# University of Mumbai 2020-2021

**M.H. SABOO SIDDIK COLLEGE OF ENGINEERING 8, Saboo Siddik Road, Byculla, Mumbai - 400 008**

**Department of Computer Engineering**

Project Report On

**“Diabetes Prediction System”**

By

**Faizaan Lakdawala**

**Nusrat Fatima Ansari**

**Mehvash Khan**

Guided By

**Prof. Saiqa Khan**



**M.H. SABOO SIDDIK COLLEGE OF ENGINEERING 8, Saboo Siddik Road, Byculla, Mumbai -400 008.**

This is to certify that,

|  |  |
| --- | --- |
| Roll No | Name |
| **3219067** | **Faizaan Lakdawala** |
| **3219069** | **Nusrat Fatima Ansari** |
| **3219070** | **Mehvash Khan** |

Of Third year (T.E) Semester VI degree course in Computer Engineering, have completed the specified project report on,

**“Diabetes Prediction System”**

As a partial fulfillment of the project work in a satisfactory manner as per the rules of the curriculum laid by the University of Mumbai, during the Academic Year July 2020 — June 2021.



Internal Guide



Internal Examiner External Examiner



Head of Department Principal

This project report entitled **Diabetes Prediction System** by **Faizaan Lakdawala, Nusrat Fatima Ansari and Mehvash Khan** is approved for the degree of Computer Engineering.

### EXAMINERS

**1.**

**2.**

**SUPERVISORS**

**1.**

**2.**

Date:

Place:

**ACKNOWLEDGEMENT**

We would like to express our gratitude and appreciation to our parents for motivating and encouraging us throughout the career.

We wish to express our sincere thanks to our principal Dr. Ganesh Kame, M.H. Saboo Siddik College of Engineering for providing us all the facilities, support and wonderful environment to meet our project requirements.

We would also take the opportunity to express our humble gratitude to our Head of Department of Computer Engineering Dr. Zainab Pirani for supporting us in all aspects and for encouraging us with her valuable suggestions to make our project success.

We our highly thankful to our internal project guide Prof. Saiqa Khan whose valuable guidance helped us understand the project better, their constant guidance and willingness to share their vast knowledge made us understand this project and its manifestations in great depths and helped us to complete the project successfully.

We would also like to acknowledge with much appreciation the role of the staff of the Computer Department, especially the Laboratory staff, who gave the permission to use the labs when needed and the necessary material to complete the project.

We would like to express our gratitude and appreciate the guidance given by other supervisors and project guides, their comments and tips helped us in improving our presentation skills. Although there may be many who remain unacknowledged in this humble note of appreciation but there are none who remain unappreciated.

### Faizaan Lakdawala

### Nusrat Fatima Ansari

### Mehvash Khan

Table of Content

|  |  |  |
| --- | --- | --- |
| **Sr.no** | **Contents** | **Page no.** |
|  | Abstract | 7 |
| 1. | Introduction and Motivation   * 1. Introduction   2. Aim   3. Objective   4. Motivation   5. Diagnosis   6. Effects | 8 |
| 2. | Problem Statement | 11 |
| 3. | Requirement Analysis   * 1. Review of Literature   2. Existing System   3. Drawbacks of Existing System | 12 |
| 4. | Design Details   * 1. Design of the System   2. Flow Chart   3. Block Diagram   4. Data Flow Diagram   5. Use Case Diagram | 14 |
| 5. | Implementation Details   * 1. Process Model | 19 |
| 6. | Technology Used   * 1. Hardware Requirement   2. Software Requirement   3. Programming Language used | 20 |
| 7. | Test Case Design   * 1. Unit Testing   2. Acceptance Testing   3. Alpha Testing   4. Beta Testing | 21 |
| 8. | Results | 22 |
| 9. | Project Timeline | 24 |
| 10. | Task Distribution | 25 |
| 11. | Conclusion and Future Scope   * 1. Conclusion   2. Future Scope | 26 |
| 12. | References | 27 |

|  |  |  |
| --- | --- | --- |
| 13. | Source Code | 28 |

**List of Figures**

**[Indicate the figures wherever used]**

|  |  |  |
| --- | --- | --- |
| **Fig. No.** | **Figures** | **Page No.** |
| **4.1.1.1** | Dataset Description | 15 |
| **4.2** | Flowchart Diagram | 16 |
| **4.3** | Block Diagram | 17 |
| **4.4** | Use Case Diagram | 17 |
| **5.1.1** | Waterfall Model | 18 |

**Abstract**

Diabetes Mellitus is among critical diseases and lots of people are suffering from this disease. Age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, high blood pressure, etc. can cause Diabetes Mellitus. People having diabetes have high risk of diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc. Current practice in hospitals is to collect required information for diabetes diagnosis through various tests and appropriate treatment is provided based on diagnosis. Big Data Analytics plays a significant role in healthcare industries. Healthcare industries have large volume databases. Using big data analytics one can study huge datasets and find hidden information, hidden patterns to discover knowledge from the data and predict outcomes accordingly. In the existing method, the classification and prediction accuracy is not so high. In this paper, we have proposed a diabetes prediction model for better classification of diabetes which includes few external factors responsible for diabetes along with regular factors like Glucose, BMI, Age, Insulin, etc. Classification accuracy is boosted with new dataset compared to existing dataset. Further with imposed a pipeline model for diabetes prediction intended towards improving the accuracy of classification.

# Chapter 1

**Introduction and Motivation**

**1.1 Introduction**

Diabetes is a noxious disease in the world. Diabetes caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves the level of sugar in the blood. Diabetes occurs when the body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffer from diabetes particularly from low- or idle-income countries. And this could be increased to 490 billion up to the year of 2030. However, prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabetics in India is 40 million. Diabetes is the major cause of death in the world. Early prediction of diseases like diabetes can be controlled and save human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose, we use the Pima Indian Diabetes Dataset, we apply various Machine Learning classification and ensemble Techniques to predict diabetes. Machine Learning Is a method that is used to train computers or machines explicitly. Various Machine Learning Techniques provide efficient results to collect Knowledge by building various classification and ensemble models from collected dataset. Such collected data can be useful to predict diabetes. Various techniques of Machine Learning are capable of prediction, however it’s tough to choose the best technique. Thus, for this purpose we apply popular classification and ensemble methods on dataset for prediction.

**1.2 Aim**

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. This project also aims to propose an effective technique for earlier detection of the diabetes disease

**1.3 Objective**

The objective is set to achieve the aims of the project through a Research on statistical models in machine learning and to understand how the algorithms works. This study aims to propose a new model for diabetic’s classification. Numerous algorithms and different approaches have been applied, such as traditional machine learning algorithms, ensemble learning approaches and association rule learning in order to achieve the best classification accuracy.

**1.4 Motivation**

There has been a drastic increase in the rate of people suffering from diabetes since a decade ago. Current human lifestyle is

the main reason behind growth in diabetes. In current medical diagnosis method, there can be three different types of

errors-

1. The false-negative type in which a patient in reality is already a diabetic patient but test results tell that the person is not having diabetes.

2. The false-positive type. In this type, the patient in reality is not a diabetic patient but test reports say that he/she is a diabetic patient.

3. The third type is an unclassifiableWe would like to express our gratitude and appreciation to our parents for motivating and encouraging us throughout the career.

**1.5 Diagnosis levels of diabetes**

* A1C tests/tested It is a blood trail of a person for recent months. The range of the various classes is recorded in the table. It is prescribed for diabetes and prediabetes.
* FPG tests/tested Fasting plasma glucose test level is utilized to recognize prediabetes and diabetes.
* OGT It is oral glucose, the blood test used to analyze prediabetes, diabetes and gestational diabetes.

### 1.6 Effects of diabetes

Diabetes is influenced by different parts of the body which incorporates

* a.Loss of vision Retinopathy retina is a condition where the retina, optic nerve, the focal point is harmed. A result of finishing night visual impairment issues, swelling in the region of the retina, lessening the contact the mind may happen. A Diabetic individual should deal with eye vision through a few tests and pharmaceuticals at the beginning times. The treatment incorporates visual sharpness testing, tonometry, student enlargement, and optic intelligibility tomography (OCT). Different medicines incorporate Anti-VEGF infusion therapy, focal/lattice macular laser medical procedure, corticosteroid.
* b.Kidney neuropathy Chronic kidney infection or diabetic neuropathy is where the high sugar level in blood harms the vessels in the kidney. The usefulness of the kidney is to channel the waste and abundant water in the blood. Because of hypertension and sugar level in Kidney endeavours to have overhead to clean the blood this may prompt kidney disappointment or successive dialysis of blood is required. The treatment may incorporate kidney substitution treatment, kidney and pancreas transplant.
* c.Liver problems Liver assumes an indispensable job in adjusting the blood glucose level in blood through starch digestion by methods gluconeogenesis and glycogenolysis . Sort 2 diabetes expands the danger of liver issues. Fatty liver assumes the stipulated job in creating a liver tumour. The difficulties incorporate renal debilitation, modified metabolism, Insulin opposition and hyperglycaemia, malnutrition. Affected individuals need to experience different antitoxin drugs and administration of the liver incorporates other treatments like the way of life alteration, pharmacological treatment, insulin secretagogues, biguanides, α-glucosidase inhibitors, TZDs, weight to decrease.
* d.Heart problems Cardiovascular ailment : According to American heart affiliation, 68% of individuals will experience the ill effects of heart issues to driving even to death, heart stroke, atherosclerosis or solidifying of the supply routes, stress and load on the heart make individuals to death. Because of high sugar level, blood conveys greater thickness, it adheres to the veins, supply routes and veins put more strain to proceed onward. Persistently it harms the vessels and nerves prompting disappointment of circulatory framework or organ disappointment in person. Hazard for creating cardiovascular illness incorporates hypertension, unusual cholesterol and high triglycerides, corpulence, the absence of physical activity. The effect of different clinical parameters like poor glycaemic control, insulin opposition of diabetes greatly affects heart issues.
* e.The different issues may incorporate foot issues and so on.

# 

# 

# 

# 

# 

# Chapter 2

**Problem Statement**

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to the International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes mellitus or simply diabetes is a disease caused due to the increased level of blood glucose. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. However, early prediction of diabetes is quite a challenging task for medical practitioners due to complex interdependence on various factors as diabetes affects human organs such as kidney, eye, heart, nerves, foot etc. Data science methods have the potential to benefit other scientific fields by shedding new light on common questions. One such task is to help make predictions on medical data. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience.

# Chapter 3

**Requirement Analysis**

## Review of Literature

Diabetes, a non-communicable disease is leading to long-term complications and serious health problems. A report from the World Health Organisation addresses diabetes and its complications that impact on individuals physically, financially, economically over the families. The survey says about 1.2 million deaths due to the uncontrolled stage of health lead to death. About 2.2 million deaths occurred due to the risk factors of diabetes like cardiovascular and other diseases.Diabetes is an ailment caused due to the extended level of sugar obsession in the blood.

K.VijayaKumar et al. proposed random Forest algo- rithm for the Prediction of diabetes develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by using Random Forest algorithm in ma- chine learning technique. The proposed model gives the best results for diabetic prediction and the result showed that the prediction system is capable of predicting the dia- betes disease effectively, efficiently and most importantly, instantly. Nonso Nnamoko et al.presented predicting diabetes onset: an ensemble supervised learning approach they used five widely used classifiers are employed for the ensembles and a meta-classifier is used to aggregate their outputs. The results are presented and compared with simi- lar studies that used the same dataset within the literature. It is shown that by using the proposed method, diabetes onset prediction can be done with higher accuracy. Tejas

N. Joshi et al.presented Diabetes Prediction Using Machine Learning Techniques aims to predict diabetes via three different supervised machine learning methods in- cluding: SVM, Logistic regression, ANN. This project proposes an effective technique for earlier detection of diabetes disease. Deeraj Shetty et al.proposed diabe- tes disease prediction using data mining assemble Intelli- gent Diabetes Disease Prediction System that gives analy- sis of diabetes malady utilizing diabetes patients database. In this system, they propose the use of algorithms like Bayesian and KNN (K-Nearest Neighbor) to apply on dia- betes patients' databases and analyze them by taking various attributes of diabetes for prediction of diabetes disease. Muhammad Azeem Sarwar et al. [10] proposed study on prediction of diabetes using machine learning algorithms in healthcare they applied six different machine learning algo- rithms Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes. Diabetes Prediction is becoming the area of interest for researchers in order to train the program to identify if the patient is diabetic or not by applying a proper classifier on the dataset. Based on previous research work, it has been observed that the classification process is not much im-proved.

In this project , analysis of a dataset is done using various classification techniques like Decision tree , Logistic Regression and deep learning. Comparison and prediction of positive and negative diabetes. Diagnosing diabetes through a data mining tool using the WEKA tool, in terms of accuracy and performance MLP is better.

* 1. **Existing System**

Machines can predict diseases but cannot predict the sub types of the diseases caused by occurrence of one disease. It fails to predict all possible conditions of the people. Existing system handles only structured data. The prediction system are broad and ambiguous. In the current past, countless disease estimate classifications have been advanced and in procedure. The standing organizations arrange a blend of machine learning algorithms which are judiciously exact in envisaging diseases. However the restraint with the prevailing systems are speckled. First, the prevailing systems are dearer only rich people could pay for to such calculation systems. And also, when it comes to folks, it becomes even higher. Second, the guess systems are non-specific and indefinite so far. So that, a machine can envisage a positive disease but cannot expect the subtypes of the diseases and diseases caused by the existence of one bug. For occurrence, if a group of people are foreseen with Diabetes, doubtless some of them might have complex risk for Heart viruses due to the actuality of Diabetes. The remaining schemes fail to foretell all possible surroundings of the tolerant.

**Chapter 4**

**Design Details**

* 1. **DesigChapter Design Details of the system**

The following steps were undertaken in order to design our system:

* + 1. **Data Collection**

The data is gathered from the Github repository which is named as Diabetes Dataset. The dataset has many attributes of 768 patients.

Dataset Description:Table 4.1.1.1

|  |  |
| --- | --- |
| **S No.** | **Attributes** |
| **1** | **Pregnancy** |
| **2** | **Glucose** |
| **3** | **Blood Pressure** |
| **4** | **Skin thickness** |
| **5** | **Insulin** |
| **6** | **BMI(Body Mass Index)** |
| **7** | **Diabetes Pedigree Function** |
| **8** | **Age** |

* + 1. **Data Preparation and Training**

Data preparation is the process of readying data for the training, testing, and implementation of an algorithm. It’s a multi-step process that involves data cleaning & preprocessing, feature engineering, and labeling. These steps play an important role in the overall quality of our machine learning model, as they build on each other to ensure a model performs to expectations.

* Data Preprocessing

Data preprocessing is the act of cleaning and preparing our data for training. This includes organizing and formatting, standardizing, and dealing with missing data. In terms of its importance, many experienced data scientists agree: 80% of their job is data preprocessing.

Data preprocessing is a way to make sure your training data is accurate, complete, and relevant. Sending incomplete or raw data through a model can cause a variety of different errors, which will ultimately result in a much lower overall accuracy.

* Data Labeling

Data labeling is a key part of data preparation for machine learning because it specifies which parts of the data the model will learn from. Though improvements in unsupervised learning have resulted in deep learning projects that do not require labeled data, many machine learning systems still rely on labeled data to learn and perform their given tasks.

* Data Quality

In machine learning, the data preparation process leads into the training of our model, so it’s important to be thorough. To help put ourself in a strong position for smoother data preparation and model training, we make sure we take the time to ensure that we have quality training data from the start, so that we get highest accuracy from our model.

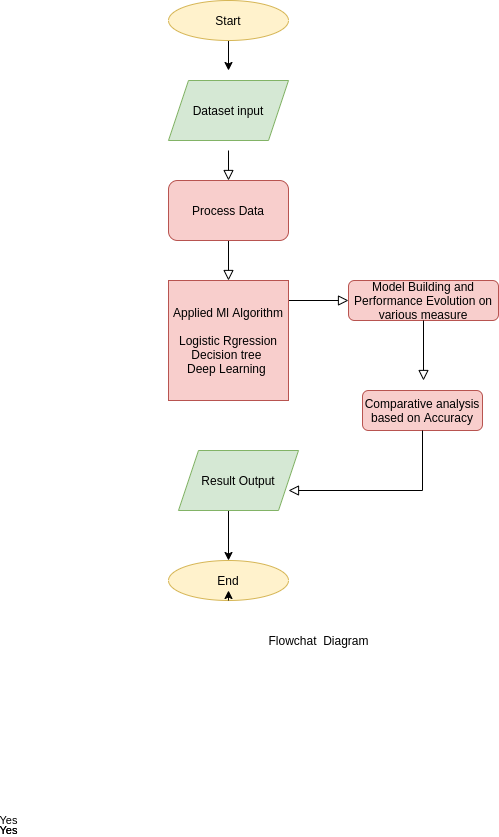
* + 1. **Model selection**

Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset.

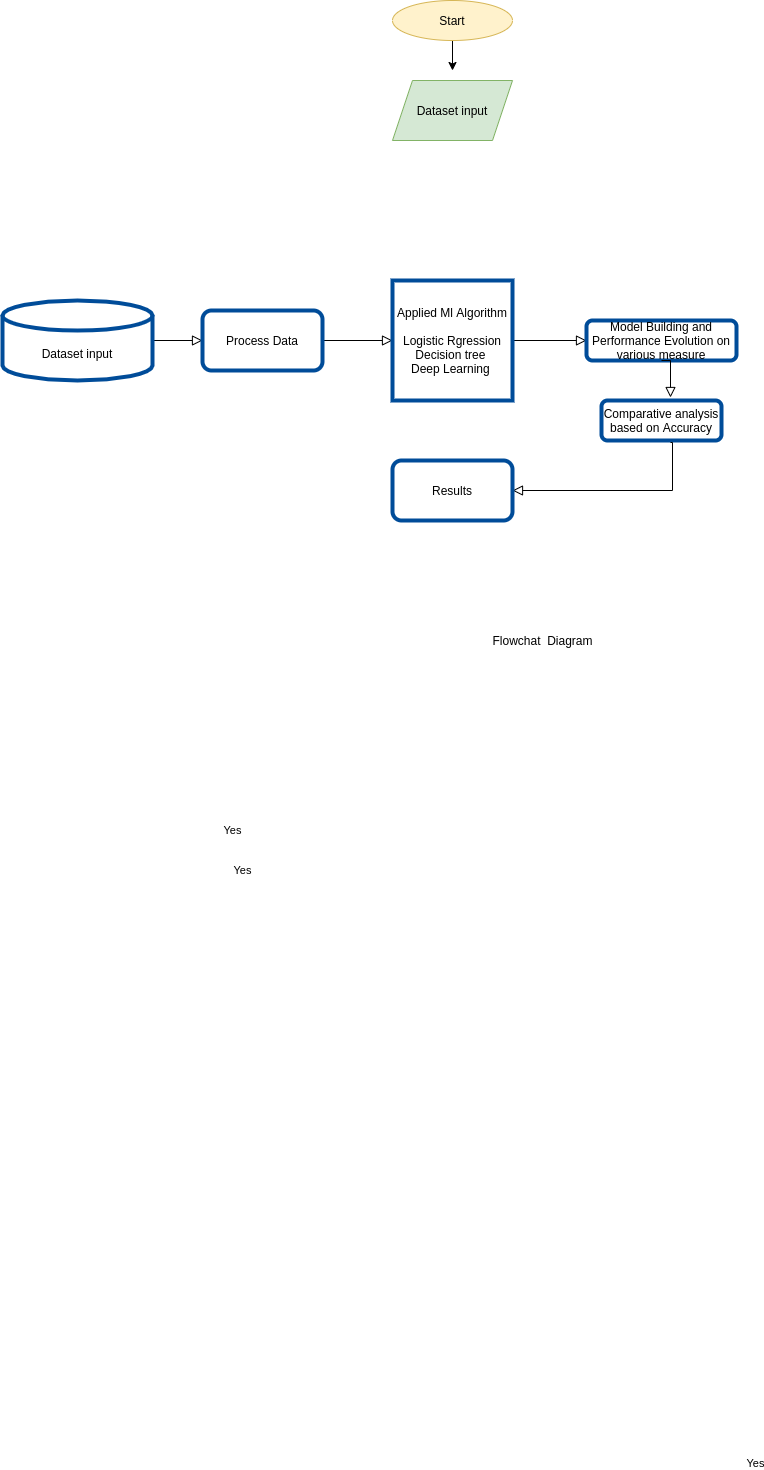
For example, we may have a dataset for which we are interested in developing a classification or regression predictive model. We do not know beforehand as to which model will perform best on this problem, as it is unknowable. Therefore, we fit and evaluate a suite of different models on the problem.

For our project, we have selected and implemented three models on our dataset which are logistic regression, decision tree and deep learning.

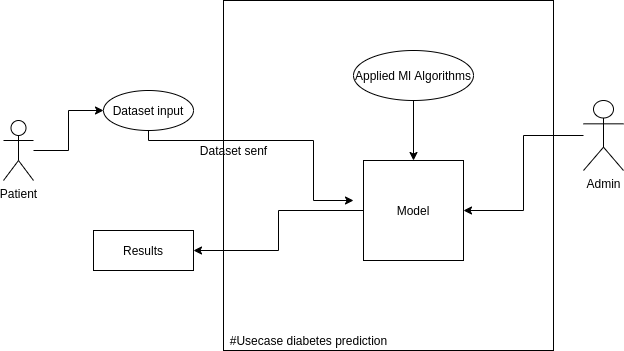
* 1. **Flow Chart**

****

* 1. **Block Diagram**

****

* 1. **Use Case Diagram**

****

**Chapter 5**

**Implementation Details**

* 1. **Process Model**

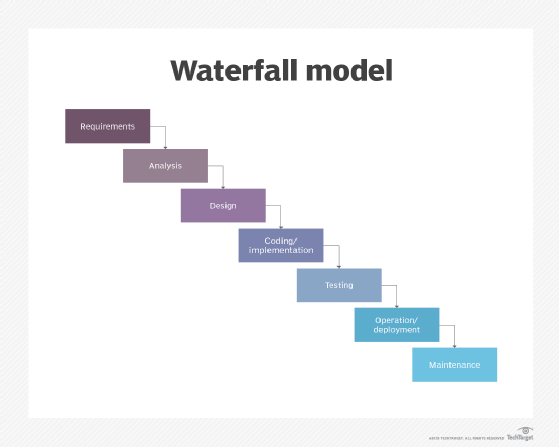
****

Fig.5.1.1 Waterfall Model

The Waterfall Model was the first Process Model to be introduced. It is also referred to as a linear-sequential life cycle model. It is very simple to understand and use. In a waterfall model, each phase must be completed before the next phase can begin and there is no overlapping in the phases.

The Waterfall model is the earliest SDLC approach that was used for software development.

The waterfall Model illustrates the software development process in a linear sequential flow. This means that any phase in the development process begins only if the previous phase is complete. In this waterfall model, the phases do not overlap.

The Waterfall approach was the first SDLC Model to be used widely in Software Engineering to ensure success of the project. In "The Waterfall" approach, the whole process of software development is divided into separate phases. In this Waterfall model, typically, the outcome of one phase acts as the input for the next phase sequentially.

The sequential phases in Waterfall model are −

* Requirement Gathering and analysis − All possible requirements of the system to be developed are captured in this phase and documented in a requirement specification document.
* System Design − The requirement specifications from the first phase are studied in this phase and the system design is prepared. This system design helps in specifying hardware and system requirements and helps in defining the overall system architecture.
* Implementation − With inputs from the system design, the system is first developed in small programs called units, which are integrated in the next phase. Each unit is developed and tested for its functionality, which is referred to as Unit Testing.
* Integration and Testing − All the units developed in the implementation phase are integrated into a system after testing of each unit. Post integration the entire system is tested for any faults and failures.
* Deployment of system − Once the functional and non-functional testing is done; the product is deployed in the customer environment or released into the market.
* Maintenance − There are some issues which come up in the client environment. To fix those issues, patches are released. Also to enhance the product some better versions are released. Maintenance is done to deliver these changes in the customer environment.

**Chapter 6 Technology Used**

* 1. **Hardware Specification**

Workstation: Windows , Linux/Ubuntu and Mac.

Memory: You should have at least 500 Mb for data and 32 Mb RAM.

* 1. **Software Requirements**

Software: Colab, Jupyter and any other notebook.

Python Compiler

* 1. **Programming Language Used**

Language: Python

**Chapter 7 Test Case Design**

* 1. **Unit testing**

In this, the programs that made up the system were tested. This is also called program testing. This level of testing focuses on the modules, independently of one another. The purpose of unit testing is to determine the correct working of the individual modules. For unit testing, we first adopted the code testing strategy, which examined the logic of the program. During the development process itself all the syntax errors etc. got rooted out. For this we developed a test case that results in executing every instruction in the program or module i.e. every path through program was tested. (Test cases are data chosen at random to check every possible branch after all the loops.).Unit testing involves a precise definition of test cases, testing criteria, and management of test cases.

User Input- Each element is tested for valid range and invalid range of data in dataset.

Error Handling- In this system we have tried to handle all the errors that have occurred while running the code.

* 1. **Acceptance Testing**

This refers to the system testing that is performed by the customer to determine whether or not to accept the delivery of the system.

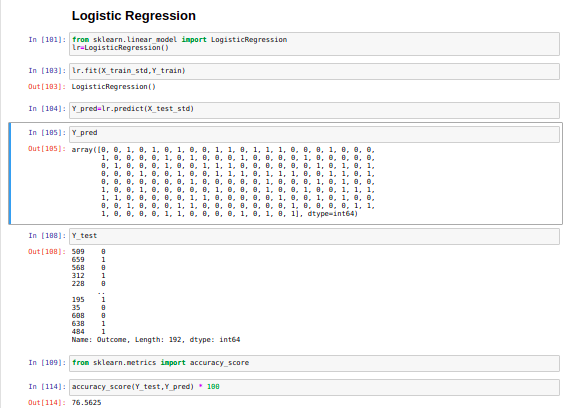
* 1. **Alpha Testing**

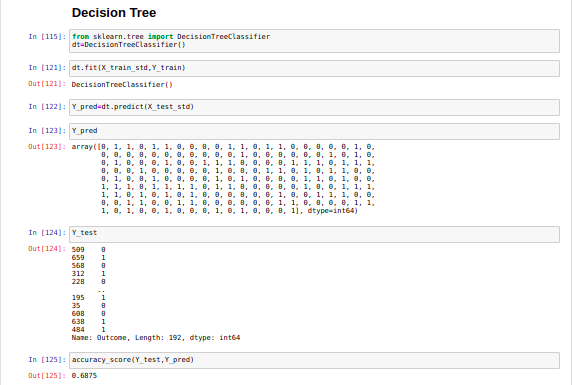
This refers to the system testing that is carried out by the test team with the organization.

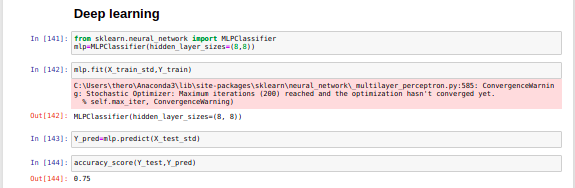
* 1. **Beta Testing**

This refers to the system testing that is performed by a select group of friendly customers.

**Chapter 8 Results**

****

****

****

**Chapter 9**

**Project Timeline**

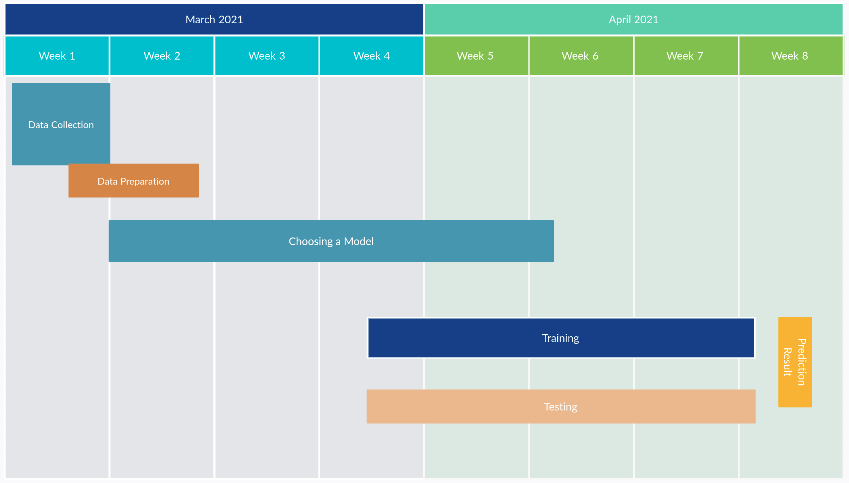
****

Fig.9.1 Timeline Chart

**Chapter 10 Task Distribution**

|  |  |  |  |
| --- | --- | --- | --- |
| **Task No** | **Task Name** | **Name** | **Completion Date** |
| 1. | Data Collection | Faizaan Lakdawala  Nusrat Fatima Ansari | 21 March 2021 |
| 2. | Data Preprocessing | Mehvash Khan | 25 March 2021 |
| 3. | Applying ML Algorithms | Faizaan Lakdawala  Nusrat Fatima Ansari  Mehvash Khan | 11 April 2021 |
| 4. | Choosing Model with Best Accuracy | Faizaan Lakdawala  Nusrat Fatima Ansari  Mehvash Khan | 12 April 2021 |
| 5. | Training | Faizaan Lakdawala  Nusrat Fatima Ansari  Mehvash Khan | 23 April 2021 |
| 6. | Testing | Faizaan Lakdawala  Nusrat Fatima Ansari  Mehvash Khan | 23 April 2021 |

**Chapter 11**

**Conclusion and Future Scope**

* 1. **Conclusion**

The main aim of this project was to design and implement Diabetes Prediction Using Machine Learning Methods and Performance Analysis of those methods and it has been achieved successfully. The proposed approach uses various classification and ensemble learning methods in which Logistic Regression , Decision Tree and Deep Learning are used. And 77% classification accuracy has been achieved. The Experimental results can assist health care to take early predictions and make early decisions to cure diabetes and save humans life.

* 1. **Future Scope**

In Future we can improve the project by including following applications such as :

* By adding UI in the project it makes project more attractive
* We can add hybrid algorithms for better accuracy of result.
* By adding more information to the dataset in the future we can predict Type of diabetes .

**References**

1. <https://www.ijert.org/diabetes-prediction-using-machine-learning-techniques>
2. <https://www.irjet.net/archives/V7/i8/IRJET-V7I8812.pdf>
3. <https://www.tutorialspoint.com/sdlc/sdlc_waterfall_model.htm>
4. <https://www.google.com/search?q=task+distribution&sxsrf=ALeKk03y_jX_WtQ9oz_SLvuf5gfuKpwtfw:1621268984491&tbm=isch&source=iu&ictx=1&fir=aMxxinly5h-HKM%252COUV41Ikyi2dCxM%252C_&vet=1&usg=AI4_-kQAfeUtIAb7MK7Ck9k0-pI440iEnw&sa=X&ved=2ahUKEwiFrJ_PkdHwAhXD4jgGHeMRAv4Q_h0wAHoECAsQAw&biw=1366&bih=619#imgrc=t3SraQy0BPYVXM>

**Source Code**

**Link:**

<https://github.com/NusratAnsari/Diabetes-Prediction-in-ML-3219067-3219069-3219070>