Mathematics for Intelligent Systems

Prediction of The Onset of Diabetes

PIMA Indians
Diabetes data set

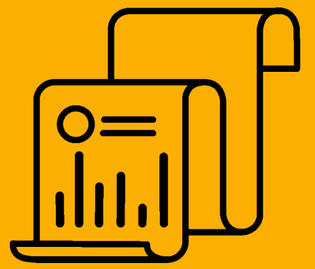
ABSTRACT

This project aims at the prediction of the onset of diabetes using the, 'PIMA Indian diabetes data set'.

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 Pima Indians Diabetes dataset includes information about attributes that are related to the onset of diabetes and its future complications.



OBJECTIVE

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

ATTRIBUTES

1 NUMBER OF PREGNANCIES

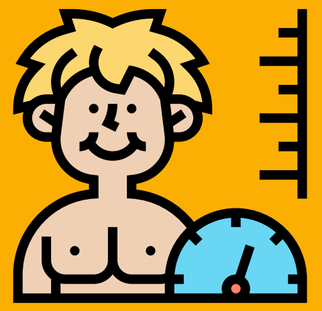
2 GLUCOSE

3 BLOOD PRESSURE

4 SKIN THICKNESS

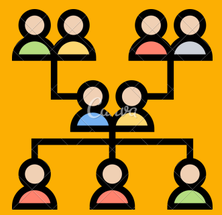
5 INSULIN

6 BMI



7 AGE

8 PEDIGREE DIABETES FUNCTION



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 centre, consult their doctor, and sit tight for a day or more to get their reports.

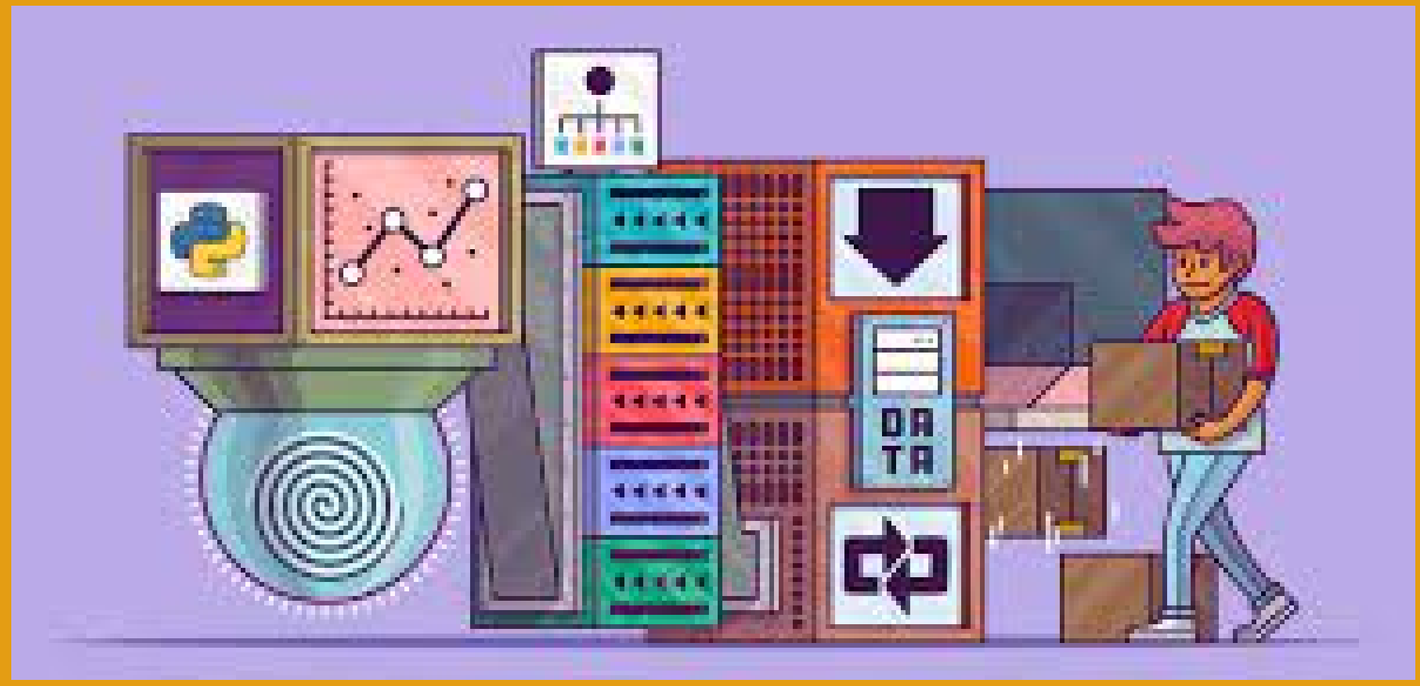
So, the objective of this project is to identify whether the patient has diabetes or not based on diagnostic measurements.

ALGORITHMS

The Machine Learning algorithm used 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).

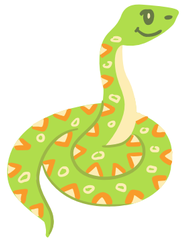
Logistic regression is easier to implement, interpret, and very efficient to train



Logistic regression is not exactly a Regression model, but it's a Classification model to be used.

Language Used

Python



Libraries

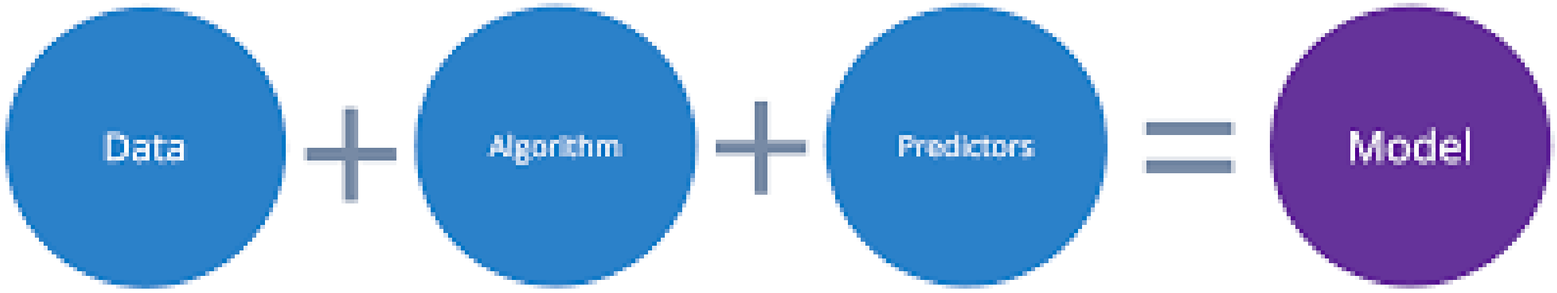
pandas

NumPy

matplotlib.pyplot

sklearn

SOLUTION APPROACH



Historical data

Formula or set of rules to be applied on data

Parameters that directly influence the result

The AI system

1

Loading the necessary libraries

DATA SET

2

Importing the Data set to the workspace

DATA

3

Data Preparation and Cleaning

Splitting into Train Data and Test Data

4

Confusion matrix, classification report and accuracy matrix are used to get a clear idea of the accuracy, precision, and recall.

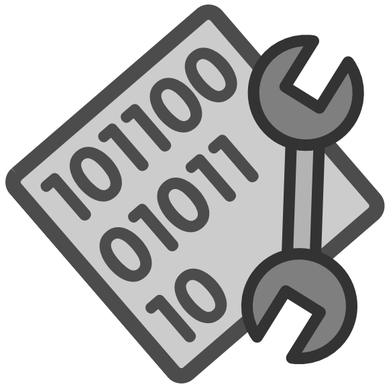
Accuracy of the Model is implemented and cross-validation is calculated

DATA SET

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

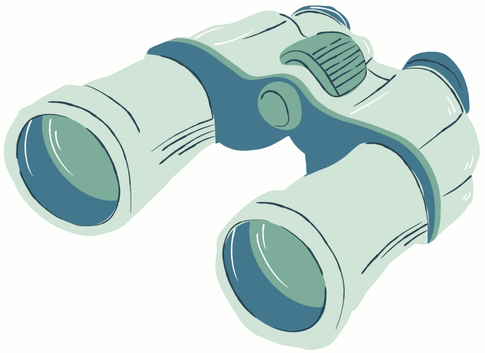
The accuracy obtained by implementing the Logistic regression on the Pima Indian data set is
around the value of 75 % to 80 % and the Cross-validation Score is also around 75 % to 80 %.

Data Implementation

Github Link

CLICK HERE

Observation of The Data prediction



Nutritional Status :

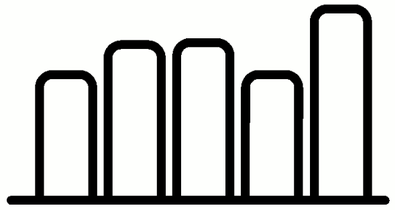
Obese : 472

Overweight : 179

Underweight : 4

Normal : 102

NA 11



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
- 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 aged 22 yrs (72 nos)and one each of 72, 64, 68, 70, and 81 years.

RESULTS

The Result 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 %.

REFERENCES

https://colab.research.google.com/drive/1duJLwYRbHXB5p1NHoGYdzC_5pp6AFGJK?usp=sharing

Kaggle

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

Pima Indians Diabetes Dataset.

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


THANK YOU
Have a good day!