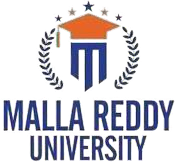
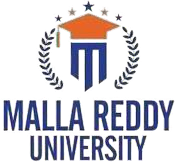
DATA ANAL YTICS WITH PYTHON



# Department of Computer Science &Engineering

## MALLA REDDY UNIVERSITY, HYDERABAD

**2023-2024**



# DIABETES PREDICITON

**Designed and Developed by**

|  |  |
| --- | --- |
| Y Dheeraj Reddy -2111CS010122 |  |
| G Gowtham -2111CS010157 |
| P Dinesh -2111CS010126 |
| K Chandra Prakash -2111CS010104 |
| N Giridhar -2111CS010148 |

**Guided by**

**Mr.D.I.P.MANIKUMAR**

# Department of Computer Science &Engineering

## MALLA REDDY UNIVERSITY, HYDERABAD

**2023-2024**



**CERTIFICATE**

This is to certify that this is the Data Analytics with Python lab report entitled “**DIABETES PREDICTION**” , submitted by **Y DHEERAJ REDDY- 2111CS010122,G GOWTHAM-2111CS010157,N GIRIDHAR- 2111CS010148,P DINESH-2111CS010126,K CHANDRA PRAKASH -2111CS010104** B. Tech **III** year II semester, Department of CSE during the year 2022-23. The results embodied in this report have not been submitted to any other university or institute for the award of any degree or diploma.

**GUIDE**

**Mr.D.I.P.MANIKUMAR**



# DECLARATION

I declare that this project report titled **DIABETES PREDICTION** submitted in partial fulfillment of the degree of B. Tech in CSE is a record of original work carried out by me under the supervision of **MR.D.I.P.MANIKUMAR** and has not formed the basis for the award of any other degree or diploma, in this or any other Institution or University. In keeping with the ethical practice in reporting scientific information, due acknowledgements have been made wherever the findings of others have been cited.

|  |
| --- |
|  |

|  |
| --- |
| Y Dheeraj Reddy -2111CS010122 |
| G Gowtham -2111CS010157 |
| P Dinesh -2111CS010126 |
| K Chandra Prakash -2111CS010104 |
| N Giridhar -2111CS010148 |

# ACKNOWLEDGEMENT

We would like to express my deepest gratitude to my guide, for her valuable guidance, consistent encouragement, personal caring, timely help and providing us with an excellent atmosphere for doing research. All through the work, in spite of her busy schedule, she has extended cheerful and cordial support to us for completing this research work. She has been helpful at every step from beginning till the end of this project. We are thankful to her for the encouragement she had given us in completing the project. We express our heart full thanks to the Head of the Department, Department of Computer Science and Engineering, for all the help and infrastructure provided to us to complete the project successfully and her valuable guidance. We are also thankful to all other faculty and staff members of our department for their kind co- operation and help.

|  |
| --- |
| Y Dheeraj Reddy -2111CS010122 |
| G Gowtham -2111CS010157 |
| P Dinesh -2111CS010126 |
| K Chandra Prakash -2111CS010104 |
| N Giridhar -2111CS010148 |

# ABSTRACT

Diabetes is a big health problem worldwide. It's important to predict who might get it early, so we can help them better. Computers are getting really good at this prediction job. They use lots of different info like medical records, genes, lifestyle, and environment.In this review, we talk about how computers are learning to predict diabetes. They use special tricks called algorithms. Some are like decision trees, others like random forests or neural networks. These algorithms learn from lots of examples to make better predictions.We also look at how computers pick the best info to use and make sure they're making good guesses. They even use new stuff like wearable gadgets and electronic records to get better at predicting.But there are challenges, too. Sometimes the info isn't good enough, or it's hard to understand why the computer made a certain prediction. Fixing these problems will make the predictions even better.

**TOOLS USED:** python, Jupyter notebook, pandas, numpy, streamlit

,

|  |  |
| --- | --- |
| **TABLE OF CONTENT** | |
| **DESCRIPTION** | **PAGE NUMBER** |
| **CERTIFICATE** | **iii** |
| **DECLARATION** | **iv** |
| **ACKNOWLEDGEMENTS** | **v** |
| **ABSTRACT** | **vi** |
|  |  |
| **LIST OF TABLES** |  |
| **Chapter 1 Introduction** |  |
| **1.1Introduction** | **1** |
| **1.2Problem Statement** | **2** |
| **1.3Objective of Project** | **3** |
| **1.4Goal of Project** | **5** |
| **Chapter 2 Problem Identification** |  |
| **2.1Existing System** | **6** |
| **2.2Proposed System** | **6** |
| **Chapter 3 Requirements** |  |
| **3.1 Software Requirements** | **7** |
| **3.2 Hardware Requirements** | **8** |
|  |  |
|  |  |
|  |  |

|  |  |
| --- | --- |
| **Chapter 4 Code** |  |
| **4.1Source Code** | **13** |
| **4.2Screenshot of Application** | **21** |
| **Chapter 5 Results & Conclusion** |  |
| **5.1Results**  **5.2Conclusion** | **22** |
| **REFERENCES** | **23** |
|  |  |
|  |  |
|  |  |

* 1. **INTRODUCTION**

# CHAPTER-1 INTRODUCTION

The Diabetes Prediction Project represents a proactive approach in tackling the global burden of diabetes mellitus, a chronic metabolic disorder characterized by elevated blood sugar levels. With its increasing prevalence and significant impact on public health, early detection and intervention are crucial for effective management and prevention of complications.

This project aims to harness the power of data-driven predictive modeling techniques, particularly machine learning, to identify individuals at risk of developing diabetes before clinical symptoms manifest. By analyzing diverse datasets encompassing demographic information, medical history, lifestyle factors, genetic predisposition, and environmental influences, the project seeks to construct robust predictive models capable of accurately stratifying individuals based on their likelihood of developing diabetes.

The utilization of advanced machine learning algorithms enables the project to leverage complex interactions among various risk factors, thereby enhancing the accuracy and precision of predictions. Additionally, the integration of emerging technologies such as wearable sensors and electronic health records expands the scope and depth of available data, enabling more comprehensive risk assessments.

Through this project, we aim to empower healthcare providers with actionable insights to implement targeted interventions, lifestyle modifications, and personalized management strategies for individuals identified as high-risk for diabetes. Moreover, by facilitating early detection and intervention, the project aims to reduce the incidence of diabetes-related complications, improve patient outcomes, and alleviate the socioeconomic burden associated with this chronic condition.

## PROBLEM STATEMENT

## Despite significant advancements in healthcare, diabetes mellitus remains a major global health concern, with a rising prevalence and substantial socioeconomic burden. Early detection of individuals at risk of developing diabetes is critical for implementing timely interventions and personalized management strategies to mitigate its adverse effects and prevent complications.

## However, traditional risk assessment methods often rely on limited clinical parameters and lack the ability to capture the complex interplay of diverse risk factors associated with diabetes onset. Moreover, the existing approaches may not adequately leverage the wealth of available data sources, including demographic information, medical history, lifestyle behaviors, genetic predisposition, and environmental factors.

## Therefore, the problem statement for diabetes prediction revolves around the need to develop accurate, scalable, and interpretable predictive models capable of identifying individuals at high risk of developing diabetes before the onset of clinical symptoms. These models should leverage advanced machine learning techniques to analyze diverse datasets comprehensively and extract meaningful patterns and associations among various risk factors.

## OBJECTIVE OF PROJECT

Objectives of Diabetes Prediction:

**Early Detection:** The primary objective of diabetes prediction is to identify individuals at risk of developing diabetes before the onset of clinical symptoms. Early detection allows for timely interventions and lifestyle modifications to prevent or delay the progression of the disease.

**Risk Stratification:** Another objective is to stratify individuals based on their likelihood of developing diabetes, enabling healthcare providers to prioritize interventions for those at the highest risk. This risk stratification facilitates targeted and personalized management strategies.

**Accuracy and Reliability:** The predictive models developed for diabetes prediction aim to achieve high accuracy and reliability in identifying at-risk individuals. This involves leveraging advanced machine learning algorithms and comprehensive datasets to capture the complex interactions among various risk factors.

**Integration of Multifactorial Data:** The integration of diverse data sources, including demographic information, medical history, genetic predisposition, lifestyle factors, and environmental influences, is crucial for developing comprehensive predictive models. The objective is to capture the multifactorial nature of diabetes risk and improve prediction accuracy.

**Generalizability:** Diabetes prediction models should be generalizable across different populations and settings to ensure their applicability and effectiveness in diverse healthcare contexts. Achieving generalizability requires robust model development methodologies and validation across various demographic groups.

**Interpretability and Explainability:** Ensuring the interpretability and explainability of predictive models is essential for gaining the trust of healthcare providers and patients. The objective is to develop models that provide insights into the underlying factors contributing to diabetes risk, facilitating informed decision-making.

**Translation into Clinical Practice:** Ultimately, the objective of diabetes prediction is to translate predictive models into clinical practice, enabling healthcare providers to integrate risk assessment into routine care. This involves overcoming barriers related to data privacy, model deployment, and integration with existing healthcare systems.

**Prevention of Complications:** By accurately identifying individuals at risk of developing diabetes, the objective is to prevent or delay the onset of complications associated with the disease. Early intervention and lifestyle modifications can help reduce the incidence of diabetes-related complications and improve overall health outcomes.

In summary, the objectives of diabetes prediction encompass early detection, risk stratification, accuracy, integration of multifactorial data, generalizability, interpretability, translation into clinical practice, and prevention of complications. Achieving these objectives requires collaborative efforts among researchers, healthcare providers, policymakers, and technology experts.

## GOAL OF PROJECT

Goal of Diabetes Prediction Project:

The main aim of the Diabetes Prediction Project is to make accurate guesses about who might get diabetes before they even show signs of it. This helps doctors to start helping them earlier, maybe even before they get sick.

Here are the main goals of the project:

**Find People Early:** The project wants to find people who might get diabetes early on, so they can start getting help sooner. This helps to prevent diabetes or at least slow it down.

**Personalized Help:** It wants to give people personalized help based on their own risks. Everyone is different, so the project wants to make sure the help each person gets fits them best.

**Sort People by Risk:** The project wants to group people into different categories based on how likely they are to get diabetes. This helps doctors know who needs the most help first.

**Use Lots of Information:** By looking at a lot of different information like age, family history, lifestyle, and genes, the project aims to make better guesses about who might get diabetes.

**Help Doctors Help People:** Ultimately, the project wants to give doctors tools they can use to spot diabetes risk in their patients early on. This way, they can start treatments and lifestyle changes to prevent diabetes problems.

**Less Diabetes Problems:** By finding and helping people early, the project hopes to stop diabetes from causing big problems like heart issues or eye damage.

In simple terms, the Diabetes Prediction Project wants to guess who might get diabetes early and give them the right help to prevent it or make it less severe. It's like giving people a heads-up so they can stay healthy.

# CHAPTER-2 PROBLEM IDENTIFICATION

## EXISTING SYSTEM

**Reliance on Basic Risk Factors:** Current methods primarily depend on common risk factors such as family history, age, weight, and blood sugar levels to predict diabetes risk.

**Diagnostic Tests after Symptom Appearance:** Healthcare providers often use blood tests like fasting blood sugar levels or oral glucose tolerance tests to diagnose diabetes only after symptoms appear, rather than predicting it beforehand.

**Limited Precision:** These traditional methods may not accurately identify everyone who might develop diabetes, and they might not provide precise risk estimates for individuals.

**Opportunity for Improvement:** There's a need to enhance prediction accuracy and early detection by incorporating advanced technology and comprehensive data analysis techniques, which projects like the Diabetes Prediction Project aim to address.

## PROPOSED SYSTEM

The proposed system for Diabetes prediction involves integrating cutting-edge technologies and methodologies to enhance accuracy, efficiency and patient outcomes. Some key components of the proposed system include:

**Advanced Machine Learning Techniques:** The proposed system will leverage advanced machine learning algorithms to analyze comprehensive datasets containing demographic, clinical, genetic, and lifestyle information for more accurate and personalized diabetes risk prediction.

**Integration of Multifactorial Data:** It will incorporate diverse data sources, including wearable sensors, electronic health records, and genomic data, to capture the complex interactions among various risk factors contributing to diabetes onset.

**Early Detection and Prevention:** By identifying individuals at high risk of developing diabetes before symptoms appear, the proposed system aims to enable early interventions and lifestyle modifications to prevent or delay the onset of the disease.

**Translation into Clinical Practice:** The system will be designed for seamless integration into routine clinical practice, providing healthcare providers with actionable insights to facilitate personalized management strategies and improve patient outcomes.

# CHAPTER-3 REQUIREMENTS

## SOFTWARE REQUIREMENTS

To develop a Diabetes prediction system using machine learning, you'll need a combination of software tools and libraries for various tasks such as data preprocessing, model training, and evaluation. Here's a list of essential software requirements:

1. **Programming Language:**

* Python: A versatile and widely used programming language for machine learning, widely used in the machine learning community, with extensive libraries and frameworks.

1. **Integrated Development Environment (IDE):**

* Jupyter Notebooks: Great for interactive development and data exploration.

1. **Data Manipulation and Analysis:**

* Pandas: For data manipulation and analysis.

1. **Data Visualization:**

* Matplotlib and Seaborn: For creating visualizations of the data.

1. **Machine Learning Libraries:**

* Scikit-learn: Provides simple and efficient tools for data mining and data analysis.

## HARDWARE REQUIREMENTS

### CPU (Central Processing Unit):

A multi-core CPU with good clock speed is essential for preprocessing data and training machine learning models. An Intel i5 or i7 processor, or an equivalent AMD processor, isa good starting point.

### RAM (Random Access Memory):

Adequate RAM is crucial, especially when working with large datasets and complex models. A minimum of 8 GB is recommended, but 16 GB or more is preferable for more extensive projects.

### GPU (Graphics Processing Unit):

For faster training of deep learning models, especially if using frameworks like TensorFlow or PyTorch, having a GPU is highly beneficial. NVIDIA GPUs, such as GeForce GTX or RTX series, or professional GPUs like NVIDIA Quadro or Tesla, are commonly used for machine learning tasks.

### Storage:

A solid-state drive (SSD) is recommended for faster data access during preprocessing and training. Ensure sufficient storage capacity for your datasets and model files. If working with large datasets, consider having multiple drives or using cloud storage solutions.

### Internet Connection:

A reliable internet connection is essential for downloading datasets, model libraries, and updates. Additionally, if you plan to use cloud-based services for training or deployment, a stable internet connection is crucial.

### Motherboard:

Ensure compatibility with the selected CPU and GPU. A motherboard with multiple Pistols may be beneficial if considering multiple GPUs for parallel processing.

### Power Supply:

Adequate power supply is essential, especially if using a dedicated GPU, to ensure stable and consistent performance.

### Cooling System:

Depending on the intensity of your machine learning tasks, consider additional cooling solutions to prevent overheating, especially if using a high-performance GPU.

### Backup System:

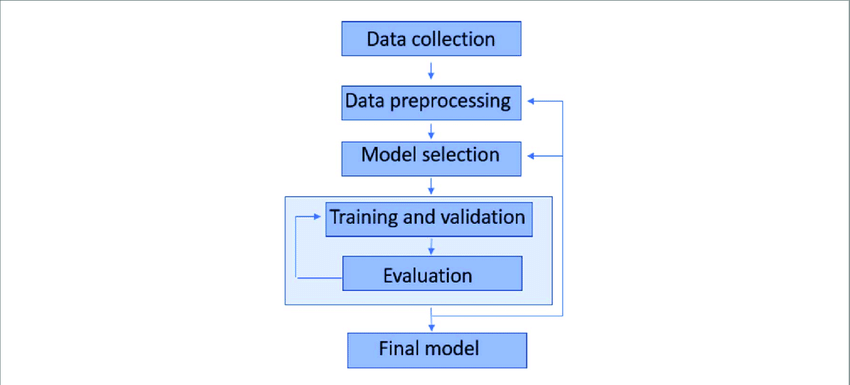
Regularly backup your work to prevent data loss, especially if working with valuable datasets.

CHAPTER-4

FLOW CHART

* **Data Collection:** It is about understanding the requirements needed for the project like datasets ,different ML algorithms**.**
* **Data preprocessing:** it is analyzing of the different features present in dataset and their importance.
* **Model selection:** It is integration of data taken by removing excessive features and data cleaning.
* **Training:** Multiple machine learning models were used to predict the popularity of song.
* **Evaluation:** Models developed were evaluated based on their performance and accuracy.
* **Final model:** Once the models were evaluated they were integrated with code developed for user interface .

## Figure : Flow chart



**Algorithm Used**

**Decision Tree Classification Algorithm:**

**Tree-based Structure:** Decision trees create a hierarchical structure of decision rules, with each node representing a feature and each leaf node representing a class label, making it easy to interpret and visualize.

**Splitting Criteria:** The algorithm selects the best feature to split the data at each node based on criteria such as Gini impurity or information gain, recursively partitioning the dataset to create a predictive model.

**Handling Categorical and Numerical Data:** Decision trees can handle both categorical and numerical features. For categorical features, the algorithm evaluates each possible value as a potential split. For numerical features, it selects a threshold value to divide the data into two subsets.

**Applications:** Decision tree classification is widely used in various domains, including healthcare (e.g., disease diagnosis), finance (e.g., credit risk assessment), and marketing (e.g., customer segmentation). Its simplicity, interpretability, and ability to handle both categorical and numerical data make it a versatile choice for classification task

# CHAPTER – 4

**4.1 Source Code**

# CODE

# 

# 

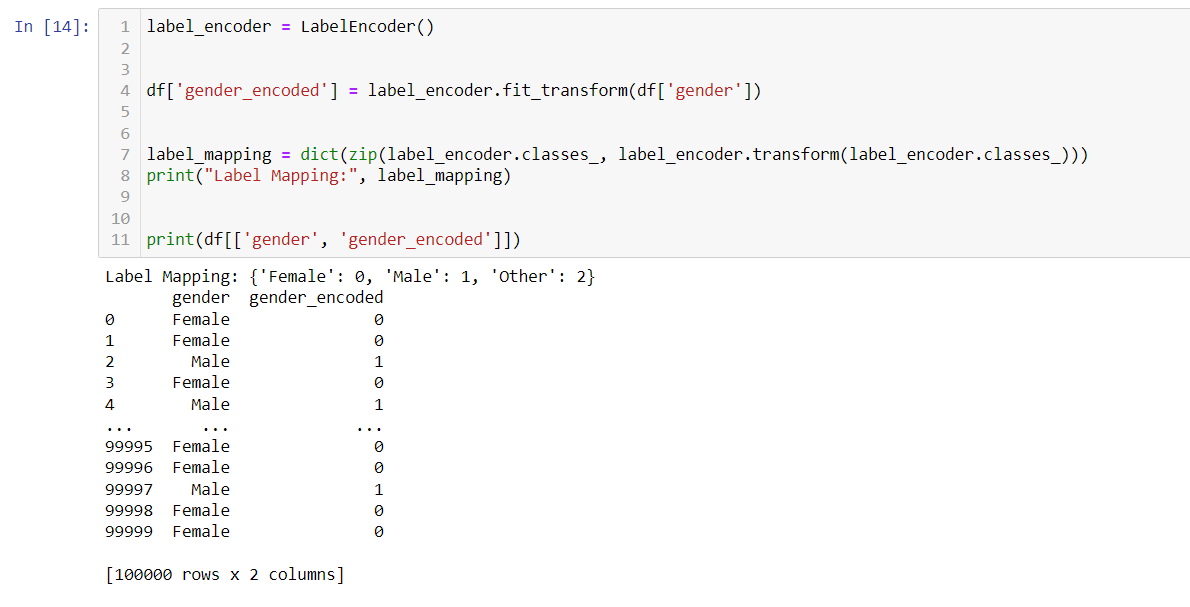
# 

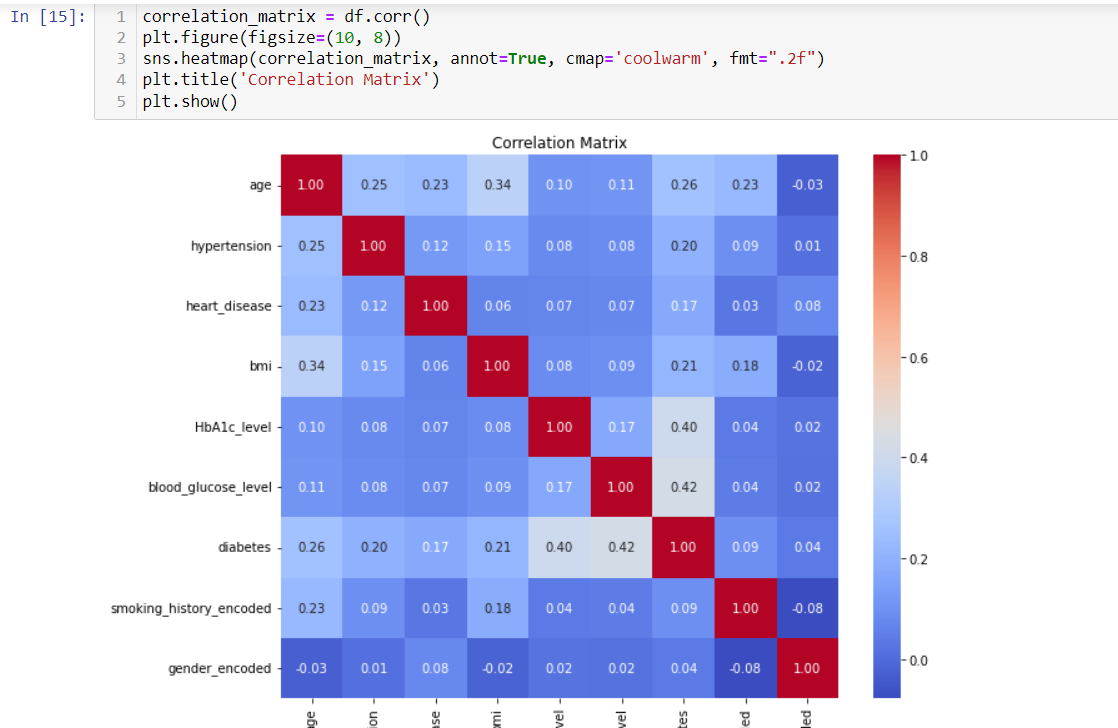
# 

# 

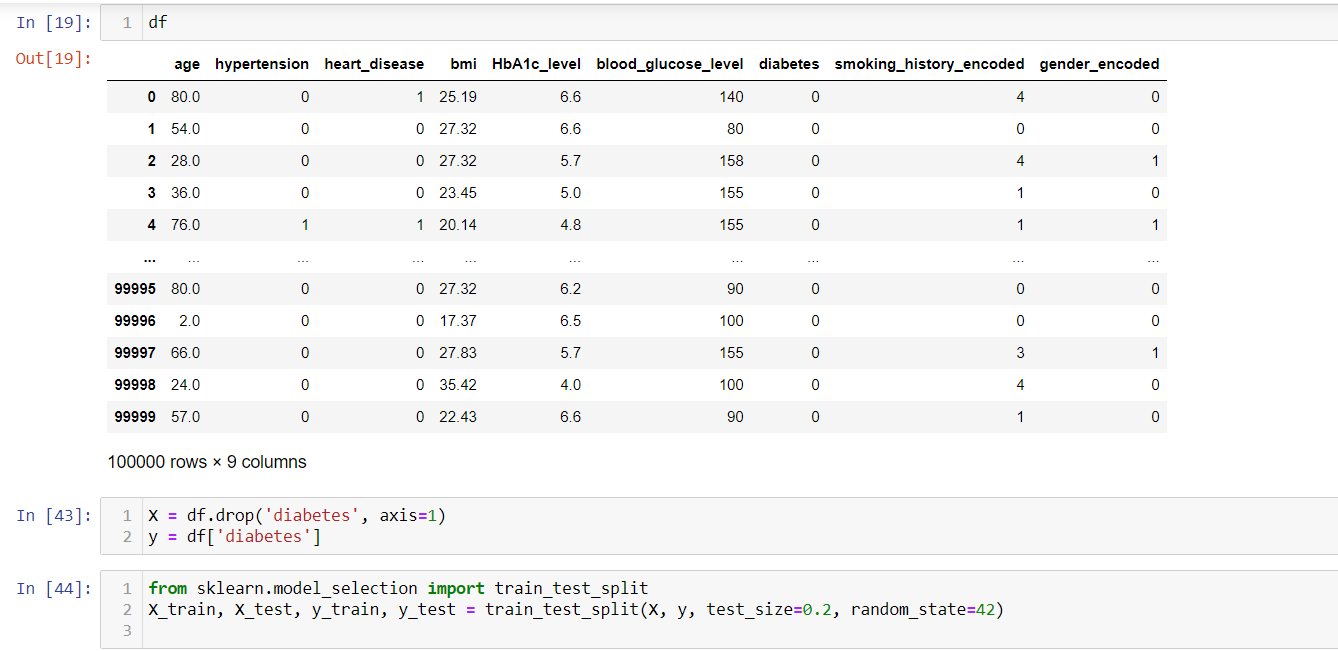
# 

# 

****

****

****

****

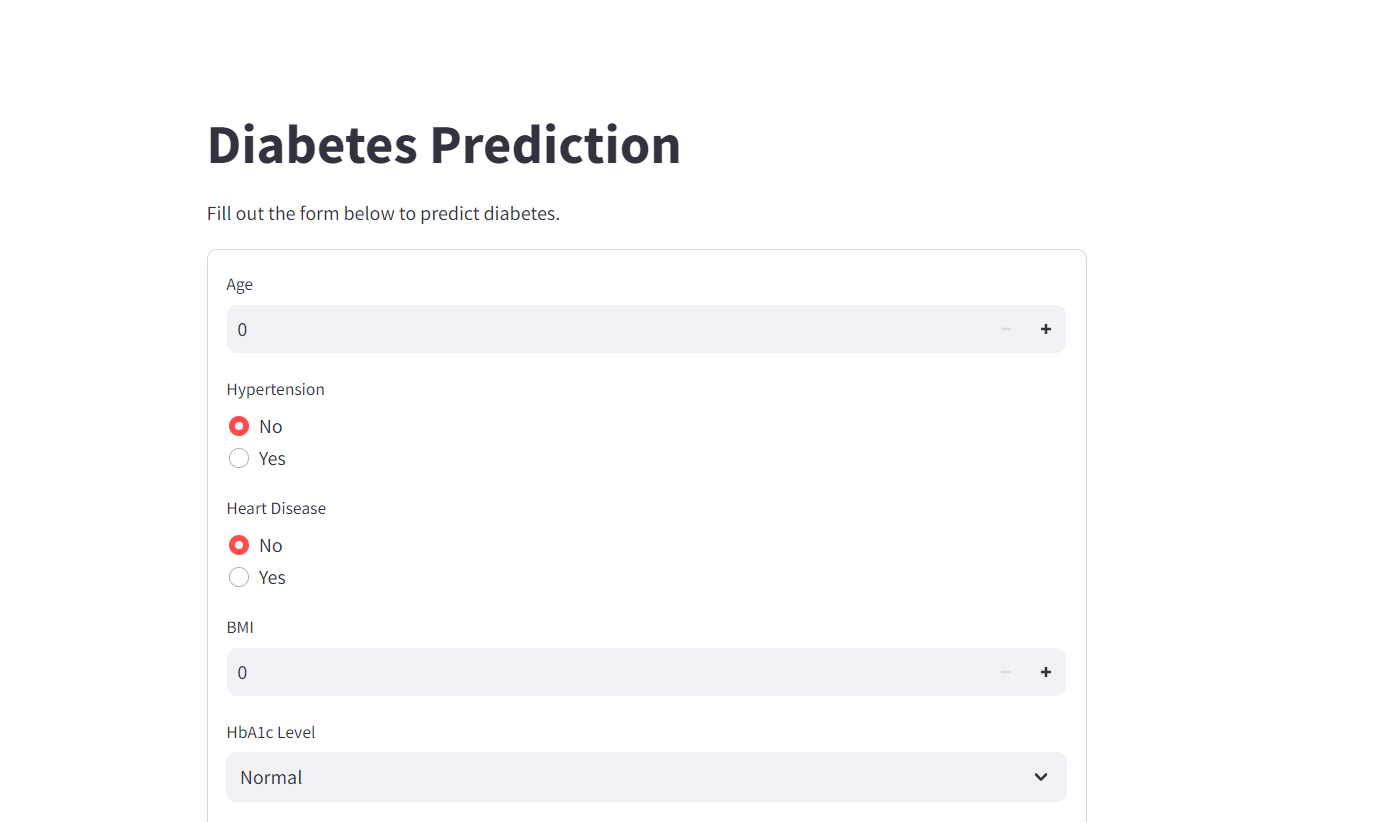
****

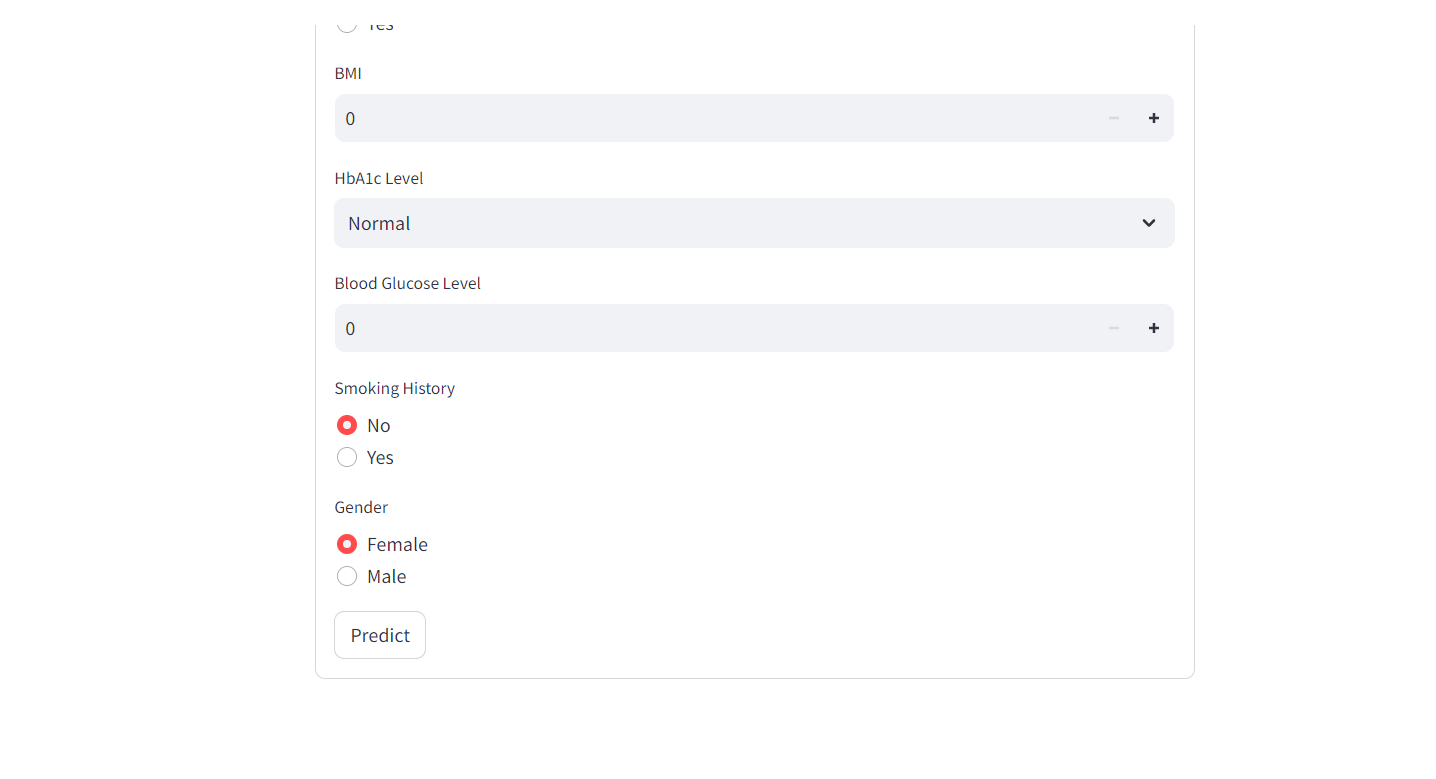
****

****

****

**Screen short of application**

****

****

**CHAPTER-5**

**RESULTS AND CONCLUSION**

## 5.1RESULTS

The result of a Healthcare prediction using machine learning relays on various factors, including the quality of the data, the features used for prediction, the chosen machine learning algorithm, and the overall model performance. The outcome will typically be a prediction i.e., the present condition of the patient, enhance healthcare delivery.

## CONCLUSION

In conclusion, the healthcare prediction model, built using a Random Forest, demonstrates its efficacy in predicting the condition of the patient based on key features .In conclusion, the healthcare prediction project represents a significant step forward in knowing the condition of the patient at the early stage. By leveraging advanced data analytics techniques, real-time decision-making capabilities, and personalized retention strategies.

# REFERENCES

* + 1. Abdi, F., & Abdollahzadehgan, A. (2018). Customer churn prediction in telecom using machine learning in big data platform. International Journal of Database Theory and Application, 11(10), 107-118.
    2. Fatima, I., Shafiq, M. Z., & Bashir, M. A. (2020). A Comprehensive Survey of Machine Learning Techniques for Telecommunication Churn Prediction. Journal of Network and Computer Applications, 169, 102765.
    3. Yin, P. Y., & Liao, S. H. (2010). Customer churn prediction by hybrid neural network and decision tree. Expert Systems with Applications, 37(1), 479-486.