VISVESVARAYA TECHNOLOGICAL UNIVERSITY

BELAGAVI-590018

A Internship Project Report

on

Diabetes Prediction using ML

Submitted in partial fulfillment of the requirements for the VI semester

Computer Science and Engineering of Visvesvaraya Technological University, Belagavi

Submitted by:

Tejas R 1RN20CS167

Under the Guidance of:

Mrs.Chethana H R
Associate Professor
Dept. of CSE

ESTD : 2001
An Institute with a Difference

Department of Computer Science and Engineering
(Accredited by NBA up to 30/6/2025)
RNS Institute of Technology
Channasandra, Dr.Vishnuvardhan Road, Bengaluru-560 098
2023

RNS INSTITUTE OF TECHNOLOGY

Channasandra, Dr.Vishnuvardhan Road, Bengaluru-560098

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

(Accredited by NBA up to 30/6/2025)

ESTD : 2001
An Institute with a Difference

CERTIFICATE

This is to certify that the mini project work entitled **Diabetes Prediction using ML** has been successfully carried out by **Tejas R** bearing USN **1RN20CS167**, bonafide student of **RNS Institute of Technology** in partial fulfillment of the requirements for the 7th semester **Computer Science and Engineering of Visvesvaraya Technological University"**, Belagavi, during academic year 2023. It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report deposited in the departmental library. The internship project report has been approved as it satisfies the internship requirements of 7th semester BE, CSE.

Signature of the Guide	Signature of the HoD	Signature of the Principal
Mrs.Chethana H R	**Dr. Kiran P**	**Dr. Ramesh Babu H S**
Associate Professor	Professor & Head	Principal
Dept. of CSE	Dept. of CSE	

External Viva:

Name of the Examiners	Signature with Date
1.	
2.	

Acknowledgement

The successful completion of any achievement is not solely dependent on individual efforts but also on the guidance, encouragement, and cooperation of intellectuals, elders, and friends. We would like to take this opportunity to express our heartfelt gratitude to all those who have contributed to the successful execution of this project.

First and foremost, we extend our profound thanks to **Sri. Satish R Shetty**, Managing Trustee of R N Shetty Trust and Chairman of RNS Group of Institutions, and **Sri. Karan S Shetty** , CEO of RNS Group of Institutions, Bengaluru, for providing a conducive environment that facilitated the successful completion of this project.

We would also like to express our sincere appreciation to our esteemed Director, **Dr. M K Venkatesha**, for providing us with the necessary facilities and support throughout the duration of this work.

Our heartfelt thanks go to our respected Principal, **Dr. Ramesh Babu H S**, for his unwavering support, guidance, and encouragement that played a vital role in the completion of this project.

We would like to extend our wholehearted gratitude to our HOD, **Dr. Kiran P**, Professor, and Head of the Department of Computer Science & Engineering, RNSIT, Bangalore, for his valuable suggestions and expert advice, which greatly contributed to the success of this endeavor.

A special word of thanks is due to our project guide,**Mrs.Chethana H R**, Associate Professor in the Department of CSE, RNSIT, Bangalore, for her exceptional guidance, constant encouragement, and unwavering assistance throughout the project.

We would also like to express our sincere appreciation to all the teaching and non-teaching staff of the Department of Computer Science & Engineering, RNSIT, for their consistent support and encouragement.

Once again, we express our deepest gratitude to everyone involved, as their support and cooperation were instrumental in the successful completion of this project.

Abstract

Diabetes mellitus is a chronic metabolic disorder with a growing global prevalence, making early detection and prevention crucial for public health. This abstract presents a comprehensive overview of the application of Machine Learning (ML) techniques in predicting the risk of diabetes.Several studies have demonstrated the effectiveness of ML-based diabetes prediction models, achieving high accuracy, sensitivity, and specificity. These models not only aid in early diagnosis but also contribute to personalized medicine by tailoring interventions and treatment plans. Additionally, ML models can highlight the significant risk factors contributing to diabetes, facilitating a better understanding of the disease's etiology.In conclusion, ML-based diabetes prediction models hold significant promise in revolutionizing healthcare by enabling early risk assessment and personalized intervention strategies. Continued research and development in this field are critical for improving model performance, enhancing interpretability, and ultimately reducing the global burden of diabetes.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 ORGANIZATION/INDUSTRY

1.1.1 COMPANY PROFILE

NASTECH is formed with the purpose of bridging the gap between Academia and Industry Nastech is one of the leading Global Certification and Training service providers for technical and management programs for educational institutions. We collaborate with educational institutes to understand their requirements and form a strategy in consultation with all stakeholders to fulfill those by skilling, reskilling and upskilling the students and faculties on new age skills and technologies.

1.1.2 DOMAIN/TECHNOLOGY

The domain chosen for our project is AI/ML. Machine learning, the fundamental driver of AI, is possible through algorithms that can learn themselves from data and identify patterns to make predictions and achieve your predefined goals, rather than blindly following detailed programmed instructions, like in traditional computer programming. This technology allows the machine to perceive, learn, reason and communicate through observation of data, like a child that grows up and acquires knowledge from examples. Machines also have the advantage of not being limited by our inherent biological limitations. With machine learning, manufacturing companies have increased production capacity up to 20Nowadays, the revolutionary AI technology evolved from rule-based expert systems to machine learning and more advanced subcomponents such as deep learning (learning representations instead of tasks), artificial neural networks (inspired by animal brains) and reinforcement learning (virtual agents rewarded if they made good decisions). The AI can master the complexity of the intertwining industrial processes to enhance the whole flow of production instead of isolated processes. This enormous cognitive capacity gives the AI the ability to consider the spatial organization of plants and the timing constraints of live production. Another key advantage is the capability of AI algorithms to think probabilistically with all the subtlety this allows in edge cases, instead of traditional rule-based methods that require rigid theories and a full comprehension of problems.

- The representation and manipulation of image data by a computer.
- The various technologies used to create and manipulate images.
- The sub-field of computer science which studies methods for digitally synthesizing and manipulating visual content.

Today, computers and computer-generated images touch many aspects of daily life. Computer images is found on television, in newspapers, for example in weather reports, in all kinds of medical investigation and surgical procedures. A well-constructed graph can present complex statistics in a form that is easier to understand and interpret. In the media such graphs are used to illustrate papers, reports, thesis, and other presentation material.Many powerful tools have been developed to visualize data. Computer generated imagery can be categorized into several different types: 2D, 3D, 4D, 7D, and animated graphics.
As technology has improved, 3D computer graphics have become more common. Computer graphics has emerged as a sub-field of computer science which studies methods for digitally synthesizing and manipulating visual content. Over the past decade, other specialized fields have been developed like information visualization, and scientific visualization more concerned with the visualization of three dimensional phenomena (architectural, meteorological, medical, biological, etc.), where the emphasis is on realistic renderings of volumes, surfaces, illumination sources, and so forth, perhaps with a dynamic component.

1.2 Problem Statement

The problem at hand is to develop an accurate and interpretable machine learning model for diabetes prediction that addresses the following key challenges:

- **High Accuracy:**
  Create a predictive model that can accurately identify individuals at risk of developing diabetes, ensuring a low rate of false positives and false negatives.

- **Early Detection:** Develop a model capable of predicting diabetes risk well in advance of clinical diagnosis, enabling timely interventions and lifestyle modifications.

- **Feature Selection:** Determine the most relevant and informative features for diabetes prediction, taking into account various data sources, such as demographic information, medical history, genetic markers, and lifestyle factors.

- **Interpretability:** Ensure that the model's predictions are interpretable, providing insights into the factors contributing to an individual's diabetes risk. This will aid healthcare professionals in making informed decisions and patients in understanding their risk factors.

- **Data Quality:** Address data quality issues by implementing robust preprocessing techniques to handle missing values, outliers, and noisy data, thus improving model performance.

- **Ethical Considerations:** Develop the model and associated protocols with a strong ethical framework, emphasizing informed consent, transparency, and fairness in predictions, particularly in sensitive healthcare applications.

- **Generalization:** Develop a model that can generalize well across different populations, ethnicities, and geographic regions, taking into account the variability in diabetes risk factors.

Chapter 2

Resource Requirements

2.1 Hardware Requirements

The Hardware requirements are very minimal and the program can be run on most of the machines.Table 2.1 gives details of hardware requirements.

Table 2.1: Hardware Requirements

Processor	Intel Core i3 processor
Processor Speed	1.70 GHz
RAM	4 GB
Storage Space	40 GB
Monitor Resolution	1024*768 or 1336*768 or 1280*1024

2.2 Software Requirements

The software requirements are description of features and functionalities of the system.Table 2.2 gives details of software requirements.

Table 2.2: Software Requirements

Operating System	Windows 8.1
IDE	Anaconda
Libraries	Pandas,NumPy,Streamlit,Matplotlib,Seaborn.

2.2.1 Anaconda

Anaconda is the birthplace of Python data science. It 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. It is developed and maintained by Anaconda, Inc., which was founded by Peter Wang and Travis Oliphant in 2012. Anaconda distribution comes with over 250 packages automatically installed, and over 7,500 additional open-source packages can be installed from PyPI as well as the conda package and virtual environment manager. It also includes a GUI, Anaconda Navigator, as a graphical alternative to the command-line interface.

2.2.2 Jupyter Notebook

Jupyter Notebook (formerly IPython Notebook) is a web-based interactive computational environment for creating notebook documents. Jupyter Notebook is built using several open-source libraries, including IPython, ZeroMQ, Tornado, jQuery, Bootstrap, and MathJax. A Jupyter Notebook document is a browser-based REPL containing an ordered list of input/output cells which can contain code, text (using Markdown), mathematics, plots and rich media. Underneath the interface, a notebook is a JSON document, following a versioned schema, usually ending with the ".ipynb" extension. Jupyter Notebook is similar to the notebook interface of other programs such as Maple, Mathematica, and SageMath, a computational interface style that originated with Mathematica in the 1980s. Jupyter interest overtook the popularity of the Mathematica notebook interface in early 2018.

Chapter 3

Design

The description of all the phases performed in the project is given below:

- **Data Collection:**

  Gather relevant data from various sources. This may include structured data from databases, spreadsheets, or APIs, as well as unstructured data like text or images.

- **Data Preprocessing:**

  Prepare and clean the data to make it suitable for analysis. This involves tasks such as: Handling missing values. Removing duplicates. Encoding categorical variables. Scaling or normalizing numeric features. Handling outliers. Feature engineering to create new informative features.

- **Data Splitting:**

  Divide your dataset into training, validation, and testing sets. The training set is used to train the model, the validation set helps tune hyperparameters, and the testing set is used to evaluate model performance.

- **Feature Selection/Extraction:**

  Choose the most relevant features for your predictive model. Feature selection helps reduce dimensionality, while feature extraction may involve creating new features from existing ones.

- **Model Selection:**

  Choose an appropriate machine learning algorithm or model for your problem. Consider factors like data type (classification or regression), model complexity, interpretability, and the size of your dataset.

  .

- **Model Training:**

  Train the selected model using the training dataset. The model learns patterns and relationships within the data to make predictions.

- **Hyperparameter Tuning:**

  Optimize the model's hyperparameters to improve its performance. Techniques like grid search or random search can be used to find the best combination of hyperparameters.

- **Model Evaluation:**

  Assess the model's performance using the validation dataset. Common evaluation metrics include accuracy, precision, recall, F1-score, and mean squared error (depending on the type of problem).

- **Model Testing:**

  Evaluate the final model on the testing dataset to assess its real-world performance. This step provides an estimate of how well the model will perform on new, unseen data.

- **Deployment:**

  Deploy the trained model in a production environment where it can make real-time predictions. This may involve creating APIs, integrating with databases, or embedding the model in software applications.

Chapter 4

Implementation

- **Step 1:Importing the necessary libraries**

  ```
  import numpy as np #numerical python
  import pandas as pd #loading,processing and analysis of data
  import matplotlib.pyplot as plt #graphical visulization
  import seaborn as sns #graphical visulization with corr,trends,patterns

  import warnings
  warnings.filterwarnings(’ignore’)
  ```

- **Step 2:Load the Dataset¶**

  ```
  df=pd.read_csv(”diabetes.csv”)
  df.head()
  ```

- **Step 3:Exploratory Data Analysis(EDA)**

  Data checking

  ```
  df.shape
  df.size
  df.dtypes
  ```

  Data Cleaning

  ```
  #Check and drop Duplicates
  df=df.drop_duplicates()
  df.shape
  ```

```
#Check for missing values
df.isnull().sum()

#replacing 0 values in dataset
df[’Glucose’]=df[’Glucose’].replace(0,df[’Glucose’].mean())
df[’BloodPressure’]=df[’BloodPressure’].replace(0,df[’BloodPressure’].mean())
df[’SkinThickness’]=df[’SkinThickness’].replace(0,df[’SkinThickness’].mean())
df[’Insulin’]=df[’Insulin’].replace(0,df[’Insulin’].mean())
df[’BMI’]=df[’BMI’].replace(0,df[’BMI’].mean())
```

- **Step 4:Data Visualization**

```
value_counts=df[’Outcome’].value_counts() #series
plt.pie(value_counts, labels=value_counts.index, autopct=’plt.axis(’equal’)
plt.title(’Distribution of Diabetes (1: Yes, 0: No)’)
plt.show()

#data is imbalanced
#34.9%= 268 people have diabetes
#65.1% =  500 people have no diabetes

sns.scatterplot(data=df,x=’Age’,y=’BloodPressure’,hue=’Outcome’)

corr_matrix=df.corr()
sns.heatmap(corr_matrix, annot=True,cmap=’terrain’,linewidths=0.1)
```

From above it is clear that features such as BloodPressure and DiabetesPedigreeFunction contribute less to the values of the Outcome label.Hence we can remove them from dataset¶

```
df.drop([’BloodPressure’,’DiabetesPedigreeFunction’],axis=’columns’,inplace=True)
```

- **Step 5:Splitting Data into X and Y (predictors and outcome)**

```
X=df.drop('Outcome',axis=1)
print(X.head())
print(X.shape)

Y=df['Outcome']
Y.shape
```

- **Step 6:Feature Scaling**

  ```
  from sklearn.preprocessing import StandardScaler scaler1=StandardScaler()
  X=df.drop('Outcome', axis=1) # Select features without the target variable
  X_scaled=scaler1.fit_transform(X)

  X=pd.DataFrame(X_scaled, columns=X.columns)
  X.head()
  ```

- **Step 7:Train Test Split**

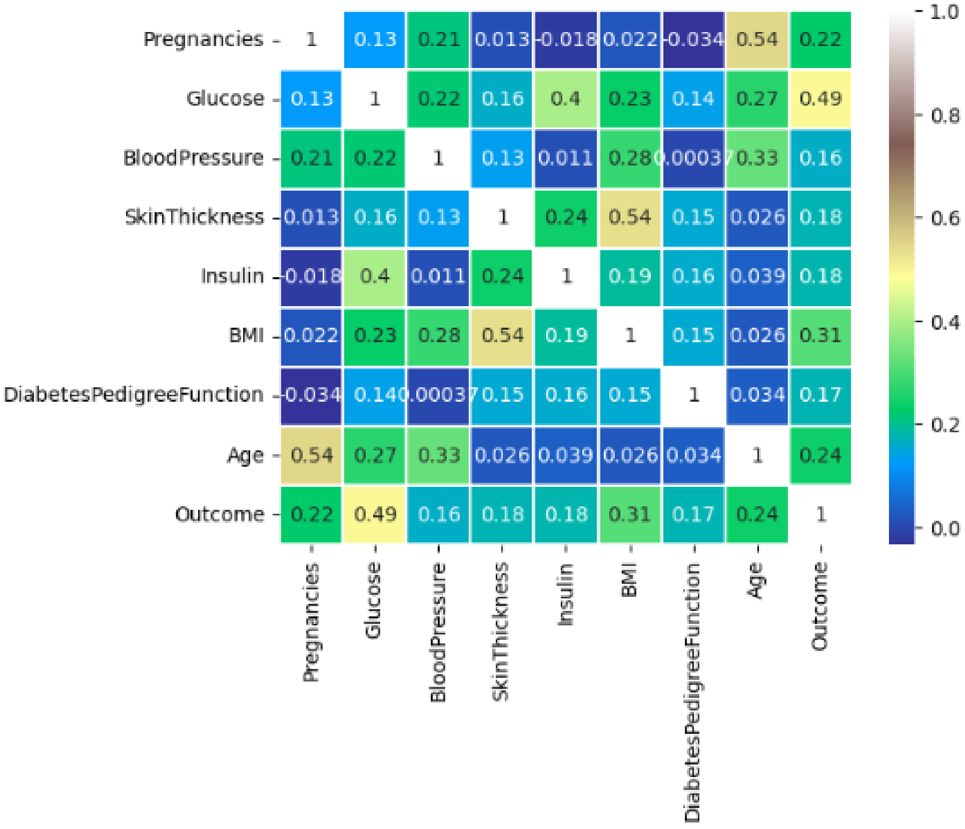
  ```
  from sklearn.model_selection import train_test_split
  x_train,x_test,y_train,y_test=train_test_split(X,Y,test_size=0.31,random_state=31)
  ```

- **Step 8:Building the Classification Algorithm**
  - **Logistic Regression**
  - **KNN Algorithm**
  - **Naive Bayes Algorithm**
  - **Decision Tree**
  - **Random Forest**
- **Step 9:Comparision of Classification Algorithms**

  From the plot,it is observed that the KNN Algorithm is the best classification algorithm that can be used to predict the presence or absence of diabetes on the basis of this dataset.¶

Chapter 5

5.1 Results & Snapshots



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1	0.13	0.21	0.013	-0.018	0.022	-0.034	0.54	0.22
Glucose	0.13	1	0.22	0.16	0.4	0.23	0.14	0.27	0.49
BloodPressure	0.21	0.22	1	0.13	0.011	0.28	0.00037	0.33	0.16
SkinThickness	0.013	0.16	0.13	1	0.24	0.54	0.15	0.026	0.18
Insulin	-0.018	0.4	0.011	0.24	1	0.19	0.16	0.039	0.18
BMI	0.022	0.23	0.28	0.54	0.19	1	0.15	0.026	0.31
DiabetesPedigreeFunction	-0.034	0.14	0.00037	0.15	0.16	0.15	1	0.034	0.17
Age	0.54	0.27	0.33	0.026	0.039	0.026	0.034	1	0.24
Outcome	0.22	0.49	0.16	0.18	0.18	0.31	0.17	0.24	1

Figure 5.1: Correlation between predictors and outcome

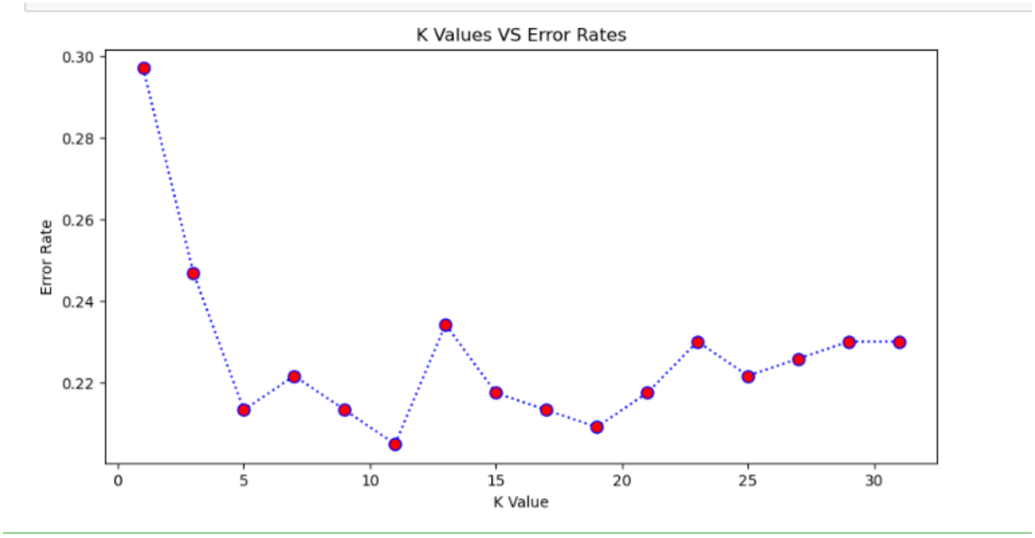


Figure 5.2: Comparision of K vs Error_rate

Figure 5.3: This is the startup and final screens of Snake and Ladder application

?? Choosing the Best Algorithms

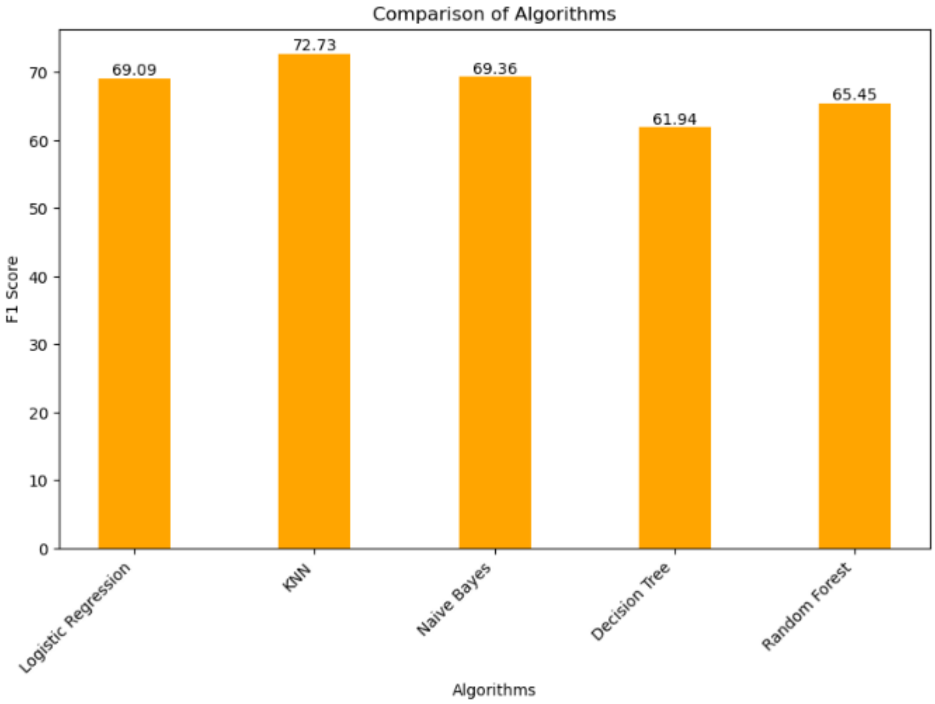
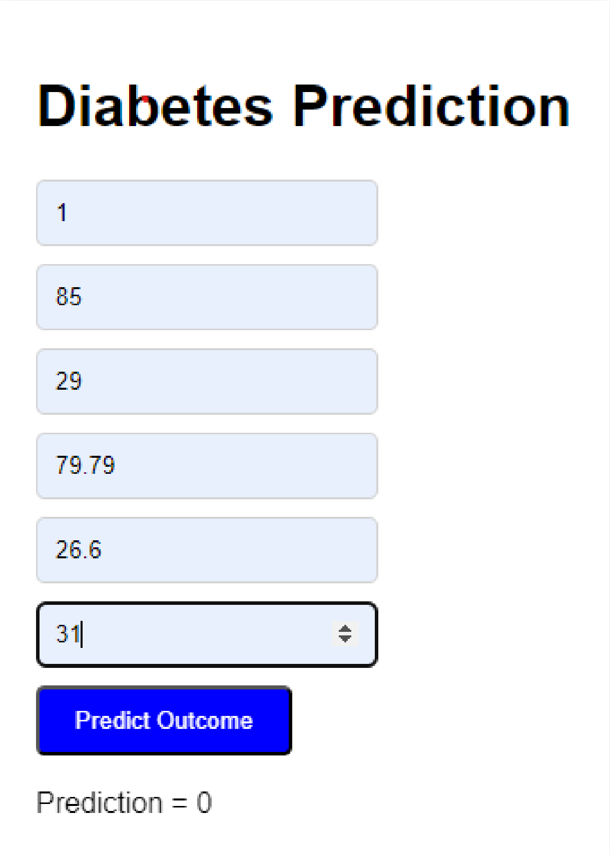


Figure 5.4: Comparision of Algorithms



Diabetes Prediction

1

85

29

79.79

26.6

31

Prediction = 0

Figure 5.5: Prediction of values

Chapter 6

Conclusion & Future Enhancements

6.1 Conclusion

The deployment of a diabetes prediction system using Machine Learning (ML) and Flask represents a significant advancement in healthcare. Such systems provide valuable tools for early risk assessment and personalized intervention, contributing to improved patient care and public health. In conclusion, the deployment of a diabetes prediction system using Flask has several key takeaways:

- Early Detection
- Accuracy
- Interpretability

6.2 Future Enhancements

While the deployment of a diabetes prediction system using Flask represents a significant achievement, there are several avenues for future enhancements and improvements:

- Continuous Model Updates
- Long-term Predictions
- Mobile Applications

References

[1] Edward Angel, *“Interactive Computer Graphics A Top-Down Approach With OpenGL”* 5th Edition, Addison-Wesley, 2008.

[2] F.S. Hill,*“Computer Graphics Using OpenGL”*, 2nd Edition, Pearson Education, 2001.

[3] James D.Foley, Andries Van Dam, Steven K. Feiner, John F Hughes, *”Computer Graphics”*, Second Edition, Addison-Wesley Professional, August 14,1995.

[4] @online OpenGL Official ,`https://www.opengl.org/`

[5] @online OpenGL Overview , `https://https://www.khronos.org/opengl/`

6.3