

Presented By: -

**Diksha Dipak Meshram**

**Diabetes Prediction using Machine Learning Algorithm**

***Project Report***

*A*

*Project Report*

*On*

*Diabetes Prediction Using Machine*

*Learning*

Presented By

*Diksha Dipak Meshram*

*Guided By*

*Shri. Vishnuvardhan Y*

Submitted To

*Exposes Data Lab*

Under

*Prof.Ram Meghe Institute of Technology*

*And Research Badnera*

**Abstract**

**Diabetes Prediction: -**

Diabetes is a chronic disease that can cause a worldwide healthcare crisis. According to International Diabetes Federation 382 million people are living with diabetes worldwide. By 2035 this will be doubled to 592 million. Diabetes is a disease caused due increased levels of blood glucose. However, early prediction of diabetes is quite a challenging task for medical practitioners due to the complex interdependence on various factors as diabetes affects human organs such as the kidney, eye, heart, nerves,etc. The Data Science method has 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. The aim of this project is to develop a system that can perform early prediction of diabetes for a patient with a higher accuracy. The project aims to predict diabetes via a logistic regression algorithm.

**Table of Content**

Page No.

Content

5

Introduction

6

Existing Method

7

Proposed method with Architecture

9

Methodology

Implementation

10

11

Conclusion

**Introduction**

Diabetes is the most common disease worldwide and keeps increasing daily due to changing lifestyles, unhealthy food habits, and overweight problems. There were studies handled in predicting diabetes through physical and chemical tests, are available for diagnosing diabetes. Data Science methods have the potential to benefit other scientific fields by shedding new light on common questions. In the proposed system, an efficient way of detecting diabetes is proposed through ML under I used Logistics Regression algorithms experiment result show that the prediction of diabetes done at high accuracy. This project achieves this by analyzing many key factors like the patient’s blood glucose level, body mass index, etc. using logistics recursion algorithm machine learning model and through retrospective analysis of patient’s medical records

**Existing Method**

Existing many research is handled for diabetes detection. Data mining approaches like clustering, and classification were studied in existing systems. Diabetes prediction using algorithms such as k- Nearest Neighbors, K-mean, branch, and bound algorithm was proposed. A basic diabetic dataset is chosen for carrying out the comparative analysis. The importance of feature analysis for predicting diabetes by employing machine learning techniques is discussed. This is proposed to achieve through a machine learning (Logistics Regression) Algorithm. The system improves the accuracy of prediction through machine learning techniques.

A general outline of a proposed method and architecture for diabetes prediction using machine learning.

1. **Data Collection and Preprocessing:**

I am using the Diabetes Dataset from Kaggle for this study. The National Institute of Diabetes and Digestive and Kidney Diseases is the original source of this database.

The next step is cleaning the dataset, which is a crucial step in data analysis. When modelling and making predictions, missing data can result in incorrect results

We found no missing values in the dataset, yet independent features like skin thickness, insulin, blood pressure,; and glucose each have some 0 values, which is practically impossible.

A particular column's mean or median scores must be used to replace unwanted 0 values.

1. **Feature Selection :**

According to the data, women having diabetes have given birth to healthy infants. However, the risk for future complications can be decreased by managing diabetes. The risk of pregnancy issues, such as hypertension, depression, preterm birth, birth abnormalities, and pregnancy loss, is increased if women have uncontrolled diabetes. The likelihood of developing diabetes gradually climbs with glucose levels.

**3. Splitting the Dataset:**

Divide the dataset into training, validation, and test sets. A common split ratio is around 70-80% for training, 10-15% for validation, and 10-15% for testing.

* .

**Proposed method with Architecture**

4. Model Selection and Architecture:

I choose a Machine Learning Algorithm suitable for classification tasks. For example, Logistic regression

5**. Model Training:**

I use a model. fit(X\_train,y\_train) method for training the train data set.

**6. Model Evaluation:**

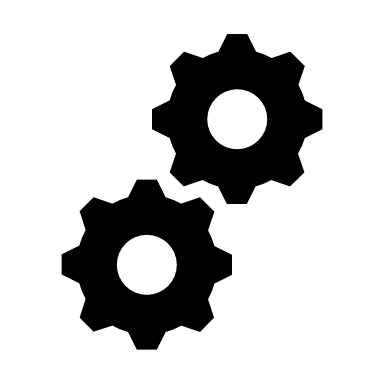
* Evaluate the model's performance using the validation dataset.

Monitor metrics like accuracy, precision, recall, F1-score, and ROC

-**7. Model Validation and Testing:**

* After obtaining satisfactory results on the validation dataset, test the model on the separate test dataset to ensure its generalization

**Methodology**



**Training evaluation results**

**Test Evaluation results**

**Model**

**Model Training**

**Test Data**

**Training Data**

**Historical Data**

Evaluate Results

Logistics Regression (ML Algo)

Split Train (80%) and Test (20%) Dataset

Data Visualization

Data Collection

**Implementation**

I will now divide the data into a training and testing dataset. We will

We will test machine learning models using Logistics Regressions and compare their accuracy in this part. We will use sklearn. preprocessing to convert the data into quantiles before dividing the dataset. We first have to import the most popular Python libraries, which we will use for implementing machine learning algorithms in the first step of building the project, including Pandas, Seaborn, Matplotlib, and others.

I will use Python because it is the most adaptable and powerful programming language for data analysis purposes. In the world of software development, we also use Python.

**Data Splitting**

I will now divide the data into a training and testing dataset. We will use the training and testing datasets to train and evaluate different models. We will also perform cross-validation for multiple models before predicting the testing data.

*I* split the dataset into the train (70%) and test (30%) datasets

A list of machine learning models is passed to the 'model' function, which provides a graph of the cross-validation scores based on the mean of the accuracy values of various models supplied to the function.

we have discovered that the Logistic Regression, models have higher accuracy.

The Objective of the project was to develop a model which would identify patients with diabetes. Many factors influence this process and the outcome. There is presently a serious need for methods that can increase healthcare institution’s understanding of what is important in predicting the diabetes. This project is a small contribution to the present existing method of diabetes detection by processing a system that can be used as an assistive tool in identifying the patients at high risk of being early diabetes. The project predicts the onset of diabetes in aperson based on the relevant medical datils that are collected using Kaggle.com, this data is passed on to the trained model for it to make predictions whether the person is diabetic or nondiabetics.

**Conclusion**