

Mathematics for Intelligent system project Sem 1

PROJECT SEMESTER - 1



NAME: ANOOP BOBY MANUEL ROLL NO: AM.EN.U4AIE21015

NAME: ADITHYA KRISHNA ROLL NO: AM.EN.U4AIE21005

NAME: NAVNEETH KRISHNA ROLL NO: AM.EN.U4AIE210047

NAME: ATHUL GIREESH ROLL NO: AM.EN.U4AIE21020

NAME: ADITHYA S NAIR ROLL NO: AM.EN.U4AIE21006



**Topic**:- Prediction of the Onset of Diabetes.

**Project Abstract**

The Project aims at using the Pima Indians Diabetes Data set to analyze the patient condition and make predictions on the onset of diabetes using Machine Learning techniques and by using Logistic Regression…

**Introduction**

The Pima Indian Diabetes Dataset, originally from the National Institute of Diabetes and Digestive and Kidney Diseases, contains information on 768 women from a population near Phoenix, Arizona, USA.

The outcome tested was Diabetes, 258 tested positive and 500 tested negative.

Therefore, there is one target (dependent) variable and the 8 attributes (TYNECKI, 2018):

1 ) Number of pregnancies

2 ) GLUCOSE (Plasma glucose concentration after 2 hours in an oral glucose tolerance test)

3 ) Blood pressure (Diastolic blood pressure (mm Hg))

4 ) Skin thickness (Triceps skinfold thickness (mm))

5 ) Insulin (2-Hour serum insulin (mu U/ml))

6 ) BMI (Body Mass Index)

7 ) Age

8 ) Pedigree diabetes function

The Pima population has been under study by the National Institute of Diabetes and Digestive and Kidney Diseases at intervals of 2 years since 1965. As epidemiological evidence indicates that type 2 diabetes results from the interaction of genetic and environmental factors, the Pima Indians Diabetes dataset includes information about attributes that could and should be related to the onset of diabetes and its future complications.

📌 Language used for Coding:- Python

📌 Libraries used:

* pandas
* NumPy
* matplotlib.pyplot
* sklearn

## OBJECTIVE

The objective of this project is:

1. Analyze the dataset from the point of view of a Dietitian.
2. Apply machine learning techniques resulting in bridging the gap between datasets and human knowledge.

**Problem Statement:**

Diabetes is one of the deadliest diseases in the world. It is not only a disease but also a creator of different kinds of diseases like heart attack, blindness, etc. The normal identifying process is that patients need to visit a diagnostic center, consult their doctor, and sit tight for a day or more to get their reports. So, the objective of this project is to identify whether the patient has diabetes or not, based on diagnostic measurements. Dataset Used: Pima Indian Diabetes Dataset, from Kaggle.

Data Cleaning will take place as data has got a lot of missing values. Handling missing values can be done either by replacing null values with mode or mean or replacing the null value with a random variable.

**Algorithms Used**

The Machine Learning algorithm used for regression of Pima Indian data set is Logistic regression which understands the relationship between the dependent variable (Outcome ) and one or more independent variables ( Attributes ) by estimating probabilities using a logistic regression equation.

* Logistic regression holds well with binary output data(0,1).
* Logistic regression is **easier to implement, interpret, and very efficient to train**
* Logistic regression is not exactly a Regression model, but it’s a Classification model to be used.

**Advantage of this project:**

The rules derived will be helpful for dietitians to identify patients suffering from diabetes. Further predicting the disease early leads to treating the patient before it becomes critical.

**Solution Approach**

The approach towards solving the problem, by using the Logistic regression as its useful to get Relationship between the dependent and independent variable

* Firstly, Loaded the necessary libraries for solving the linear algebra and data processing

and importing it to the workspace

* The Pima Indian dataset was loaded in the Google colab., then the data set was imported to the workspace
* The data preparation and cleaning are done.
* The processed data is split into the train and test data in the ratio 75:25.
* Confusion matrix, classification report, and accuracy matrix are used to get a clear idea of the accuracy, precision, and recall.

The confusion matric creates a 2x2 matrix in which -

a11 = True Positive (TP)

a22 = True Negative (TN)

a12= False Positive(FP)

a21= False Negative (FN)

Accuracy = TP+TN / TP+TN+FP+FN

Precision = TP / TP+FP

Recall = TP +TP+FN

F1 score = 2(Precision x Recall ) / Precision + Recall ,

Then the accuracy of the prediction is calculated.

* By importing the Logistic regression model from Sk Learn Library, the proceed to find the Accuracy of the Model implemented ..and cross validation is also done .

**Data**

Dataset used is the Pima Indian data set from Kaggle. That includes the attributes such as:-

Pregnancies, Glucose, Blood pressure, Skin thickness, Insulin, BMI, Diabetes pedigree function, and Age.

**Data Validation**

The accuracy obtained by implementing the Logistic regression on the Pima Indian data set is

around the value of 75 % to 80 %

and the Cross-validation Score is also around 75 % to 80 %.

**Data Implementation**

[Github Link](https://github.com/ADITHYASNAIR2021/Pima-analysis-using-Logistic-Regression.git)

**Observation of The Data prediction**

**Nutritional Status**

* Obese 472
* Overweight 179
* Normal 102
* NA 11
* Underweight 4
* Name: Nutritional Status,
* dtype: int64
* 11 women don't have information about BMI. Only 106 of 758 women have normal weight. Most of the women present as overweight or obese.

**Reference on the Level of diabetes:-**

* Normal 571
* Impaired Glucose Tolerance 192
* NA 5
* Name: Glucose Result,
* dtype: int64

Not every woman with impaired glucose tolerance has diabetes. That can show that the ones with impaired glucose tolerance might be at risk of developing diabetes or are diabetic, but not already diagnosed.

* In the given dataset 134 Women were recorded as non-diabetic.
* Most of the Women who participated in the creation of the data set are 22 yrs ( 72 )and the least to be 72, 64, 68, 70, and 81.

**Reference**

Taken reference from -

https://colab.research.google.com/drive/1duJLwYRbHXB5p1NHoGYdzC\_5pp6AFGJK?usp=sharing

Kaggle

https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python In this project

Pima Indians Diabetes Database has been used.

URL: https://www.kaggle.com/uciml/pima-indians-diabetes-database