**Text

Description automatically generated**

**A Project Report and Requirement**

**on**

**Diabetic Management System**

**(CSE 6th Semester Miniproject )**

**2021-2022**

**In**

**COMPUTER SCIENCE AND TECHNOLOGY**

**UNDER THE SUPERVISION OF**

**MR. MANOJ KUMAR.(RESOURCE PERSON)**

***By***

**Nandini Gupta (2014421)**

**Enrollment number : GE -192014421**

**Admission number : 19012744**

**Class roll number : 32**

Department of Computer Science and Engineering

**Graphic Era University, Dehradun**

**Introduction:**

Diabetes is an illness caused because of high glucose level in a human body. Diabetes should not be ignored if it is untreated then Diabetes may cause some major issues in a person like: heart related problems, kidney problem, blood pressure, eye damage and it can also affects other organs of human body. Diabetes can be controlled if it is predicted earlier. To achieve this goal this project work we will do early prediction of Diabetes in a human body or a patient for a higher accuracy through applying, Various Machine Learning Techniques. Machine learning techniques Provide better result for prediction by con- structing models from datasets collected from patients. In this work we will use Machine Learning Classification and ensemble techniques on a dataset to predict diabetes. Which is Logistic Regression (LR). The accuracy is different for every model when compared to other models. The Project work gives the accurate or higher accuracy model shows that the model is capable of predicting diabetes effectively.

**Motivation:**

Diabetes is noxious diseases 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 level of sugar in the blood. Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering 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 major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the 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 result 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 can capable to do prediction, however its tough to choose best technique. Thus for this purpose we apply popular classification and ensemble methods on dataset for prediction.

**Problem  Statement:**

**Diabetic Management System:** Diabetes Prediction is becoming the area of interest for researchers in order to train the program to identify the patient are diabetic or not by applying proper classifier on the dataset. Based on previous research work, it has been observed that the classification process is not much improved. Hence a system is required as Diabetes Prediction is important area in computers, to handle the issues identified based on previous research.

**Software requirements:**

* Language used: Python 3.7
* Operating system: Windows 10
* Tool used: Jupyter Notebook

**Hardware Requirements:**

* Processor: Intel Core i5 10 Gen

**PROPOSED METHODOLOGY**

Goal of the paper is to investigate for model to predict diabetes with better accuracy. We experimented with different classification and ensemble algorithms to predict diabetes. In the following, we briefly discuss the phase.

**1.Dataset Description-** the data is gathered from UCI repository which is named as Pima Indian Diabetes Dataset. The dataset have many attributes of 768 patients.

Dataset Description

|  |  |
| --- | --- |
| 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 |

The 9th attribute is class variable of each data points. This class variable shows the outcome 0 and 1 for diabetics which indicates positive or negative for diabetics.

Distribution of Diabetic patient- We made a model to predict diabetes however the dataset was slightly imbalanced having around 500 classes labeled as 0 means negative means no diabetes and 268 labeled as 1 means positive means diabetic.

**2.Data Preprocessing-** Data preprocessing is most important process. Mostly healthcare related data contains missing vale and other impurities that can cause effective- ness of data. To improve quality and effectiveness obtained after mining process, Data preprocessing is done. To use Machine Learning Techniques on the dataset effectively the process is essential for accurate result and successful prediction. For Pima Indian diabetes dataset we need to perform pre processing in two steps.

* + **Missing Values removal-** Remove all the instances that have zero (0) as worth. Having zero as worth is not possible. Therefore this instance is eliminated. Through eliminating irrelevant features/instances we make feature subset and this process is called features subset selection, which reduces diamentonality of data and help to work faster.
  + **Splitting of data**- After cleaning the data, data is normalized in training and testing the model. When data is spitted then we train algorithm on the training data set and keep test data set aside. This training process will produce the training model based on logic and algorithms and values of the feature in training data. Basically, aim of normalization is to bring all the attributes under same scale.

**3.Apply Machine Learning-** When data has been ready we apply Machine Learning Technique. We use different classification and ensemble techniques, to predict diabetes. The methods applied on Pima Indians diabetes dataset. Main objective to apply Machine Learning Techniques to analyze the performance of these methods and find accuracy of them, and also been able to figure out the responsible/important feature which play a major role in prediction. The Techniques is as follows-

1. **Logistic Regression-** Logistic regression is also a supervised learning classification algorithm. It is used to estimate the probability of a binary response based on one or more predictors. They can be continuous or discrete. Logistic regression used when we want to classify or distinguish some data items into categories.
2. It classify the data in binary form means only in 0 and 1 which refer case to classify patient that is positive or negative for diabetes.
3. Main aim of logistic regression is to best fit which is responsible for describing the relationship between target and predictor variable. Logistic regression is a based on Linear regression model. Logistic regression model uses sigmoid function to predict probability of positive and negative class.

**4,MODEL BUILDING**

This is most important phase which includes model building for prediction of diabetes. In this we have implemented various machine learning algorithms which are discussed above for diabetes prediction.

Procedure of Proposed Methodology-

**Step1**: Import required libraries, Import diabetes dataset.

**Step2:** Pre-process data to remove missing data.

**Step3:** Perform percentage split of 80% to divide dataset as Training set and 20% to Test set.

**Step4:** Select the machine learning algorithm i.e. Logistic regression.

**Step5:** Build the classifier model for the mentioned ma- chine learning algorithm based on training set.

**Step6:** Test the Classifier model for the mentioned ma- chine learning algorithm based on test set.

**Step7:** Perform Comparison Evaluation of the experimental performance results obtained for each classifier.

**Step8:** After analyzing based on various measures conclude the best performing algorithm.

**Libraries used:**

**Pandas**: It is an open-source library that is made mainly for working with relational or labeled data both easily and intuitively. It provides various data structures and operations for manipulating numerical data and time series. This library is built on top of the NumPy library.

**Scikit-learn (Sklearn):** It is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

**Pickle:** It is primarily used in serializing and deserializing a Python object structure. In other words, it's the process of converting a Python object into a byte stream to store it in a file/database, maintain program state across sessions, or transport data over the network.

**Tkinter :** It is the standard GUI library for Python. Python when combined with Tkinter provides a fast and easy way to create GUI applications. Tkinter provides a powerful object-oriented interface to the Tk GUI toolkit.

**Joblib :** The aim is to render tools competently and get better presentation when operating with long-running tasks. Evade computing several times the same thing. Code is run again and again, for example, when prototyping computational-heavy jobs, but the solution to mitigate this problem is error-prone and often leads to unprolific outcomes.

**Screens:**

**HOMEPAGE:**

Chart

Description automatically generated

This Page shows the basic view of how the application looks like and the list of modules present in the application.

**USER LOGIN PAGE :**

Chart

Description automatically generated

This page shows the user to log in by giving valid username and password.

**DIABETIC PREDICTION :**

Shape

Description automatically generated with low confidence

**CONCLUSION**

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 ac- curacy from data

The main aim of this project was to design and implement Diabetes Prediction System Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully.The Experimental results can be asst health care to take early prediction and make early decision to cure diabetes and save humans life.