6
	2.1. RELATED WORK	6
3	METHODOLOGY	8
	3.1. EXISTING SYSTEM	8
	3.2. PROPOSED SYSTEM	8
	3.3. OBJECTIVE	8
	3.4. SOFTWARE AND HARDWARE REQUIREMENTS	9
	3.4.1. SOFTWARE REQUIREMENTS	9
	3.4.2. HARDWARE REQUIREMENTS	9
	3.4.3. LIBRARIES	9
	3.5. PROGRAMMING LANGUAGES	10
	3.5.1 PYTHON	10
	3.5.2. DOMAIN	10
	3.6. SYSTEM ARCHITECTURE	10
	3.7. ALGORITHMS USED	11
	3.7.1. NAIVE BAYES	11
	3.7.2. DECISION TREE	11

	3.7.3 LOGISTIC REGRESSION	12
	3.7.4 RANDOM FOREST	12
	3.7.5 SUPPORT VECTOR MACHINES	13
	3.7.6 K-NEAREST NEIGHBORS	14
	3.8 MODULES	15
	3.8.1 DATASET COLLECTION	15
	3.8.2 TRAIN THE ML MODELS	15
	3.8.3 DEPLOY ML MODELS AS A WEB APP	15
4	RESULTS AND DISCUSSION	16
	4.1 WORKING	16
5	CONCLUSION	17
	5.1 CONCLUSION	17
	REFERENCES	18
	APPENDICES	19
	A. SOURCE CODE	19
	B. SCREENSHOTS	23
	C. PLAGIARISM REPORT	26
	D. JOURNAL PAPER	27

LIST OF FIGURES

Figure No.	Figure Name	Page No.
1.1.	MACHINE LEARNING CLASSIFICATION	5
1.2.	MACHINE LEARNING TASK	5
3.1	SYSTEM ARCHITECTURE	10
3.2	DECISION TREE	12
3.3	RANDOM FOREST	13
3.4	SUPPORT VECTOR MACHINES	14
4.1	ACCURACIES	16

LIST OF ABBREVIATIONS

ABBREVIATIONS	EXPANSION
ML	Machine Learning
NB	Naive Bayes
DT	Decision Tree
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
RF	Random Forest

CHAPTER 1

INTRODUCTION

Diabetes is a disease caused due to lack or increase of sugar level in blood. It can also be caused when the insulin is not produced by pancreas. And also when body cells are not being affected by produced insulin.

Diabetes is divided into 3 namely, Juvenile-onset, Adult onset and Gestational Diabetes. It also affects our body very badly. It can also cause diseases related to kidney, heart, eye and so on.

Nowadays, diabetes has become the most recognized diseases in the world, not only because of its serious effects on the human body, but also because of its difficulty of diagnosing it.

Till 2019, 4630 lakh individuals are affected by adult-onset diabetes worldwide, with adult-onset diabetes. Male and Females are affected at the same rate. It also suggests that there will be a drastic rise in diabetes in the coming years. Diabetes also leads to a risk of early death.

Juvenile-onset diabetes have the symptoms like Thirst, Frequent urination, Bed-wetting in children who have no habit of doing it previously, Hunger, Weight Loss, Irritability, Fatigue, Blurred vision and so on.

Juvenile onset diabetes can cause either by genetics or more exposure to viruses. There are no preventive measures to reduce Juvenile-onset diabetes.

Adult-onset diabetes has the same signs as that of juvenile-onset diabetes but it additionally includes slow healing of wounds, frequent infections, weight gain or loss and so on. Treatment consists of doing physical exercise, quit smoking, following diabetic diet and so on.

Make sure that diabetes will not enter your body. Once it enters, it will be so difficult to get rid of it. Take care of yourself. Nowadays, even children are getting diabetes easily. So, follow a good diet and maintain regularity in taking food, doing physical exercise and so on.

1.1 OVERVIEW

A disease is a condition that affects the individual functioning of body totally. Diseases if neglected will lead to the death of an individual. Diseases can be identified by the symptoms of the body of an individual. Health is the most important in every human's life. Weekly or monthly check up of one's health is most important for the prevention and also to stay healthy.

Nowadays, the individual is not having that much time to go for health check-up. Recently, due to covid-19, no one are willing to go to hospital for health check-up due to the fear of spreading virus. As we all know the saying which tells that "Prevention of the disease at an early stage is much better than the cure which we take after we get affected by the disease".

Healthcare is the most crucial parts of the human life. Nowadays, so many are not willing to go to hospital, due to work overload and negligence of their health. The doctors and nurses are putting up maximum efforts to save people's lives without even considering their own loves. There are also some villages which lack medical facilities.

This project totally focuses on prediction of diabetes. We will use the Pima Indian Diabetes Dataset, which contains 768 rows and 9 columns. The attributes of dataset include glucose level, level of insulin, etc. First we will find the correlation of all the features. Then, we will select highest correlation features and use the same for prediction of diabetes.

The Machine Learning techniques we will apply are K-Nearest Neighbors, Logistic Regression, Decision Tree, Support Vector Machine, Naive Bayes and Random Forest. The accuracy is calculated for every model.

The most accurate model is considered and it is used to predict the diabetes. K-Nearest Neighbors has the highest accuracy compared to other machine learning techniques. So, it is used to predict the diabetes and it is deployed it as an application which can be used using website and it is done using flask library of python. By using this web app, we can easily predict whether a person has diabetes or not by just giving the values in the website.

1.2 MACHINE LEARNING

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.

Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

1.3 Machine Learning Methods

In machine learning, tasks are generally classified into broad categories. These categories are based on how learning is received or how feedback on the learning is given to the system developed.

Two of the most widely adopted machine learning methods are **supervised learning** which trains algorithms based on example input and output data that is labelled by humans, and **unsupervised learning** which provides the algorithm with no labelled data in order to allow it to find structure within its input data.

1.3.1 Supervised Learning

In supervised learning, the computer is provided with example inputs that are labeled with their desired outputs. The purpose of this method is for the algorithm to be able to “learn” by comparing its actual output with the “taught” outputs to find errors, and modify the model accordingly. Supervised learning therefore uses patterns to predict label values on additional unlabelled data.

For example, with supervised learning, an algorithm may be fed data with images of sharks labelled as fish and images of oceans labelled as water. By being

trained on this data, the supervised learning algorithm should be able to later identify unlabelled shark images as fish and unlabelled ocean images as water.

A common use case of supervised learning is to use historical data to predict statistically likely future events. It may use historical stock market information to anticipate upcoming fluctuations, or be employed to filter out spam emails. In supervised learning, tagged photos of dogs can be used as input data to classify untagged photos of dogs.

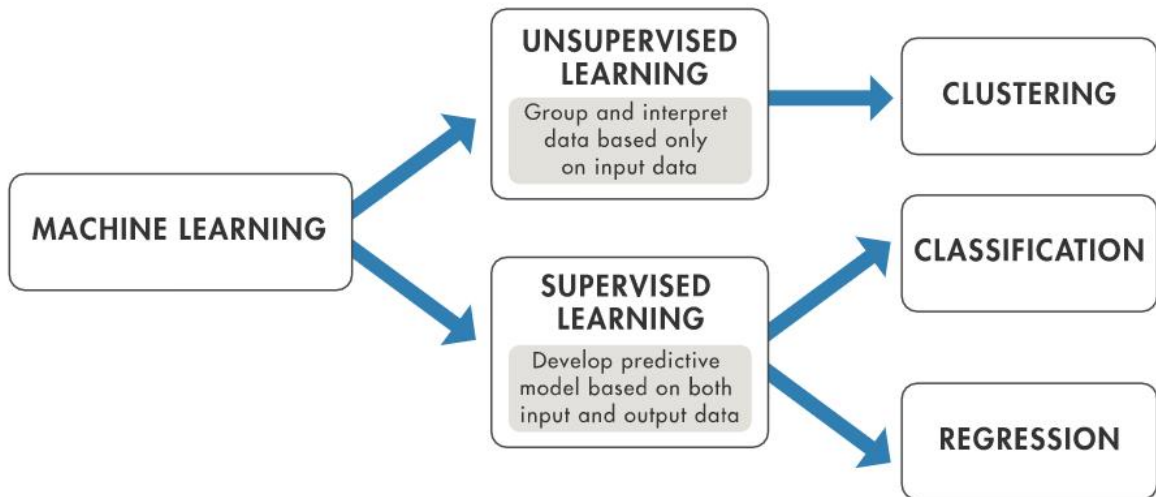
1.3.2 Unsupervised Learning

In unsupervised learning, data is unlabelled, so the learning algorithm is left to find commonalities among its input data. As unlabelled data are more abundant than labelled data, machine learning methods that facilitate unsupervised learning are particularly valuable.

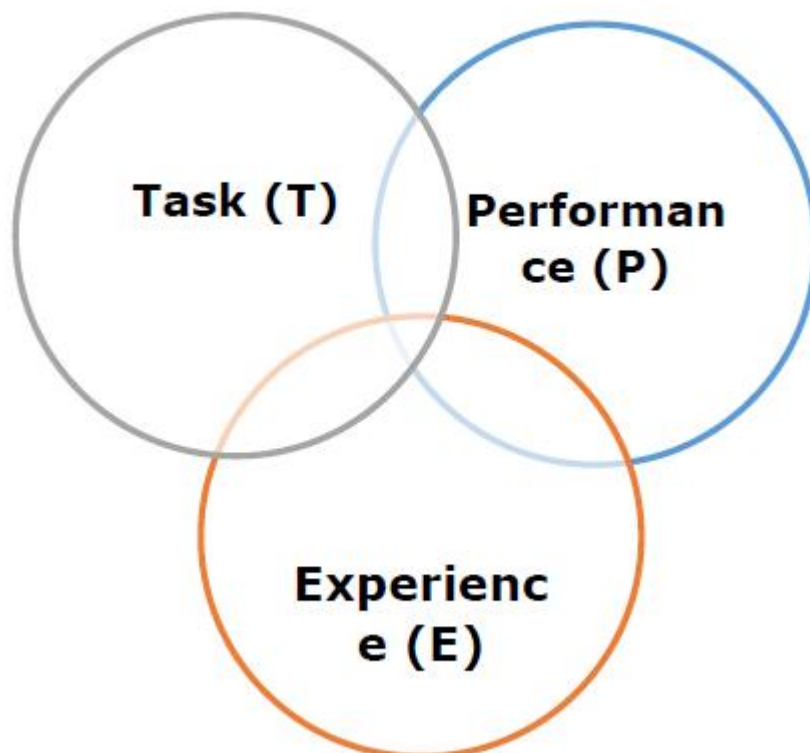
The goal of unsupervised learning may be as straightforward as discovering hidden patterns within a dataset, but it may also have a goal of feature learning, which allows the computational machine to automatically discover the representations that are needed to classify raw data.

Unsupervised learning is commonly used for transactional data. You may have a large dataset of customers and their purchases, but as a human you will likely not be able to make sense of what similar attributes can be drawn from customer profiles and their types of purchases.

With this data fed into an unsupervised learning algorithm, it may be determined that women of a certain age range who buy unscented soaps are likely to be pregnant, and therefore a marketing campaign related to pregnancy and baby products can be targeted to this audience in order to increase their number of purchases.



1.1 MACHINE LEARNING CLASSIFICATION



1.2 MACHINE LEARNING TASK

CHAPTER 2

LITERATURE SURVEY

2.1 RELATED WORK

- Olga Tsave et al. in [1] used many machine learning models, 85% of supervised machine learning algorithms and 15% of unsupervised machine learning algorithms. Of all those, SVM has the highest accuracy. By using that, useful knowledge is extracted from datasets.
- Deepti Sisodia et al. in [2] used Pima Indians Diabetes Dataset and used Decision Tree, Naive Bayes and SVM models are implemented. Accuracies of models using various measures like recall, precision and so on. Naive Bayes has the highest accuracy of all the three and it is used to predict whether the individual suffer from diabetes or not.
- D. Dutta et al. in [3] used machine learning to find the most important features which causes diabetes and predict whether the person will get diabetes or not in future.
- N. Yuvaraj et al. in [4] used machine learning to increase the automation of health care systems and give accurate disease prediction. Pima Indians Diabetes dataset is used for evaluation of working of an algorithm.
- Akhtar Jamil et al. in [5] compared machine learning algorithms with deep learning algorithms. CNN is used for deep learning algorithm. Random forest and SVM are used for machine learning algorithms. Of all these, Random Forest has the highest accuracy. So, it is used for diabetes prediction.
- Aishwarya Mujumdar et al. in [6] predicted diabetes using external factors and also regular factors like BMI, Insulin, Glucose and so on. Pipeline model which is developed gives more accurate prediction.

- Mitesh Warke et al. in [7] done Diabetes Diagnosis using Machine Learning Algorithms. There is a static prediction engine that serves prediction results for one disease.
- Maniruzzaman, M., Rahman, et al. in [8] implemented four classifiers namely LR, DT, RF, NB and the accuracy of models is calculated. Finally, they found that combination of logistic regression and random forest gives the most accurate prediction.
- M. K. Hasan et al. in [9] implemented various ML models, Adaboost, Decision tree, XGBoost, K-Nearest Neighbors, Multi-layer perceptron and so on Pima Indians Dataset. Accurate predictions are given.
- Alexi Delgado et al. in [10] designed an application which can be used through website. ML is used as a main domain. Decision Tree classifier to predict the diabetes.
- Shruti Garg et al. in [11] used six Machine Learning Classifiers namely LR, SVM, DT, RF, NB, Gradient Boosting after which accuracies are also calculated. In this way, they had predicted whether an individual suffers from type 2 diabetes or not.
- Ahan Chatterjee et al. in [12], they have used machine learning models like Decision tree, SVM, Random Forest, and so on, find the best classifier using the accuracy and use it to predict cancer disease risk in the early stage and prevent it. Simulation model is also designed to manage the patient flow in OPDs.

CHAPTER 3

METHODOLOGY

3.1 EXISTING SYSTEM

From the literature survey, we have found that existing systems considers all the variables except the outcome as independent variables and predict the diabetes disease using them.

And also they have used only logistic regression to predict the diabetes. They have restricted themselves till the prediction of diabetes using Jupyter notebook and not deployed it.

3.2 PROPOSED SYSTEM

The proposed system takes the variables having high correlation (i.e., Glucose, Age, BMI, Insulin) with outcome as independent variables and predict the diabetes using them.

Implementing various models and finding the most accurate one i.e., K-Nearest Neighbors and used it for prediction. At last, we will deploy the model as web application using flask web framework. By this, an individual can find easily whether they are suffering from diabetes or not.

3.3 OBJECTIVE OF PROJECT

- Machine learning is the method used in this project, which allows us, through its classifiers, to predict upcoming events, in this case, probability of disease.
- Various machine learning algorithms are used and the most accurate of them is found i.e., k-nearest algorithm. It is used as a classifier to predict whether the patient has diabetes or not.

The objective of this project is :

- ✓ To create a machine learning model for predicting whether a person suffers from diabetes or not.
- ✓ Deploy Machine Learning model as a web application (using Flask), which is used for the best care of patients and predicts diabetes, thus helping people who need a consultation.

3.4 SOFTWARE AND HARDWARE REQUIREMENTS

3.4.1 Software Requirements:

- ✓ Python
- ✓ Anaconda
- ✓ Jupyter Notebook

3.4.2 Hardware Requirements:

- ✓ Processor: Intel Core i5
- ✓ RAM: 8GB
- ✓ OS: Windows

3.4.3 Libraries:

The libraries required are

- ✓ Flask- It is used for deploying machine learning model as web application
- ✓ Pandas- Used to read the dataset
- ✓ Matplotlib- Used for data visualisation
- ✓ Sklearn- Used for importing models, get accuracy and confusion matrix.

3.5. PROGRAMMING LANGUAGES

3.5.1 PYTHON

- ✓ Python is the best programming language suited for Machine Learning. According to studies and surveys, Python is the fifth most important language as well as the most popular language for machine learning and data science. It is because of the following strengths that Python has –
- ✓ **Easy to learn and understand** – The syntax of Python is simpler; hence it is relatively easy, even for beginners also, to learn and understand the language.
- ✓ **Multi-purpose language** – Python is a multi-purpose programming language because it supports structured programming, object-oriented programming as well as functional programming.

- ✓ **Support of open source community** – As being open source programming language, Python is supported by a very large developer community. Due to this, the bugs are easily fixed by the Python community. This characteristic makes Python very robust and adaptive.

3.5.2 DOMAIN

Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people. Although machine learning is a field within computer science, it differs from traditional computational approaches. In traditional computing, algorithms are sets of explicitly programmed instructions used by computers to calculate or problem solve. Machine learning algorithms instead allow for computers to train on data inputs and use statistical analysis in order to output values that fall within a specific range. Because of this, machine learning facilitates computers in building models from sample data in order to automate decision-making processes based on data inputs.

3.6. SYSTEM ARCHITECTURE

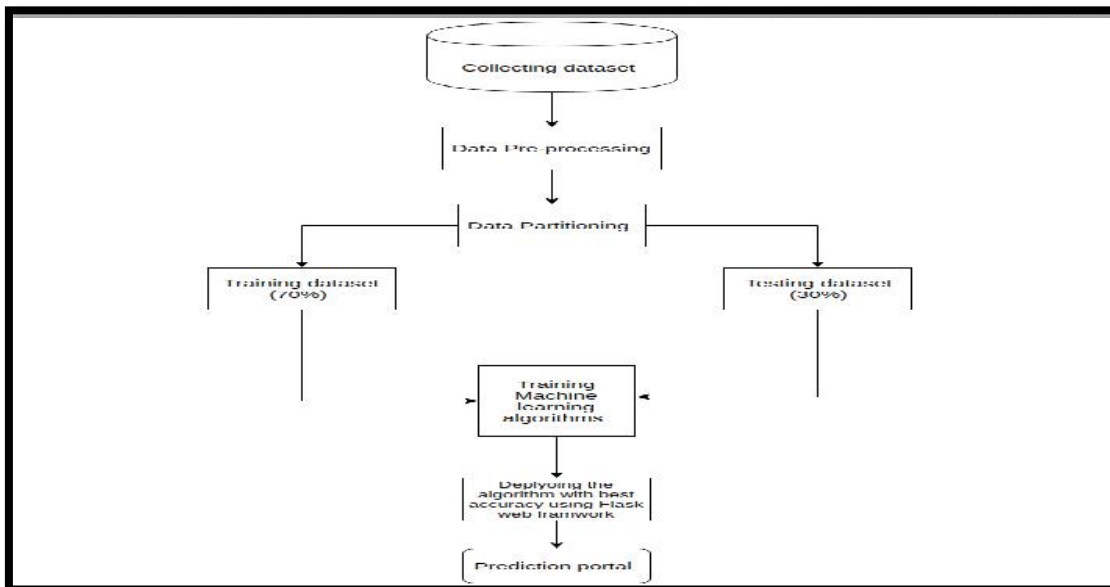


Fig 3.1 System Architecture

3.7 ALGORITHMS USED

3.7.1 NAIVE BAYES

It is a machine learning algorithm for classification problems which is based on Bayes theorem. The primary use of this is to do text classification. The Bayes theorem can be defined as:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

$P(C|X)$ is the probability of hypothesis C for the given data X . This is called the posterior probability.

$P(X|C)$ is the probability of data X given that the hypothesis C was true.

$P(C)$ is the probability of hypothesis C being true. This is called the prior probability of C .

$P(X)$ is the probability of the data and evidence of data and is called marginal probability.

3.7.2 DECISION TREE

A **decision tree** is a type of supervised learning algorithm (with a predefined target variable). While mostly used in classification tasks, it can handle numeric data as well.

This algorithm splits a data sample into two or more homogeneous sets based on the most significant differentiator in input variables to make a prediction.

With each split, a part of a tree is being generated. As a result, a tree with decision nodes and leaf nodes (which are decisions or classifications) is developed. A tree starts from a root node – the best predictor.

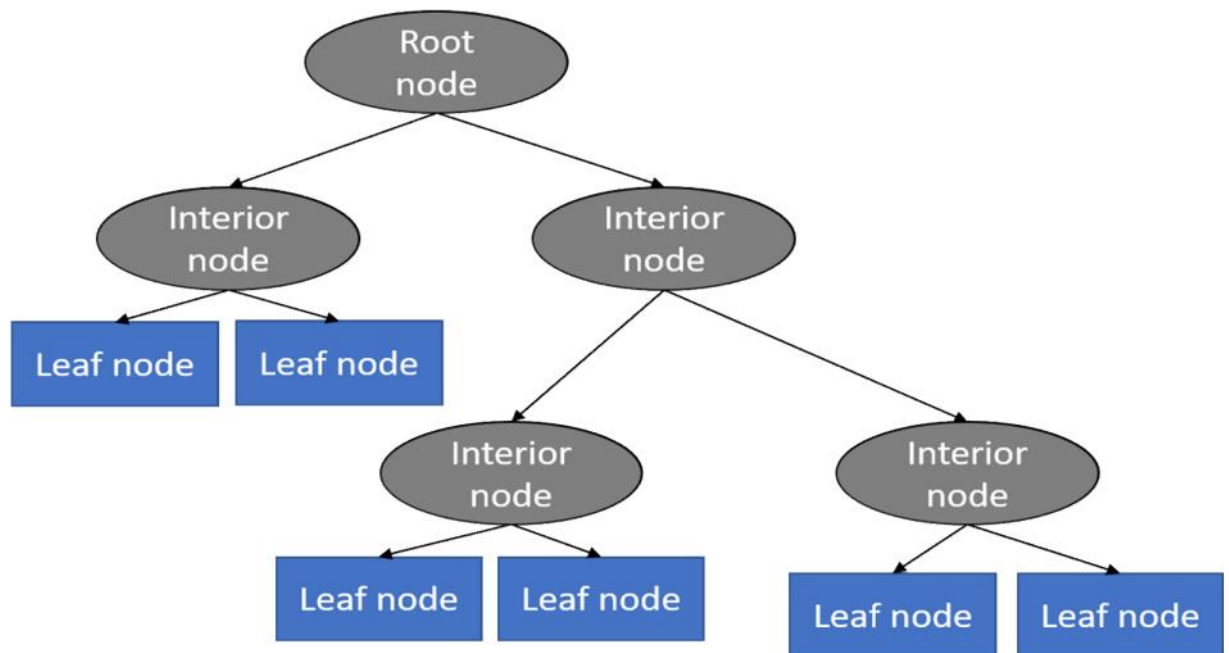


Fig 3.2. Decision Tree

3.7.3 Logistic Regression

Logistic Regression is an algorithm used for binary classification problems. It predicts the likelihood of an event by measuring the relationship between a dependent variable and one or more independent variables (features). More specifically, logistic regression will predict the possibility of an instance (data point) belonging to the default category.

The Sigmoid Function

The sigmoid function/logistic function is a function that resembles an “S” shaped curve when plotted on a graph. It takes values between 0 and 1 and “squishes” them towards the margins at the top and bottom, labelling them as 0 or 1.

The equation for the Sigmoid function is this:

$$y = 1 / (1 + e^{-x})$$

3.7.4 Random Forest

A **Random forest** is a type of an ensemble learning method that uses numerous decision trees to achieve higher prediction accuracy and model stability. This method deals with both regression and classification tasks. Every tree classifies a data

instance (or votes for its class) based on attributes, and the forest chooses the classification that received the most votes. In the case of regression tasks, the average of different trees decisions is taken.

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the "bagging" method. The general idea of the bagging method is that a combination of learning models increases the overall result.

One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems.

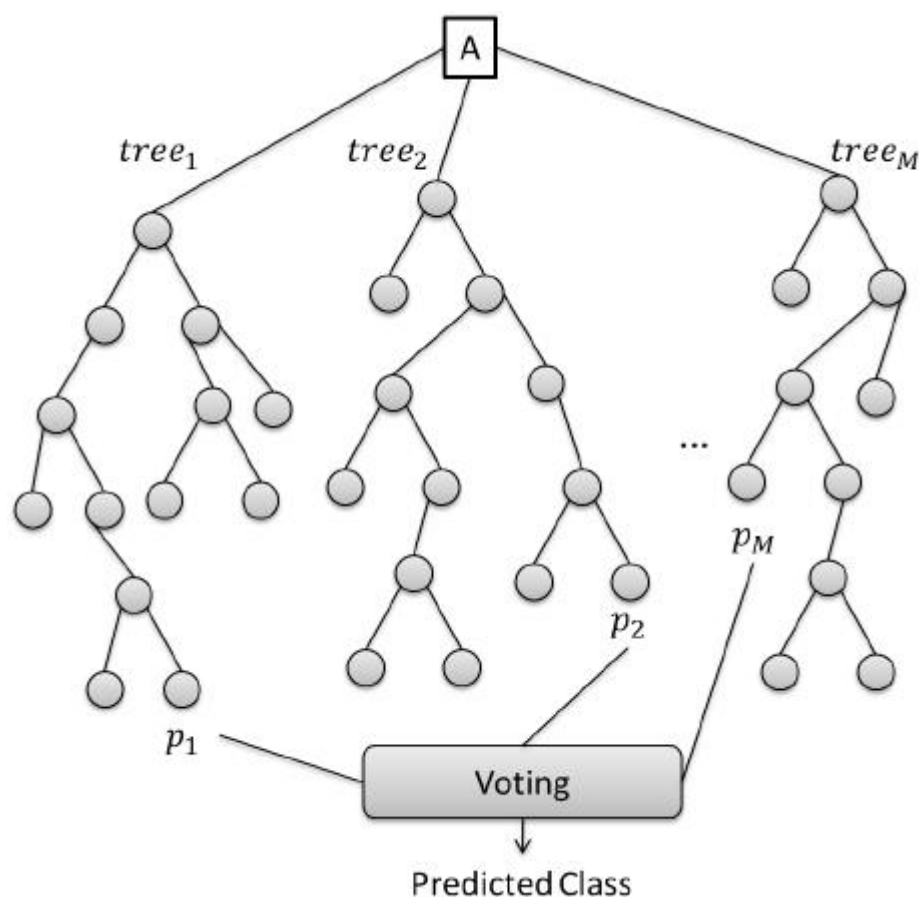


Fig 3.3. Random Forest

3.7.5. SUPPORT VECTOR MACHINES

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labelled training data for each category, they're able to categorize new

text. Support Vector Machines (SVM) is a fast and dependable classification algorithm that performs very well with a limited amount of data. Separation of classes. That's what SVM does. It finds out a line/ hyper-plane (in multidimensional space that separate outs classes).

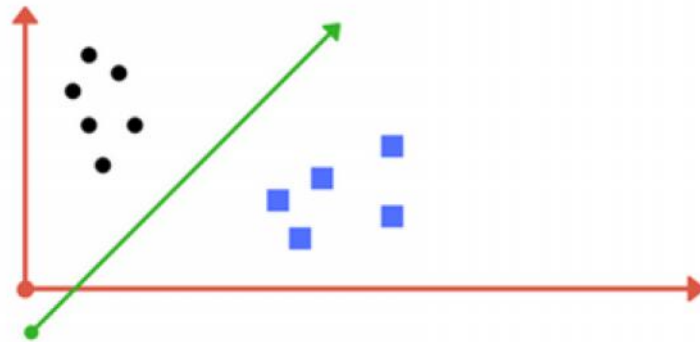


Fig. 3.4 SUPPORT VECTOR MACHINES

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labelled training data (*supervised learning*) the algorithm outputs an optimal hyperplane which categorizes new examples. In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.

3.7.6. K-NEAREST NEIGHBORS

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry. The following two properties would define KNN well –

-)] **Lazy learning algorithm** – KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.
-)] **Non-parametric learning algorithm** – KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

3.8 MODULES

Web Application on Diabetes Prediction have the following modules:

Dataset Collection

- ✓ Collect the dataset, clean it, split into train and test dataset. This is done in Jupyter Notebook

Training the machine learning algorithms

- ✓ Train the dataset using various Machine learning algorithms – the most accurate algorithm is k-nearest neighbors and use it for prediction of diabetes. This is done in Jupyter Notebook.

Deploying Machine learning model as web app

- ✓ Deploy the machine learning model as web app using flask web framework and prediction portal will appear.

The steps followed in achieving this project objective are given below:

1. Collect the dataset from kaggle notebooks.
2. Data Pre-processing like checking for missing values, out-of-range values and so on is done.
3. Data Partitioning i.e., Splitting the dataset into train and test dataset.
4. Train the dataset using Machine learning algorithms and evaluate the accuracies of all the models.
5. Deploy the most accurate machine learning model i.e., K-Nearest Neighbors as a web app using Flask web framework.
6. Now, the prediction portal appears in which we can give the values and predict whether the person has diabetes or not.

CHAPTER 4

RESULTS AND DISCUSSION

Doctors and medical professionals are always required in case of an emergency. In the current situation of COVID-19, where essential resources are unavailable and people are also not willing to go outside in fear of spreading virus, our prediction system will be very helpful for finding the diabetes based on the attributes in the early stage and get the correct diagnosis of a disease. This also helps in reduction of the cost and give the correct and fast result.

4.1. WORKING

In the current situation of COVID-19, where sufficient resources are unavailable, our prediction portal can help in prediction of diabetes easily. This also helps in reduction of the cost required.

Accuracy of different models arranged in descending order is shown in the table. Of these models K-Nearest Neighbors has the most highest accuracy, so it is used to predict. Accuracy is defined as the ratio of sum of TP and TN to the sum of TP, TN, FP, FN.

Model	Accuracy
K-Nearest Neighbors	78%
Random Forest	75%
Support Vector Machines	73%
Logistic Regression	72%
Naive Bayes	71%
Decision Tree	68%

Fig 4.1 Accuracies

CHAPTER 5

CONCLUSION

5.1. CONCLUSION

In the current situation of COVID-19, where sufficient facilities and resources are unavailable, our prediction system can prove to be helpful and can be used in the diagnosis of a disease.

Diabetes is a disease caused due to lack or increase of sugar level in blood. It can also be caused when the insulin is not produced by pancreas. And also when body cells are not being affected by produced insulin.

Real-time analysis of the PIMA Indian Dataset is done. The k-nearest neighbors gave the highest accurate prediction than the other models. Flask API has been used to deploy trained model with the HTML page.

The HTML page has been displayed in such a way that the user only needs to input the values of mass index of the body, BP, Level of glucose, Insulin level and so on. The HTML page provides the result whether the user has diabetes or not on the webpage indicating positive if the person has diabetes and negative if the person does not have diabetes.

Doctors and medical professionals are always required in case of an emergency. In the current situation of COVID-19, where sufficient facilities and resources are unavailable, our prediction system can prove to be helpful and can be used in the diagnosis of a disease. This project would help in lowering the cost required in dealing with the disease and would also make the recovery process easy.

REFERENCES

1. Olga Tsave et al. "Machine Learning and Data Mining Methods in Diabetes Research", Journal ,Volume 15, 2017
2. Deepti Sisodia et al. "Prediction of Diabetes using Classification Algorithms", Volume 132, 2018
3. D. Dutta, D. Paul and P. Ghosh, "Analysing Feature Importances for Diabetes Prediction using Machine Learning," 2018 IEEE 9th Annual Information Technology, pp.924-928, 10.1109/IEMCON.2018.8614871.
4. N. Yuvaraj et al. "Diabetes prediction in healthcare systems using machine learning on hadoop cluster", 16 January 2019.
5. Akhtar Jamil et al. "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques", 2019, pp. 1-4, doi: 10.1109/UBMYK48245.2019.8965556.
6. Aishwarya Mujumdar et al. "Diabetes Prediction using Machine Learning Algorithms", Volume 165, 2019.
7. Mitesh Warke et al. "Diabetes Diagnosis using Machine Learning Algorithms", Volume: 06 Issue: 03 Mar 2019
8. Maniruzzaman, M., Rahman, M.J., Ahammed, B. et al. "Classification and prediction of diabetes disease using machine learning" Syst 8, 7 (2020), 3 January 2020.
9. M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling Of Different Machine Learning Classifiers," in IEEE Access, vol. 8, pp. 76516-76531, 2020.
10. Alexi Delgado et al. "Design of a Web Application for the Detection of Diabetes using Machine Learning", Volume 8. No. 7, July 2020.
11. Shruti Garg et al. "Prediction of Type 2 Diabetes using Machine Learning Classification Methods" 167 (2020) 706–716, Volume 167, 2020.
12. Ahan Chatterjee et al. "A Machine Learning Approach to prevent cancer", DOI: 10.4018/978-1-7998-2742-9.ch007, 2021. 167 (2020) 706–716, Volume 167, 2020.

APPENDICES

A) SOURCE CODE

MODULES IMPLEMENTATION

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import warnings

warnings.filterwarnings('ignore')

dataset = pd.read_csv('diabetes.csv')

from sklearn.preprocessing import MinMaxScaler

sc = MinMaxScaler(feature_range = (0, 1))

dataset_scaled = sc.fit_transform(dataset_new)

dataset_scaled = pd.DataFrame(dataset_scaled)

# Selecting features - [Glucose, Insulin, BMI, Age]

X = dataset_scaled.iloc[:, [1, 4, 5, 7]].values

Y = dataset_scaled.iloc[:, 8].values

# Splitting X and Y

from sklearn.model_selection import train_test_split

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.20,
random_state = 42, stratify = dataset_new['Outcome'] )

# Logistic Regression Algorithm

from sklearn.linear_model import LogisticRegression

logreg = LogisticRegression(random_state = 42)

logreg.fit(X_train, Y_train)

# Plotting a graph for n_neighbors
```

```

from sklearn import metrics

from sklearn.neighbors import KNeighborsClassifier

X_axis = list(range(1, 31))

acc = pd.Series()

x = range(1,31)

for i in list(range(1, 31)):

    knn_model = KNeighborsClassifier(n_neighbors = i)

    knn_model.fit(X_train, Y_train)

    prediction = knn_model.predict(X_test)

    acc = acc.append(pd.Series(metrics.accuracy_score(prediction, Y_test)))

plt.plot(X_axis, acc)

plt.xticks(x)

plt.title("Finding best value for n_estimators")

plt.xlabel("n_estimators")

plt.ylabel("Accuracy")

plt.grid()

plt.show()

print('Highest value: ',acc.values.max())

# K nearest neighbors Algorithm

from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors = 24, metric = 'minkowski', p = 2)

knn.fit(X_train, Y_train)

# Support Vector Classifier Algorithm

from sklearn.svm import SVC

svc = SVC(kernel = 'linear', random_state = 42)

svc.fit(X_train, Y_train)

```

```

# Naive Bayes Algorithm

from sklearn.naive_bayes import GaussianNB

nb = GaussianNB()

nb.fit(X_train, Y_train)

# Decision tree Algorithm

from sklearn.tree import DecisionTreeClassifier

dectree = DecisionTreeClassifier(criterion = 'entropy', random_state = 42)

dectree.fit(X_train, Y_train)

# Random forest Algorithm

from sklearn.ensemble import RandomForestClassifier

ranfor = RandomForestClassifier(n_estimators = 11, criterion = 'entropy',
random_state = 42)

ranfor.fit(X_train, Y_train)

# Making predictions on test dataset

Y_pred_logreg = logreg.predict(X_test)

Y_pred_knn = knn.predict(X_test)

Y_pred_svc = svc.predict(X_test)

Y_pred_nb = nb.predict(X_test)

Y_pred_dectree = dectree.predict(X_test)

Y_pred_ranfor = ranfor.predict(X_test)

# Evaluating using accuracy_score metric from

sklearn.metrics import accuracy_score accuracy_logreg =
accuracy_score(Y_test, Y_pred_logreg) accuracy_knn =
accuracy_score(Y_test, Y_pred_knn) accuracy_svc =
accuracy_score(Y_test, Y_pred_svc) accuracy_nb =
accuracy_score(Y_test, Y_pred_nb)

```

```

accuracy_dectree = accuracy_score(Y_test, Y_pred_dectree)
accuracy_ranfor = accuracy_score(Y_test, Y_pred_ranfor)
# Accuracy on test set
print("Logistic Regression: " + str(accuracy_logreg * 100))
print("K Nearest neighbors: " + str(accuracy_knn * 100))
print("Support Vector Classifier: " + str(accuracy_svc * 100))
print("Naive Bayes: " + str(accuracy_nb * 100))
print("Decision tree: " + str(accuracy_dectree * 100))
print("Random Forest: " + str(accuracy_ranfor * 100))
# Confusion matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(Y_test, Y_pred_knn)
cm

```

CREATION OF WEB

```

import numpy as np
import pandas as pd
from flask import Flask, request, jsonify, render_template
import pickle
app = Flask(__name__)
model = pickle.load(open('model.pkl', 'rb'))
dataset = pd.read_csv('diabetes.csv')
dataset_X = dataset.iloc[:, [1, 2, 5, 7]].values
from sklearn.preprocessing import MinMaxScaler
sc = MinMaxScaler(feature_range = (0,1))
dataset_scaled = sc.fit_transform(dataset_X)
@app.route('/')

```

```

def home():

    return render_template('index.html')

@app.route('/predict',methods=['POST'])

def predict():

    float_features = [float(x) for x in request.form.values()]

    final_features = [np.array(float_features)]

    prediction = model.predict( sc.transform(final_features) )

    if prediction == 1:

        pred = "You have Diabetes, Do not neglect and please consult a Doctor immediately."

    elif prediction == 0:

        pred = "You don't have diabetes."

    output = pred

    return render_template('index.html', prediction_text='{}'.format(output))

if __name__ == "__main__":

    app.run(debug=True)

```

B) SCREENSHOTS

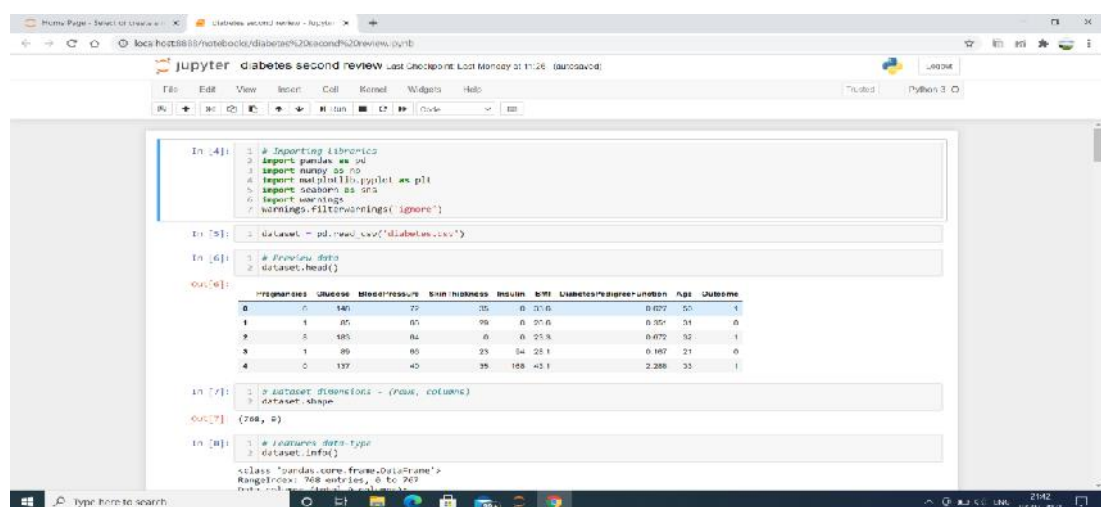


Fig B.1 Dataset

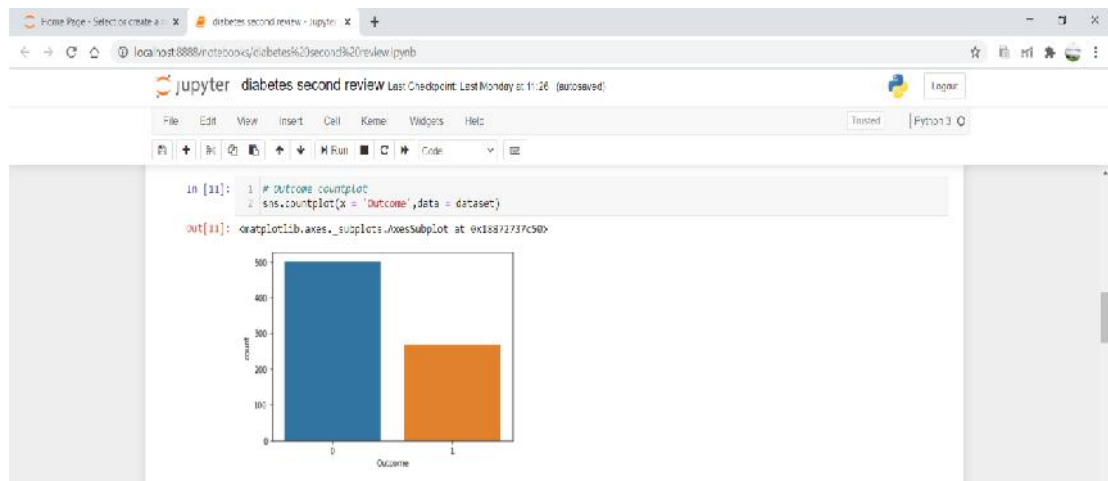


Fig B.2 Visualisation of outcome

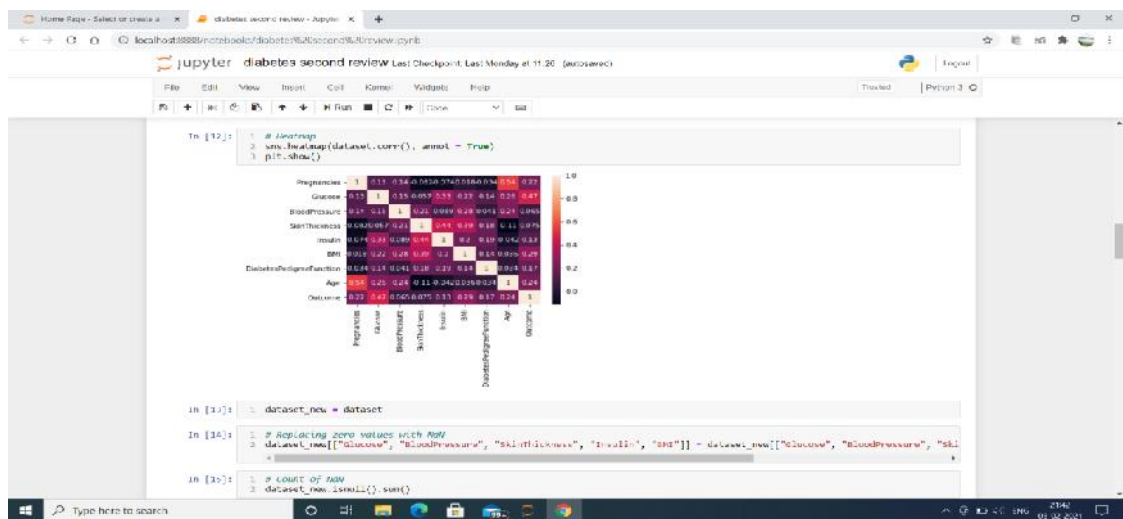


Fig B.3 Heat map of correlation of attributes

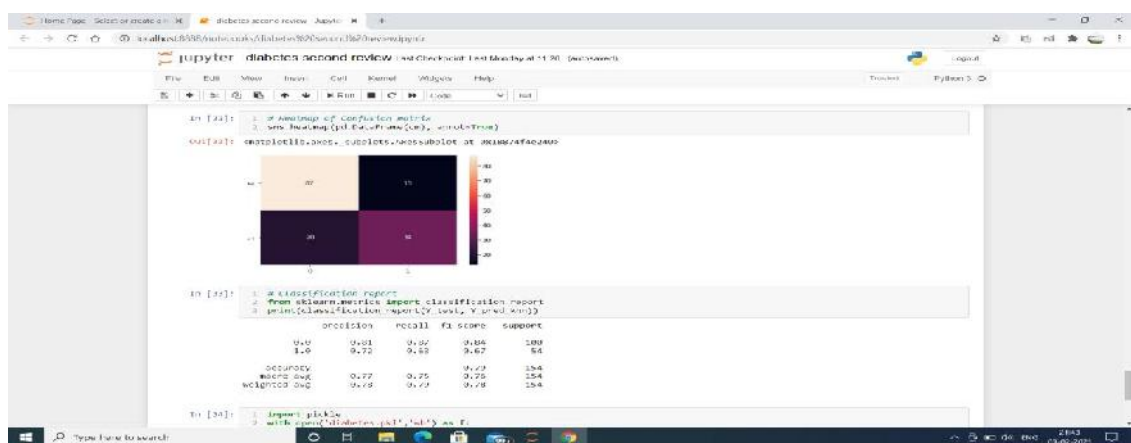


Fig B.4 Heat map of confusion matrix

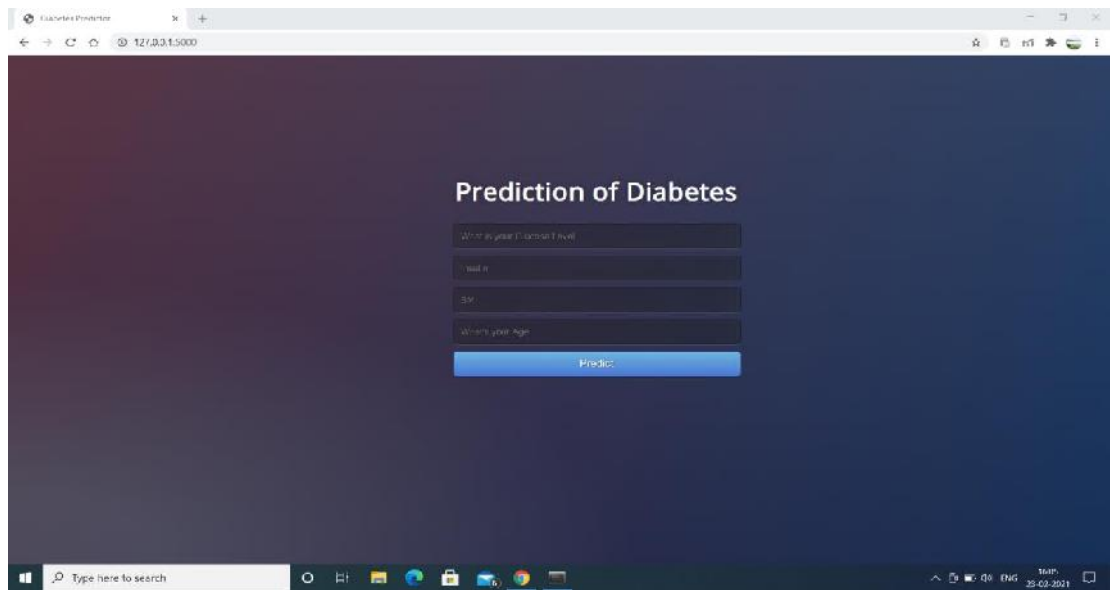


Fig B.5 Prediction Portal

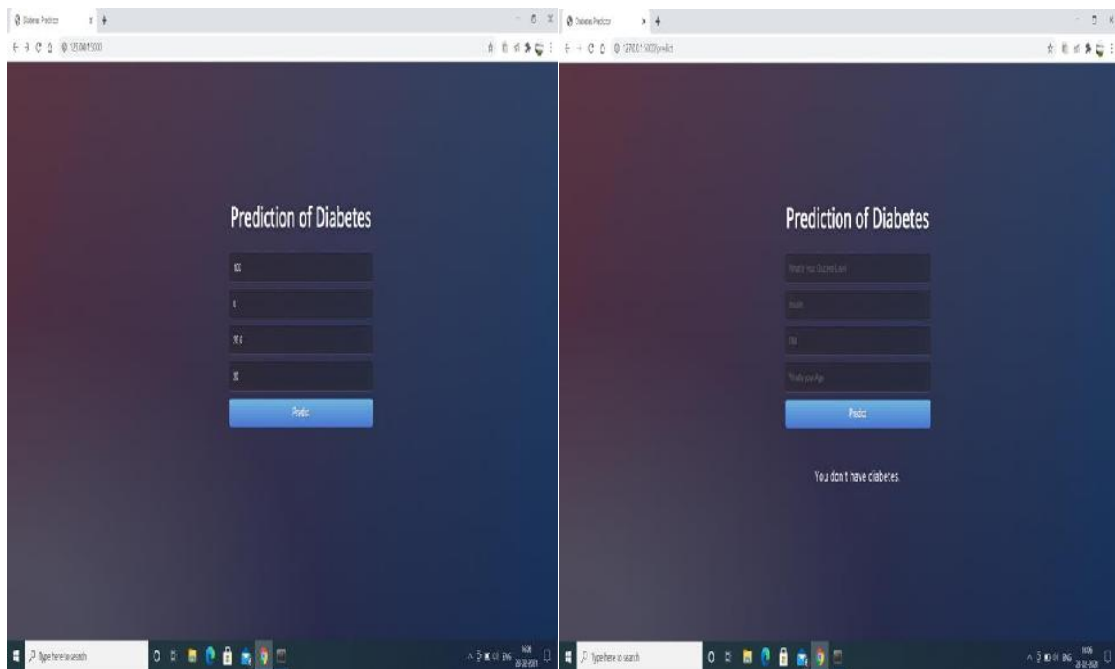


Fig B.6 Portal with inputs

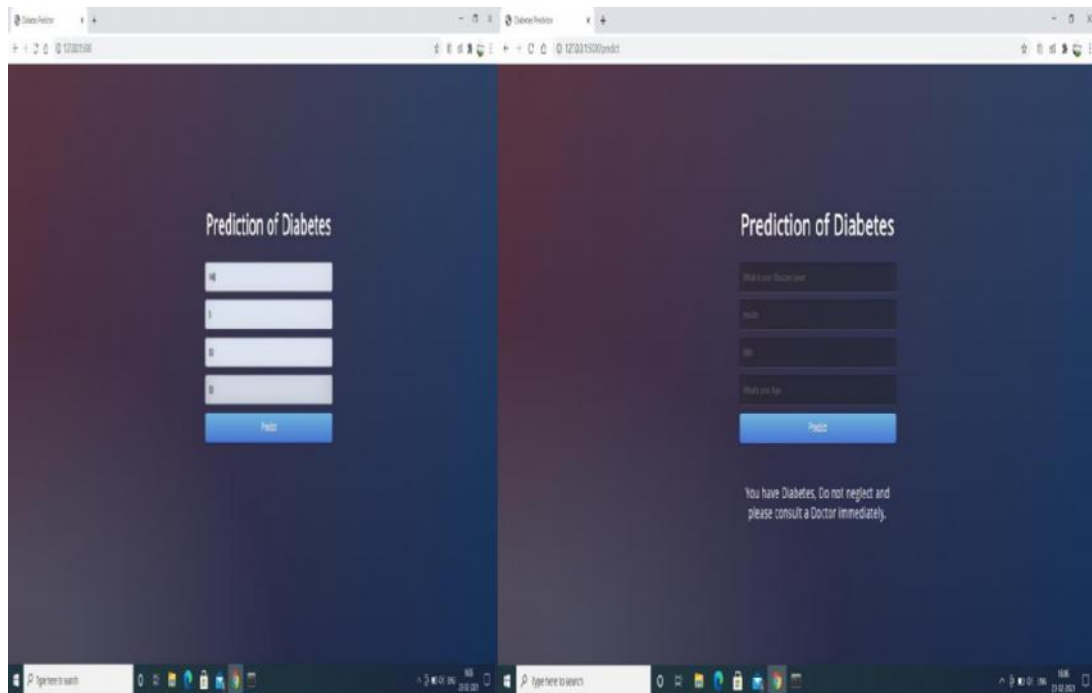
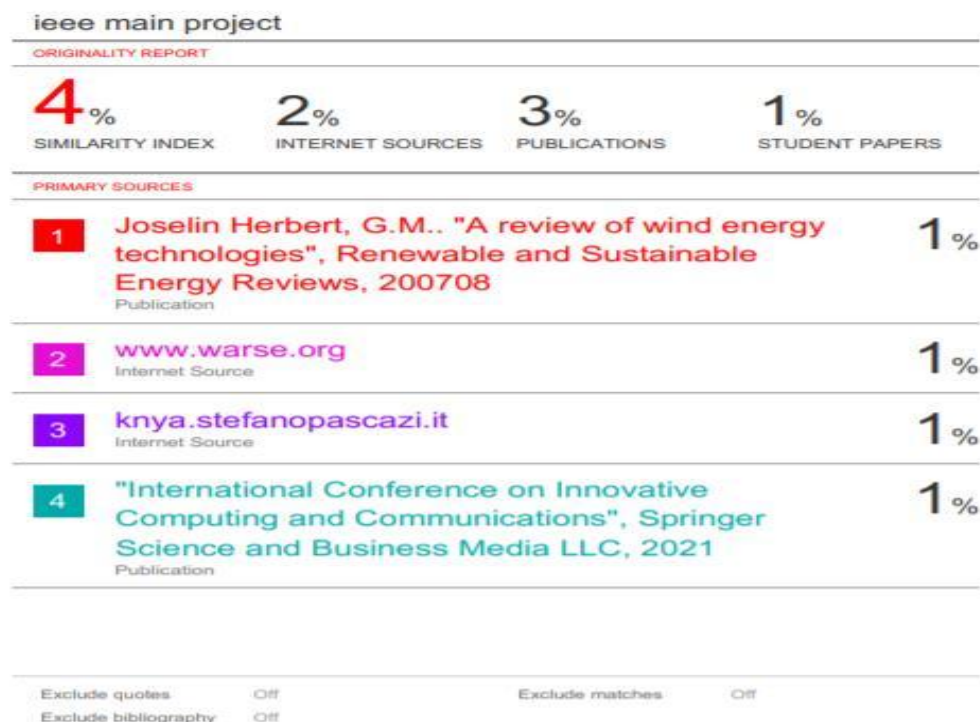


Fig B.7 Portal with inputs

C) PLAGIARISM REPORT



D) JOURNAL PAPER

WEB APPLICATION ON DIABETES PREDICTION

¹**Srujan Chowdary Putta,**

¹UG Student, CSE Department,
Sathyabama Institute of Science
and Technology, Chennai. India.

¹psrujan2000@gmail.com,

²**Rimmalapudi Pavan Sai,**

²UG Student, CSE Department,
Sathyabama Institute of Science
and Technology, Chennai. India.

²rpavansai777@gmail.com,

³**Dr.M.Selvi**

³Assistant Professor, CSE
Department, Sathyabama
Institute of Science and
Technology, Chennai.
India. ³[selvi.cse@sathyabama.ac.
in](mailto:selvi.cse@sathyabama.ac.in)

Abstract- Diabetes is a disease caused due to lack or increase of sugar level in blood. It can also be caused when the insulin is not produced by pancreas. And also when body cells are not being affected by produced insulin. Diabetes is divided into 3 types namely Juvenile onset diabetes, Adult-onset and Gestational Diabetes. Diabetes should not be ignored since it may cause many diseases which can lead to death. It also affects our body very badly. It can cause diseases related to kidney, heart, eye and so on. This paper we will do early prediction of Diabetes in a human body or a patient by applying various Machine Learning Techniques namely KNN, LR, DT, SVM, NB and RF. The accuracy is calculated for every model. The most accurate model is considered and it is used to predict the diabetes. K-Nearest Neighbors has the highest accuracy compared to other machine learning techniques. So, it is used to predict the diabetes and it is deployed as an application which can be used using website and it is done using flask library of python. By using this web app, we can easily predict whether a person has diabetes or not by just giving the values in the website.

Keywords: Diabetes Prediction, Machine Learning, Flask, Web application.

1. INTRODUCTION

Diabetes is a disease caused due to lack or increase of sugar level in blood. It can also be caused when the insulin is not produced by pancreas. And also when body cells are not being affected by produced insulin.

Diabetes is divided into 3 namely, Juvenile-onset, Adult onset and Gestational Diabetes. It also affects

our body very badly. It can cause diseases related to kidney, heart, eye and so on.

Nowadays, diabetes has become the most recognized diseases in the world, not only because of its serious effects on the human body, but also because of its difficulty of diagnosing it.

Till 2019, 4630 lakh individuals are affected by adult-onset diabetes worldwide, with adult-onset diabetes. Male and Females are affected at the same rate. It also suggests that there will be a drastic rise in diabetes in the coming years. Diabetes also leads to a risk of early death.

Juvenile-onset diabetes have the symptoms like Thirst, Frequent urination, Bed-wetting in children who have no habit of doing it previously, Hunger, Weight Loss, Irritability, Fatigue, Blurred vision and so on.

Juvenile onset diabetes can cause either by genetics or more exposure to viruses. There is no preventive measures to reduce Juvenile-onset diabetes.

Adult-onset diabetes has the same signs as that of juvenile-onset diabetes but it additionally includes slow healing of wounds, frequent infections, weight gain or loss and so on. Treatment consists of doing physical exercise, quit smoking, following diabetic diet and so on.

Make sure that diabetes will not enter your body. Once it enters, it will be so difficult to get rid of it. Take care of yourself. Nowadays, even children are getting diabetes easily. So, follow a good diet and maintain regularity in taking food, doing physical exercise and so on.

By 2030, diabetes disease rate will increase. To reduce these outcomes, early prediction of diabetes is necessary. This paper totally focuses on prediction of diabetes. We will use the Pima Indian Diabetes Dataset, which contains 768 rows and 9 columns.

The attributes of dataset include glucose level, level of insulin, etc. First we will find the correlation of all the features. Then, we will select highest correlation features and use the same for prediction of diabetes.

The Machine Learning techniques we will apply are K-Nearest Neighbors, Logistic Regression, Decision Tree, Support Vector Machine, Naive Bayes and Random Forest. The accuracy is calculated for every model.

The most accurate model is considered and it is used to predict the diabetes. K-Nearest Neighbors has the highest accuracy compared to other machine learning techniques. So, it is used to predict the diabetes and it is deployed it as an application which can be used using website and it is done using flask library of python. By using this web app, we can easily predict whether a person has diabetes or not by just giving the values in the website.

II.LITERATURE SURVEY

- Olga Tsave et al. in [1] used many machine learning models, 85% of supervised machine learning algorithms and 15% of unsupervised machine learning algorithms. Of all those, SVM has the highest accuracy. By using that, useful knowledge is extracted from datasets.
- Deepti Sisodia et al. in [2] used Pima Indians Diabetes Dataset and used Decision Tree, Naive Bayes and SVM models are implemented. Accuracies of models using various measures like recall, precision and so on. Naive Bayes has the highest accuracy of all the three and it is used to predict whether the individual suffer from diabetes or not.
- D. Dutta et al. in [3] used machine learning to find the most important features which causes diabetes and predict whether the person will get diabetes or not in future.
- N. Yuvaraj et al. in [4] used machine learning to increase the automation of health care systems and give accurate disease prediction. Pima Indians Diabetes dataset is used for evaluation of working of an algorithm.
- Akhtar Jamil et al. in [5] compared machine learning algorithms with deep learning algorithms. CNN is used for deep learning algorithm. Random forest and SVM are used for machine learning algorithms. Of all these, Random Forest has the highest accuracy. So, it is used for diabetes prediction.
- Aishwarya Mujumdar et al. in [6] predicted diabetes using external factors and also regular factors like BMI, Insulin, Glucose and so on. Pipeline model which is developed gives more accurate prediction.
- Mitesh Warke et al. in [7] done Diabetes Diagnosis using Machine Learning Algorithms. There is a static prediction engine that serves prediction results for one disease.
- Maniruzzaman, M., Rahman, et al. in [8] implemented four classifiers namely LR, DT, RF, NB and the accuracy of models is calculated. Finally, they found that combination of logistic regression and random forest gives the most accurate prediction.
- M. K. Hasan et al. in [9] implemented various ML models, Adaboost, Decision tree, XGBoost, K-Nearest Neighbors, Multi-layer perceptron and so on Pima Indians Dataset. Accurate predictions are given.
- Alexi Delgado et al. in [10] designed an application which can be used through website. ML is used as a main domain. Decision Tree classifier to predict the diabetes.
- Shruti Garg et al. in [11] used six Machine Learning Classifiers namely LR, SVM, DT, RF, NB, Gradient Boosting after which

accuracies are also calculated. In this way, they had predicted whether an individual suffers from type 2 diabetes or not.

- Ahan Chatterjee et al. in [12], they have used machine learning models like Decision tree, SVM, Random Forest, and so on, find the best classifier using the accuracy and use it to predict cancer disease risk in the early stage and prevent it. Simulation model is also designed to manage the patient flow in OPDs.

A) Existing System

From the literature survey, we have found that existing systems considers all the variables except the outcome as independent variables and predict the diabetes disease using them. And also they have used only logistic regression to predict the diabetes. They have restricted themselves till the prediction of diabetes using Jupyter notebook and not deployed it.

B) Proposed System

The proposed system takes the variables having high correlation (i.e., Glucose, Age, BMI, Insulin) with outcome as independent variables and predict the diabetes using them. Implements various models and find the most accurate one i.e., K-Nearest Neighbors and used it for prediction. At last, we will deploy the model as web application using flask web framework. By this, an individual can find easily whether they are suffering from diabetes or not.

III SYSTEM ARCHITECTURE

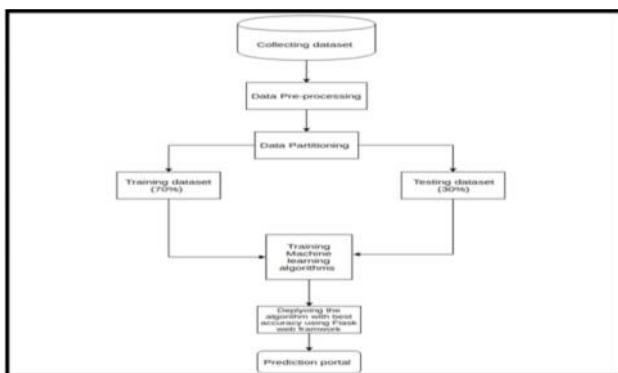


Fig 1 System Architecture

IV METHODOLOGY

The libraries used in this project are-

- Flask- It is used for deploying machine learning model as web application
- Pandas- Used to read the dataset
- Matplotlib- Used for data visualisation
- Sklearn- Used for importing models, get accuracy and confusion matrix

Web Application on Diabetes Prediction have the following modules:

1. Dataset Collection

Collect the dataset, clean it, split into train and test dataset. This is done in Jupyter Notebook.

2. Train the machine learning algorithms

Train the dataset using various ML models, we found that the most accurate algorithm is k-nearest neighbors and use it for prediction of diabetes. This is done in Jupyter Notebook.

3. Deploying Machine learning model as web app

Deploy the machine learning model as web app using flask web framework and prediction portal will appear.

The steps followed in achieving the objective are given below:

1. Collect the dataset from kaggle notebooks.
2. Data Pre-processing like checking for missing values, out-of-range values and so on is done.
3. Data Partitioning i.e., divide the dataset into holdout and training dataset.
4. Train the dataset by ML models and evaluate accuracies.
5. Deploy the most accurate machine learning model i.e., K-Nearest Neighbors as a web app using Flask web framework.
6. Now, the prediction portal appears in which we can give the values and predict whether the person has diabetes or not.

V RESULTS

In the current situation of COVID-19, where sufficient resources are unavailable, our prediction portal can help in prediction of diabetes easily. This also helps in reduction of the cost required.

Accuracy of different models arranged in descending order is shown in the table. Of these models K-Nearest Neighbors has the most highest accuracy, so it is used to predict.

Model	Accuracy
K-Nearest Neighbors	78%
Random Forest	75%
Support Vector Machines	73%
Logistic Regression	72%
Naive Bayes	71%
Decision Tree	68%

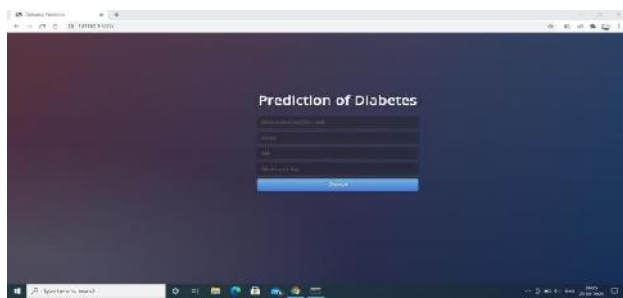


Fig 2 Prediction portal

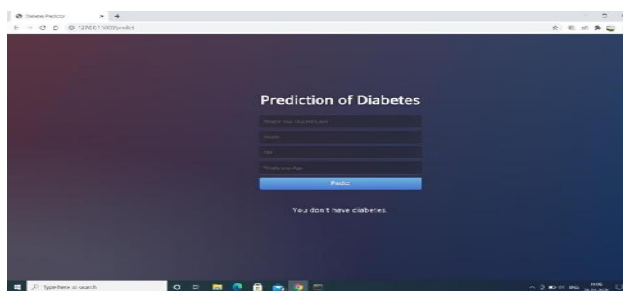


Fig 3 Output-1

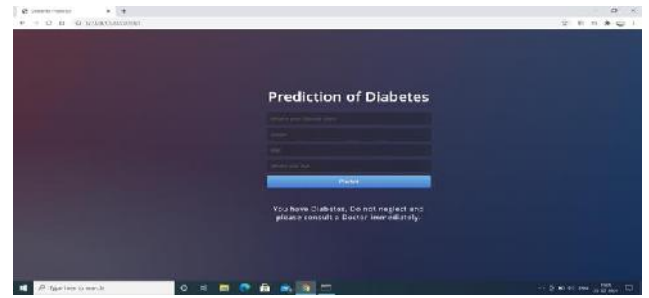


Fig 4 Output-2

VI. CONCLUSION

In the current situation of COVID-19, where sufficient facilities and resources are unavailable, our prediction system can prove to be helpful and can be used in the diagnosis of a disease.

Diabetes is a disease caused due to lack or increase of sugar level in blood. It can also be caused when the insulin is not produced by pancreas. And also when body cells are not being affected by produced insulin.

Real-time analysis of the PIMA Indian Dataset is done. The k-nearest neighbors gave the highest accurate prediction than the other models. Flask API has been used to deploy trained model with the HTML page.

The HTML page has been displayed in such a way that the user only needs to input the values of mass index of the body, BP, Level of glucose, Insulin level and so on. The HTML page provides the result whether the user has diabetes or not on the webpage indicating positive if the person has diabetes and negative if the person does not have diabetes.

REFERENCES

- [1] Olga Tsave et al. "Machine Learning and Data Mining Methods in Diabetes Research", Journal, Volume 15, 2017
- [2] Deepti Sisodia et al. "Prediction of Diabetes using Classification Algorithms", Volume 132, 2018
- [3] D. Dutta, D. Paul and P. Ghosh, "Analysing Feature Importances for Diabetes Prediction using Machine Learning," 2018 IEEE 9th Annual Information Technology, pp.924-928, 10.1109/IEMCON.2018.8614871.

- [4] N. Yuvaraj et al. "Diabetes prediction in healthcare systems using machine learning on hadoop cluster", 16 January 2019.
- [5] Akhtar Jamil et al. "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques", 2019, pp. 1-4, doi: 10.1109/UBMYK48245.2019.8965556.
- [6] Aishwarya Mujumdar et al. "Diabetes Prediction using Machine Learning Algorithms", Volume 165, 2019.
- [7] Mitesh Warke et al. "Diabetes Diagnosis using Machine Learning Algorithms", Volume: 06 Issue: 03 Mar 2019
- [8] Maniruzzaman, M., Rahman, M.J., Ahammed, B. et al. "Classification and prediction of diabetes disease using machine learning" Syst 8, 7 (2020), 3 January 2020.
- [9] M. K. Hasan, M. A. Alam, D. Das, E. Hossain and M. Hasan, "Diabetes Prediction Using Ensembling Of Different Machine Learning Classifiers," in IEEE Access, vol. 8, pp. 76516-76531, 2020.
- [10] Alexi Delgado et al. "Design of a Web Application for the Detection of Diabetes using Machine Learning", Volume 8. No. 7, July 2020.
- [11] Shruti Garg et al. "Prediction of Type 2 Diabetes using Machine Learning Classification Methods" 167 (2020) 706–716, Volume 167, 2020.
- [12] Ahan Chatterjee et al. "A Machine Learning Approach to prevent cancer", DOI: 10.4018/978-1-7998-2742-9.ch007, 2021. 167 (2020) 706–716, Volume 167, 2020.