Project Report

On

Diabetes Data Analysis and Prediction System

Submitted in partial fulfillment for the award of

Post Graduate Diploma in Big Data Analytics-(DBDA)

from

Know-IT (Pune)

Guided by:

Mr. Amey Manjrekar

Presented by:

Sarika Wandre	(220343025041)
Sayali Tandel	(220343025047)
Vinay Padwal	(220343025053)
Ajay Shewale	(220343025058)

Centre of Development of Advanced Computing (C-DAC), Pune

CERTIFICATE

TO WHOMSOEVER IT MAY CONCERN

This is to certify that

Ms. Sarika Wandre	**PRN: 220343025041**
Ms. Sayali Tandel	**PRN: 220343025047**
Mr. Vinay Padwal	**PRN: 220343025053**
Mr. Ajay Shewale	**PRN: 220343025058**

Have successfully completed their project on

Diabetes Data Analysis and Prediction System

Under the guidance of

Mr. Amey Manjrekar

Project Guide **Project Supervisor**

ACKNOWLEDGEMENT

This project "**Diabetes Prediction System**" was a great learning experience for us and we are submitting this work to CDAC Know-IT. We all are very glad to mention the name of **Mr. Amey Manjrekar** for his valuable guidance to work on this project. His guidance and support helped us to overcome various obstacles and intricacies during the course of project work.

We are highly grateful to **Mr. Vaibhav Inamdar** (Manager (Know-IT), C- DAC, for his guidance and support whenever necessary while doing this course Post Graduate Diploma in **Big Data Analytics (PG-DBDA)** through C-DAC ACTS, Pune.

From:

Sarika Wandre (220343025041)
Sayali Tandel (220343025047)
Vinay Padwal (220343025053)
Ajay Shewale (220343025058)

CONTENTS

ABSTRACT

Diabetes is considered as one of the deadliest and chronic diseases which causes an increase in blood sugar. Many complications occur if diabetes remains untreated and unidentified. The tedious identifying process results in visiting of a patient to a diagnostic Centre and consulting doctor.

It includes analyzing of data and predicting diabetes using attributes like Gender, Creatinine ratio, Cholesterol, Glucose, BMI etc.

In this project we will do prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques.

The project is processed using Python, Spark, MongoDB, Machine Learning, Power BI, Flask.

INTRODUCTION

Document Purpose

The purpose of this document is to build a system to analyze Diabetes Data for Diabetes prediction. The scope of this document is to define the functional and non-functional requirements, business rules and other constraints requirements.

Project Background

Main aim of this project is to predict the diabetes in order to make an event successful based on features like Cholesterol, Glucose, BMI, Age, HDL, Gender, Cholesterol/HDL Ratio etc.

Problem Statement

The research problem that this project try to address can be stated as follows:

How to develop a software platform to conduct descriptive, predictive, and prescriptive analysis of diverse Diabetes data.

Objectives

The main objective of this study is **to develop a machine learning (ML)-based system for predicting diabetic patients** based on features like Cholesterol, Glucose, BMI, Age, HDL, Gender, Cholesterol/HDL Ratio etc.

FUNCTIONAL REQUIREMENTS OVERVIEW

Modules used

1. Apache Spark
2. MongoDB
3. Python
4. Machine Learning

Apache Spark

- Apache Spark is an open-source big data processing engine, an in-memory, streaming-enabled. To perform stream processing, it handles the micro batching procedure by dividing the incoming stream of events into small batches and keeping the latency of stream processing under control.
- Therefore, it demands to be faster than Hadoop by achieving better performance due to its micro-batch processing. A strong point of using Apache Spark is its capacity to allow batch and streaming analysis in the same platform and its package streaming, which can process streaming data from different sources, including social media.
- Apache Spark includes Spark Streaming API that can read data from Apache Kafka and process data using difficult algorithms like a map, reduce, join, and window. Also, Apache Spark provides several interesting features, for example, iterative machine learning algorithms through the Mllib library, which gives efficient algorithms with the highest speed for streaming data analysis.

MongoDB

- MongoDB is a cross-platform, document oriented database that provides, high performance, high availability, and easy scalability. MongoDB works on concept of collection and document.

- Database
  Database is a physical container for collections. Each database gets its own set of files on the file system. A single MongoDB server typically has multiple databases.

- Collection
  Collection is a group of MongoDB documents. It is the equivalent of an RDBMS table. A collection exists within a single database. Collections do not enforce a schema. Documents within a collection can have different fields. Typically, all documents in a collection are of similar or related purpose.

- Document
  A document is a set of key-value pairs. Documents have dynamic schema. Dynamic schema means that documents in the same collection do not need to have the same set of fields or structure, and common fields in a collection's documents may hold different types of data.

Python

- Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language.
- Python supports multiple programming pattern, including object-oriented, imperative, and functional or procedural programming styles.
- Python has a large and broad library and provides rich set of module and functions for rapid application development.

Machine Learning

- Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed.
- Machine learning combines data with statistical tools to predict an output.
- Machine learning is also used for a variety of task like fraud detection, predictive maintenance, portfolio optimization, automatize task and so on.

SYSTEM REQUIREMENTS

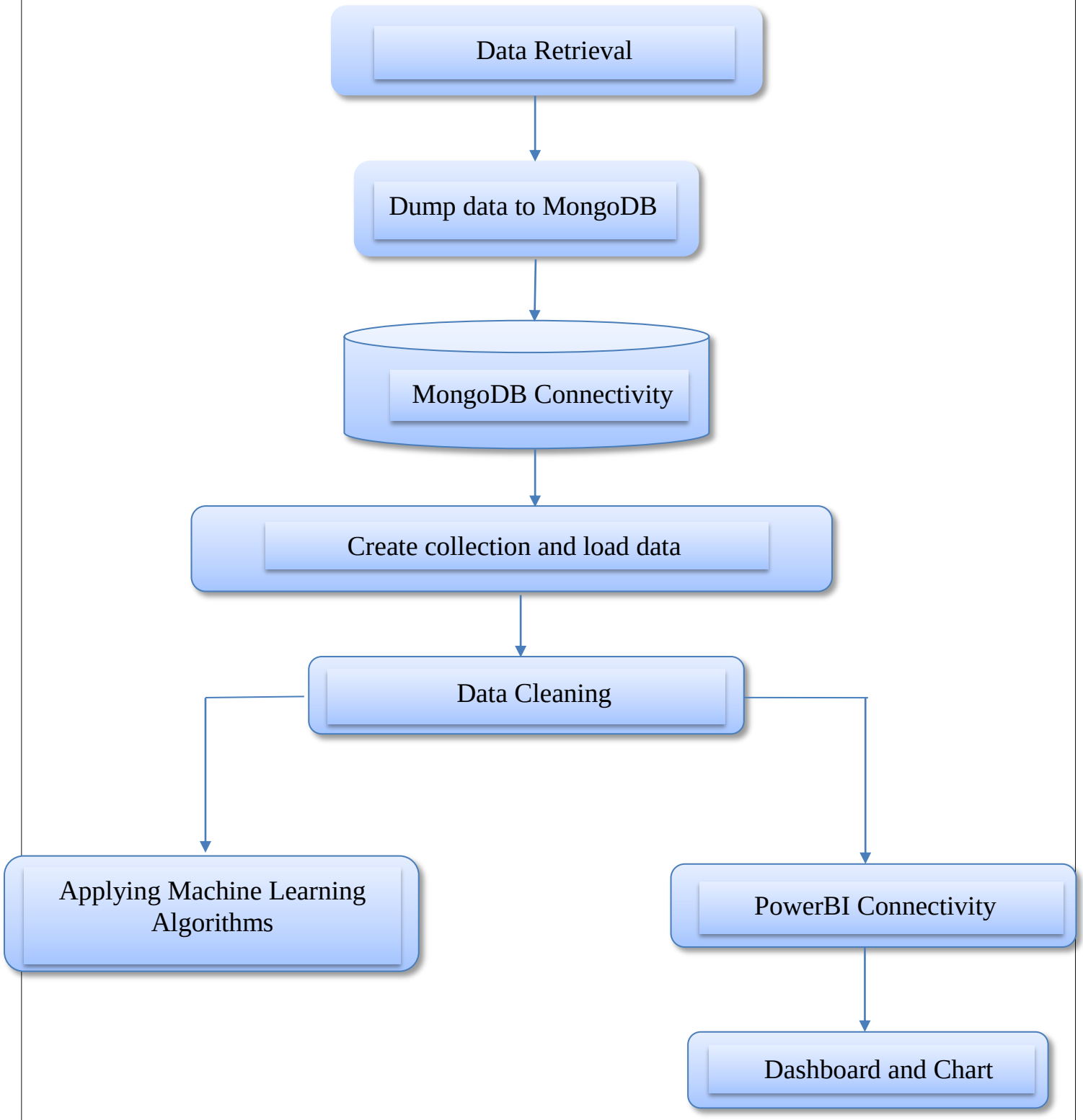
Hardware Requirements

- Platform – Windows
- RAM – At least 8 GB of RAM,
- Peripheral Devices – Mouse, Keyboard
- A network connection for data recovering over network.

Software Requirements

- Apache Spark
- MongoDB
- Python
- Machine Learning

PROJECT FLOW



METHODOLOGY

Data Used

The data used in this project was of dataworld.com open source Diabetes Data of 10MB size.

The data was in CSV format.

Data Cleaning

Data cleansing or data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from record set, tables or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

Data Dumping

Data was dumped into MongoDB using pyspark. The MongoDB was used for storing the large amount of data.

Analysis

The refined data is used to perform predictive analysis and draw appropriate conclusions from the performed data analysis.

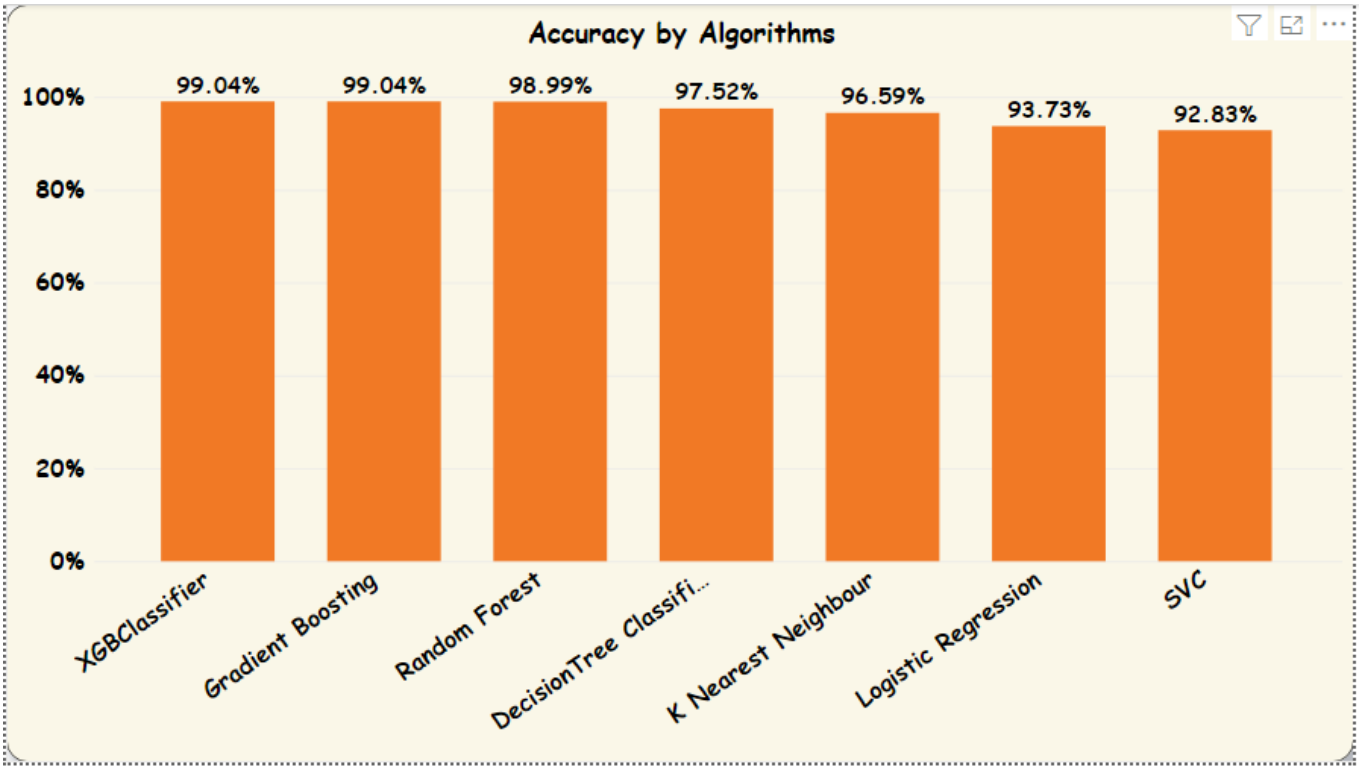
Models and Algorithms

Based on the data appropriate machine learning where applied on the data and models as solutions where created.

ALGORITHMS AND FORMULAS

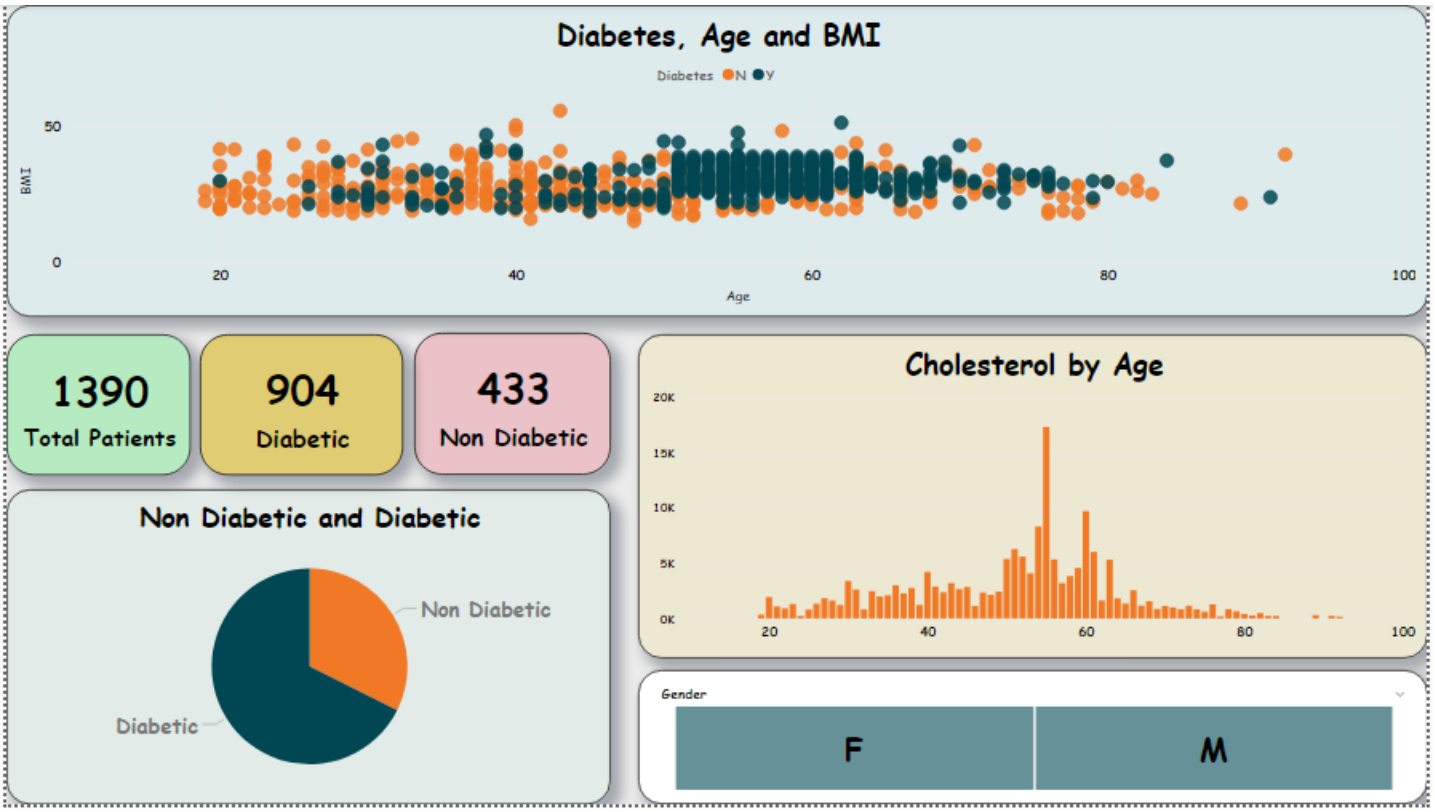
- Logistic Regression
- K-Nearest Neighbor
- Gradient Boosting
- XGBoost
- SVC
- Decision Tree Algorithm
- Random Forest Classifier

Accuracy By Algorithms:

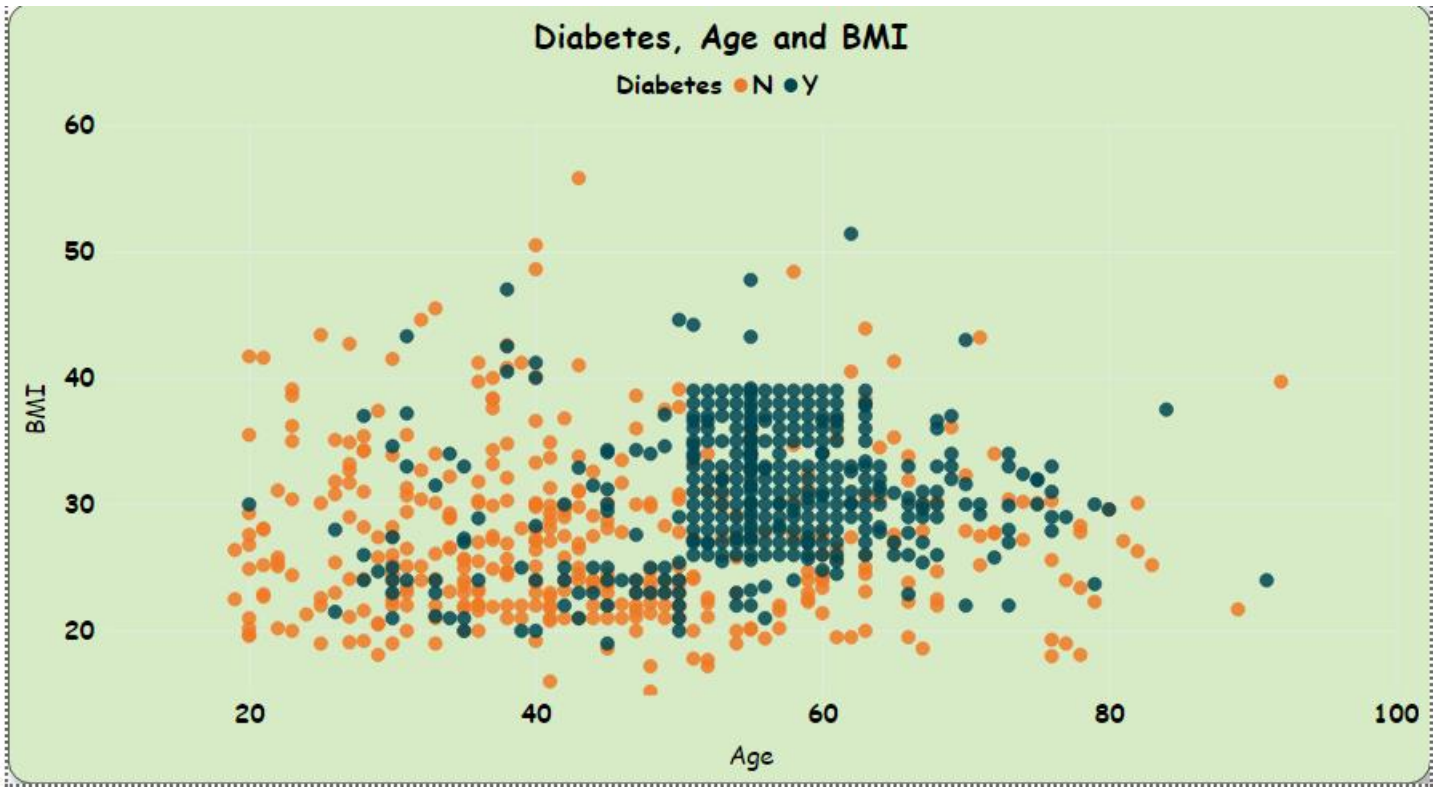


DATA VISUALIZATION

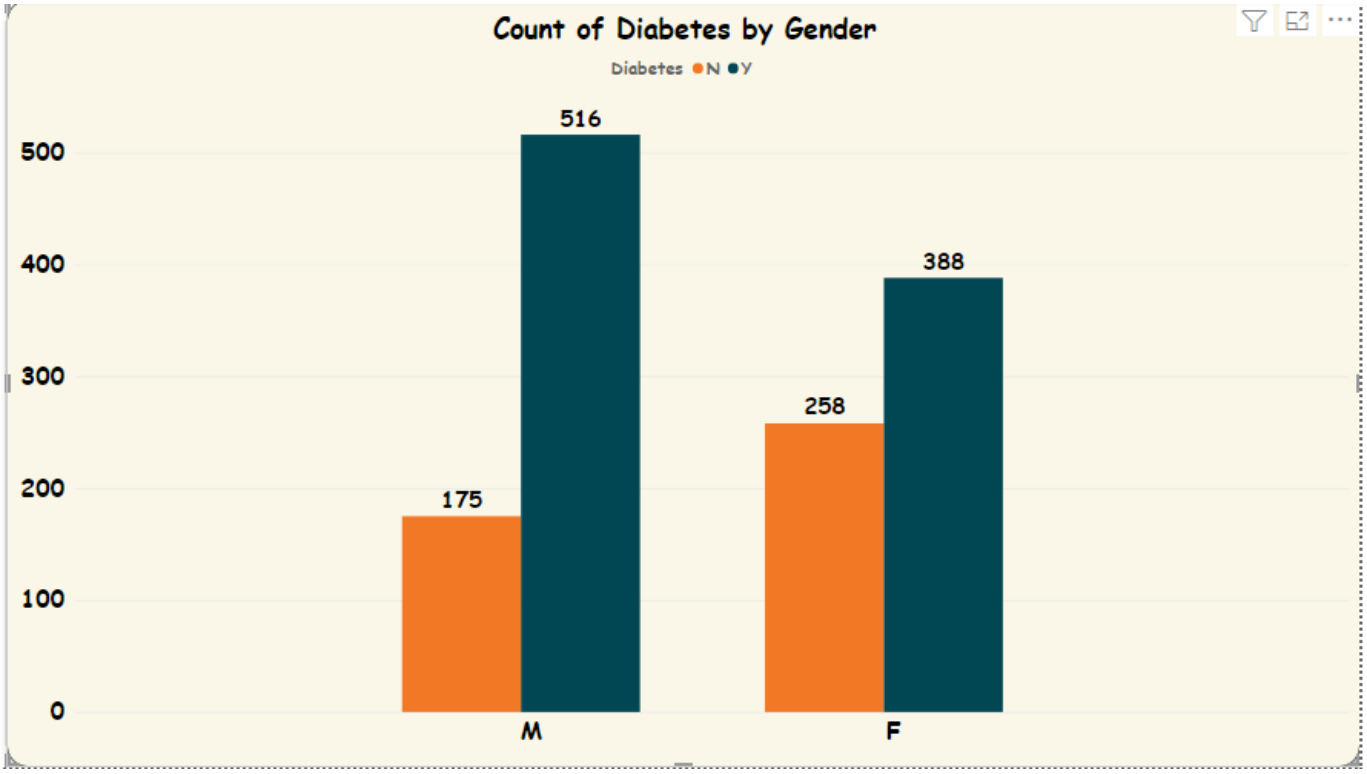
Dashboard:



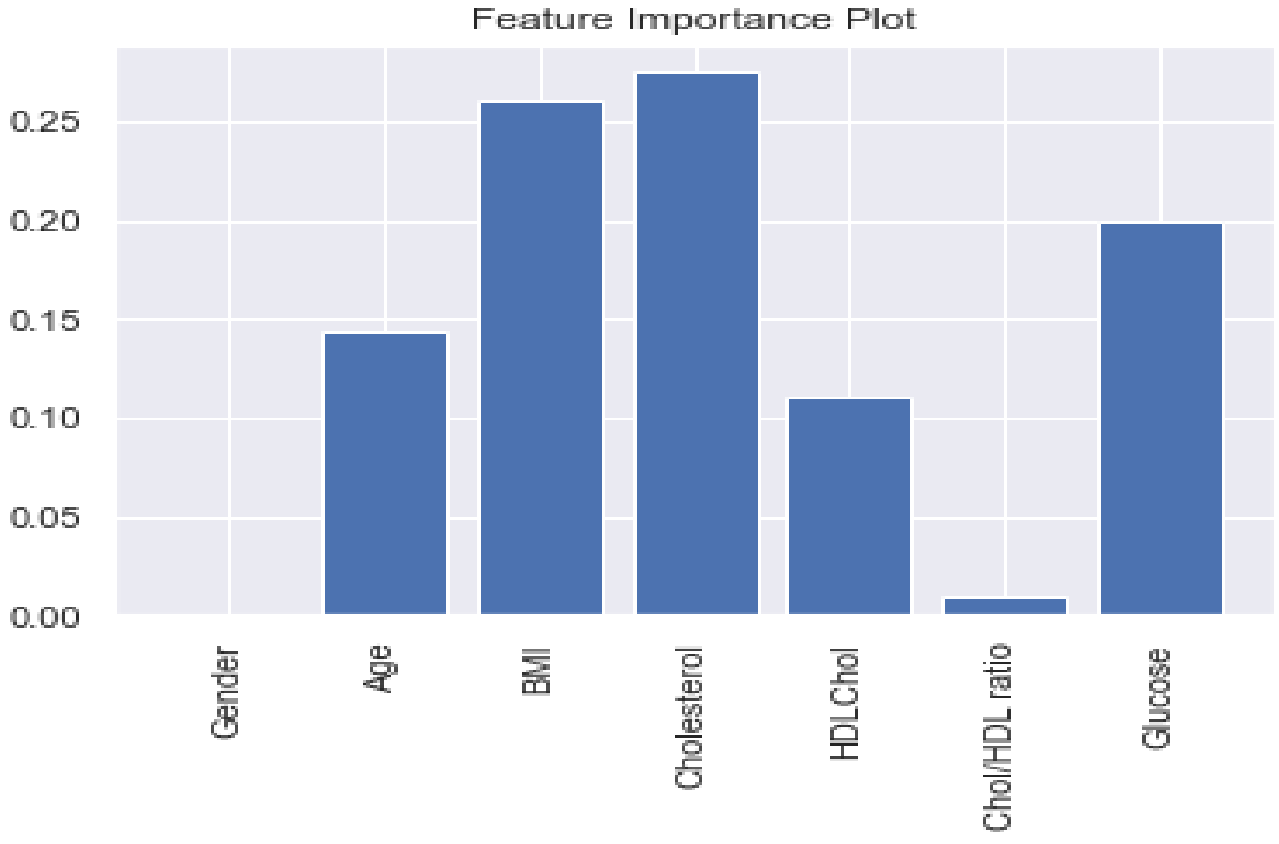
Scatter Plot:



Gender wise Count of Diabetes:



Feature Importance Plot :



CONCLUSION

After using all these patient records, we are able to build a machine learning model (XGBoost – best one) to accurately predict whether or not the patients in the dataset have diabetes or not along with that we were able to draw some insights from the data via data analysis and visualization.