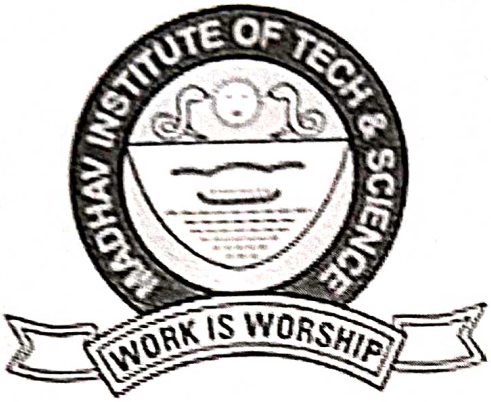
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR
(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)



Project Report

on

Diabetes Predictor

Submitted By:

Ritik Malarya

0901CS191100

Shashwat Sharma

0901CS191114

Faculty Mentor:

Mr. Mir Shahnawaz Ahmad

Assistant Professor, Computer Science and Engineering

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE
GWALIOR - 474005 (MP) est. 1957

MAY-JUNE 2022

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR
(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

Project Report

on

Diabetes Predictor

A project report submitted in partial fulfilment of the requirement for the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

Submitted by:

Ritik Maharya

0901CS191100

Shashwat Sharma

0901CS191114

Faculty Mentor:

Mr. Mir Shahnawaz Ahmad

Assistant Professor, Computer Science and Engineering

Submitted to:

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE
GWALIOR - 474005 (MP) est. 1957

MAY-JUNE 2022

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

CERTIFICATE

This is certified that **Ritik Malarya** (0901CS191100) has submitted the project report titled **Diabetes Predictor** under the mentorship of **Mr. Mir Shahnawaz Ahmad,** in partial fulfilment of the requirement for the award of degree of Bachelor of Technology in Computer Science and Engineering from Madhav Institute of Technology and Science, Gwalior.

09/05/22

Mr. Mir Shahnawaz Ahamad
Faculty Mentor
Assistant Professor
Computer Science and Engineering

09/05/2022

Dr. Manish Dixit
Professor and Head
Computer Science and Engineering

Dr. Manish Dixit
Professor & HOD
Department of CSE
M.I.T.S. Gwalior

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

CERTIFICATE

This is certified that **Shashwat Sharma (0901CS191114)** has submitted the project report titled **Diabetes Predictor** under the mentorship of **Mr. Mir Shahnawaz Ahmad,** in partial fulfilment of the requirement for the award of degree of Bachelor of Technology in Computer Science and Engineering from Madhav Institute of Technology and Science, Gwalior.

09/05/22

Mr. Mir Shahnawaz Ahamad
Faculty Mentor
Assistant Professor
Computer Science and Engineering

09/05/2022

Dr. Manish Dixit
Professor and Head
Computer Science and Engineering

Dr. Manish Dixit
Professor & HOD
Department of CSE
M.I.T.S. Gwalior

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

DECLARATION

We hereby declare that the work being presented in this project report, for the partial fulfilment of requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering at Madhav Institute of Technology & Science, Gwalior is an authenticated and original record of my work under the mentorship of **Mr. Mir Shahnawaz Ahmad, Assistant Professor**, Computer Science and Engineering.

We declare that we have not submitted the matter embodied in this report for the award of any degree or diploma anywhere else.

Ritik Malarya
0901CS191100
3rd Year
Computer Science and Engineering

Shashwat sharma
0901CS191114
3rd Year
Computer Science and Engineering

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR
(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science,** for allowing me to continue my disciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Department of Computer Science and Engineering, for allowing** me to explore this project. I humbly thank **Dr. Manish Dixit**, Professor and Head, Department of Computer Science and Engineering, for his continued support during the course of this engagement, which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance of **Mr. Mir Shahnawaz Ahmad**, **Assistant Professor**, Computer Science and Engineering for their continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department.

Ritik Malarya
0901CS191100
3rd Year
Computer Science and Engineering

Shashwat Sharma
0901CS191114
3rd Year
Computer Science and Engineering

Abstract

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes is a disease caused due to the increase level of blood glucose. This high blood glucose produces the symptoms of frequent urination, - increased thirst, and increased hunger. Diabetes is a one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy. But with diabetes, this system does not work. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. The algorithms like K nearest neighbour, Logistic Regression, Random forest, Support vector machine and Decision tree are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the diabetes.

.**Keywords:** : Machine Learning, Diabetes, Decision tree, K nearest neighbour,Logistic Regression, Support vector Machine, Accuracy

सार:

मधुमेह एक पुरानी बीमारी है जिसमें दुनिया भर में स्वास्थ्य पैदा करने की क्षमता है देखभाल संकट। इंटरनेशनल डायबिटीज फेडरेशन के अनुसार 382 मिलियन लोग पूरी दुनिया में मधुमेह के साथ जी रहे हैं। 2035 तक यह दोगुना हो जाएगा 592 मिलियन के रूप में। मधुमेह एक रोग है जो रक्त के स्तर में वृद्धि के कारण होता है ग्लूकोज। यह उच्च रक्त शर्करा बार-बार पेशाब आने के लक्षण पैदा करता है, प्यास बढ़ी, और भूख बढ़ी। मधुमेह का एक प्रमुख कारण है अंधापन, गुर्दे की विफलता, विच्छेदन, दिल की विफलता और स्ट्रोक। जब हम खाते हैं, हमारा शरीर भोजन को शर्करा या ग्लूकोज में बदल देता है। उस समय, हमारा अग्न्याशय है इंसुलिन जारी करना चाहिए। इंसुलिन हमारी कोशिकाओं को खोलने, अनुमति देने के लिए एक कुंजी के रूप में कार्य करता है ग्लूकोज में प्रवेश करने और हमें ऊर्जा के लिए ग्लूकोज का उपयोग करने की अनुमति देता है। लेकिन इसके साथ मधुमेह, यह प्रणाली काम नहीं करती है। टाइप । और टाइप 2 मधुमेह सबसे ज्यादा हैं रोग के सामान्य रूप हैं, लेकिन अन्य प्रकार भी हैं, जैसे कि गर्भावधि
मधुमेह, जो गर्भावस्था के दौरान होता है, साथ ही अन्य रूपों में भी होता है। मशीन सीखना डेटा विज्ञान में एक उभरता हुआ वैज्ञानिक क्षेत्र है, जिसमें तरीकों से निपटना है कौन सी मशीनें अनुभव से सीखती हैं। इस परियोजना का उद्देश्य एक विकसित करना है प्रणाली जो एक रोगी के लिए मधुमेह की प्रारंभिक भविष्यवाणी कर सकती है
विभिन्न मशीन लर्निंग के परिणामों को मिलाकर उच्च सटीकता तकनीक। K निकटतम पड़ोसी जैसे एल्गोरिदम, लॉजिस्टिक रिग्रेशन, रैंडम फ़ॉरेस्ट, सपोर्ट वेक्टर मशीन और डिसीजन ट्री का उपयोग किया जाता है।
प्रत्येक एल्गोरिदम का उपयोग करके मॉडल की सटीकता की गणना की जाती है। फिर एक अच्छी सटीकता के साथ मधुमेह की भविष्यवाणी के लिए मॉडल के रूप में लिया जाता है।

कीवर्ड: मशीन लर्निंग, डायबिटीज, डिसीजन ट्री, K निकटतम पड़ोसी, लॉजिस्टिक रिग्रेशन, सपोर्ट वेक्टर मशीन, एक्यूरेसी।

TABLE OF CONTENTS

CHAPTER 1: PROJECT OVERVIEW

1.1 Introduction

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 everybody, even those people with diabetes, needs carbohydrate. Carbohydrate foods include bread, cereal, pasta, rice, 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

1.2 Objective and Scope

Aim is 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. Using PIMA dataset and cross validation approach 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.

1.3 Project Features

Project 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.

1.4 Technologies used in Diabetes Predictor

- Python
- Machine Learning
- HTML
- CSS
- Flask

1.4.1 Python

Python is a high-level, interpreted, interactive, and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently whereas other languages use punctuation, and it has fewer syntactical constructions than other languages.

1.4.2 Machine Learning

Machine learning is a method of data analysis that automates analytical model building. It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

1.4.3 Jupyter Notebook

Jupyter Notebook is a web-based interactive computational environment for creating notebook documents. A Jupyter Notebook document is a browser-based REPL containing an ordered list of input/output cells which can contain code, text, mathematics, plots and rich media. Underneath the interface, a notebook is a JSON document, following a versioned schema, usually ending with the ".ipynb" extension.

1.4.4 Visual Studio Code

Visual Studio Code is a source-code editor made by Microsoft for Windows, Linux and macOS. Features include support for debugging, syntax highlighting, intelligent code completion, snippets, code refactoring, and embedded Git. Users can change the theme, keyboard shortcuts, preferences, and install extensions that add additional functionality.

1.4.5 Anaconda

Anaconda is a distribution of the Python and R programming languages for scientific computing that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows, Linux, and macOS. We used anaconda to get access to Anaconda Navigator and Anaconda Prompt.

1.4.6 Libraries Used

1.4.6.1 Numpy

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. NumPy is open-source software and has many contributors.

1.4.6.2 Pandas

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.

1.4.6.3 Seaborn

Seaborn is a data visualization library built on top of matplotlib and closely integrated with Pandas data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.

1.4.6.4 Matplotlib

Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK.

1.4.6.5 Scikit-Learn

Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

1.4.6.6 Pickle

Pickle module is used for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk. What pickle does is that it "serializes" the object first before writing it to file. Pickling is a way to convert a python object into a character stream. The idea is that this character stream contains all the information necessary to reconstruct the object in another python script.

System Requirements

Windows Based Requirements:

Computers running Microsoft Windows must meet the following minimum hardware and software requirements.

Microsoft Windows: 7/8/10/11

GB RAM minimum, 8 GB RAM recommended 1GB of available disk space minimum

1280 * 800 minimum screen resolution

Software Requirement: Python 3.10.4

Hardware Requirement:

Laptop/ComputerInternet Connectivity

CHAPTER 2: LITERATURE REVIEW

It 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

It 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. Using PIMA dataset and cross validation approach 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 after applying the resample filter over the data. The author emphasis on the class imbalance problem and the need to handle this problem before applying any algorithm to achieve better accuracy rates. The class imbalance is a mostly occur in a dataset having dichotomous values, which means that the class variable have two possible outcomes and can be handled easily if observed earlier in data preprocessing stage and will help in boosting the accuracy of the predictive model.

CHAPTER 3: PRELIMINARY DESIGN

1.5Data Flow Diagram
1.6

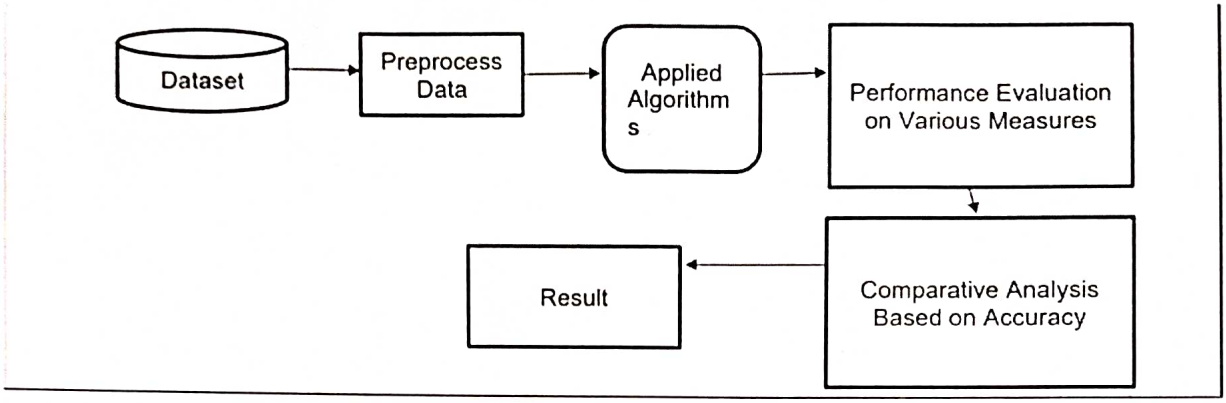


Fig 3.2.1 Data flow diagram

Dataset Description:
The diabetes data set was originated from Diabetes dataset containing 2000 cases. The objective is to predict based on the measures to predict if the patient is diabetic or not.

Link: https://www.kaggle.com/johndasilva/diabetes

Preprocess Data: here data is devided into two different sets i.e training data set and testing Data set

Applied Algorithm :

	Training Accuracy	Testing Accuracy
C=1	0.779	0.788
C=0.01	0.784	0.780
C=100	0.778	0.792

In first row, the default value of C=1 provides with 77% accuracy on the training and 78% accuracy on the test set.
In second row, using C=0.01 results are 78% accuracy on both the training and the test sets.

Using C=100 results in a little bit lower accuracy on the training set and little bit highest accuracy on the test set, confirming that less regularization and a more complex model may not generalize better than default setting

Performance Evaluation on Various Measures :

Comparison of Various machine learning Classifier models is evaluated to the Diagnosis of Diabetes. Performance accuracy of the classifiers is evaluated based on Incorrectly and Correctly Classified Instances out of a total number of instances.

Result : The entire outcomes of the experiment in terms of accuracy, precision, recall, and f1-score are presented in Fig. 7. For NB, DT, RF, SVM, LR, GB, and KNN, the accuracy of these models is 86.17%, 96.81%, 96.81%, 91.49%, 84.04%, 90.43%, and 90.43%, respectively. This table illustrates that DT and RF both provide the highest level of accuracy and exceed the other approaches.

CHAPTER 4: FINAL ANALYSIS AND RESULT

4.1 Final Application UI/UX

Home Page:

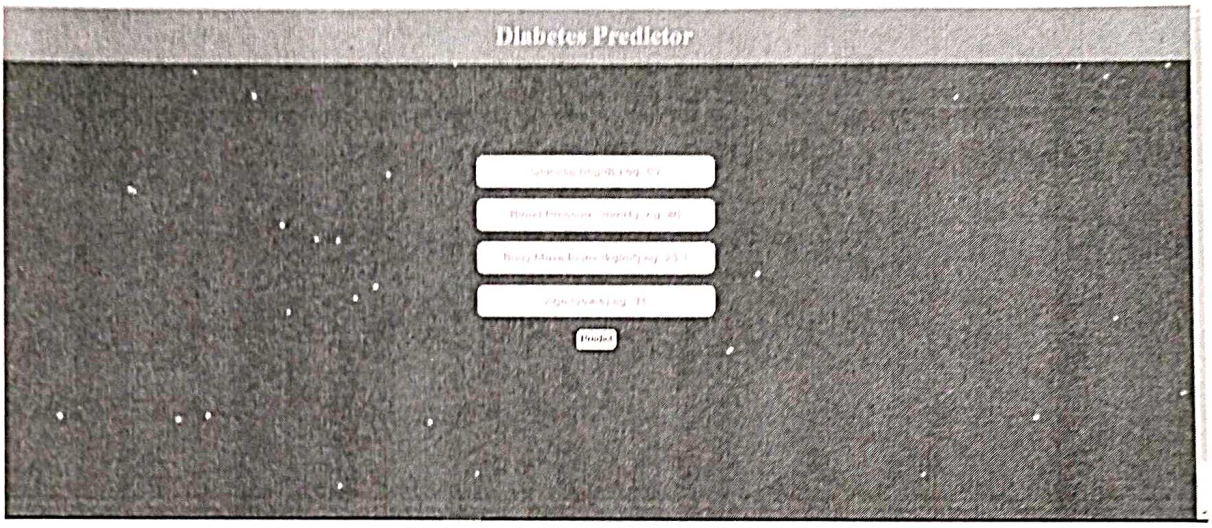


Fig 1

Output:

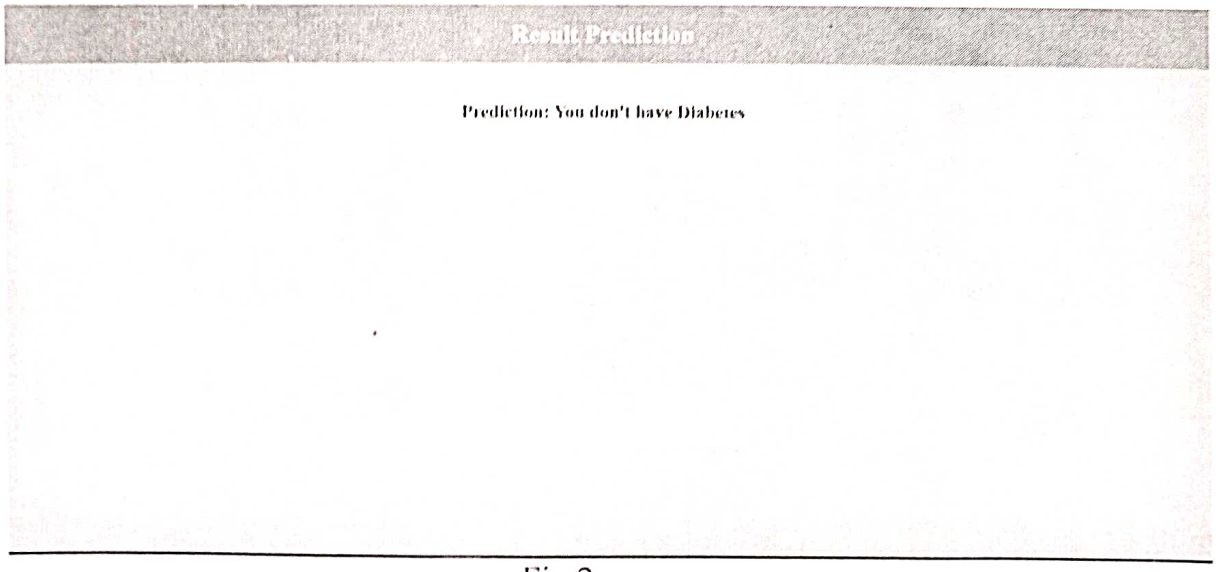


Fig 2

4.2 RESULT:

In this work different steps were taken. The proposed approach uses different classification and ensemble methods and implemented using python. These methods are standard Machine Learning methods used to obtain the best accuracy from data. In this work we see that random forest classifier achieves better compared to others. Overall we have used best Machine Learning techniques for prediction and to achieve high performance accuracy.

Here feature played important role in prediction is presented for random forest algorithm. The sum of the importance of each feature playing major role for diabetes have been plotted, where X-axis represents the importance of each feature and Y-Axis the names of the features

CHAPTER 5: CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning method in which SVM, Knn, Random Forest, Decision Tree, Logistic Regression and Gradient Boosting classifiers are used. And 77% classification accuracy has been achieved. The Experimental results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life.

In this work different steps were taken. The proposed approach uses different classification and ensemble methods and implemented using python. These methods are standard Machine Learning methods used to obtain the best accuracy from data. In this work we see that random forest classifier achieves better compared to others. Overall we have used best Machine Learning techniques for prediction and to achieve high performance accuracy. Figure shows the result of these Machine Learning methods.

5.2 Future Scope

- We will try to enhance the dataset by including more cities (like Mumbai, Hyderabad, Pune, Bangalore, etc.)
- We will try to build a user interface.
- We will try to incorporate Restaurant Recommendation System based on Collaborative filtering.
- We will try to improve our data analysis and prediction by adding more features like life expectancy, and customer reviews on the demographic location.
- We will try to add a Customer review system for the pre-existing restaurants in that location which will help to build a community.

References

1. Berumen Calderón, Mauro Felipe, Damayanti Estolano Cristerna, Angelica Selene Sterling Zozoaga, and Andreé Ricardo Berumen Calderón. "Model to assess the selection of the optimal location for a restaurant, a quantitative approach. Case study: Theme restaurants in Cancun, Mexico." *Journal of Foodservice Business Research* 24, no. 4 (2021): 457-501.

2. Hariharan, R., Arish Pitchai, and M. Dhilsath Fathima. "Prediction of Locations Using Unsupervised Learning Method to Open a Restaurant Branch." In *Conference on Multimedia, Interaction, Design and Innovation*, pp. 59-71. Springer, Cham, 2020.

3. Baksi, Bidisha Das, Varsha Rao, and C. Anitha. "A Survey on Local Market Analysis for a Successful Restaurant Yield." In *Emerging Technologies in Data Mining and Information Security*, pp. 249-257. Springer, Singapore, 2019.

4. Dock, Joel P., Wei Song, and Jia Lu. "Evaluation of dine-in restaurant location and competitiveness: Applications of gravity modeling in Jefferson County, Kentucky." *Applied Geography* 60 (2015): 204-209.