Data Science: EXPOSYS DATA SCIENCE PROJECT

By- ashutosh kumar

July 2020

**DESCRIPTION**

Diabetes is a type of chronic disease which is more common among the

people of all age groups. Predicting this disease at an early stage can help

a person to take the necessary accordingly to either prevent the disease(For people who have already have the disease)

Task:

1. Prepare the data-set

2. Build a model

**1. ABSTRACT**

**2. Table of Contents**

**3. Introduction**

**4. Existing Method**

**5. Proposed method with Architecture**

**6. Methodology**

**7. Implementation**

**8. Conclusion**

# Prediction of diabetes - Type 2 (Mellitus)

Hosted Link : [Click here](http://rpubs.com/SridharCR/diabetes-prediction)

This is a machine learning project based on the prediction of type 2 diabetes, with a given data. It uses logistic regression to classify the diabetic outcomes of each person's record.The diabetes is growing threat nowadays, one of the reasons being that there is no perfect cure for it. There are actually two types of diabetes, namely

- Type 1

- Type 2

The type 2 diabetes is commonly called diabetes mellitus. It can be defined as a chronic condition that affects the way the body processes blood sugar (glucose). We consider the mellitus here. After several researches we found that, some parameters are directly responsible for the mellitus to occur. By using the data of the people with diabetes and without diabetes, a dataset has been build.

We use that dataset to classify the people who are in the risk of getting diabetes.

## Directory structure

\*\*Code\*\*

This directory contains the source code files.

- Prediction.rmd : Source code,R Markdown file

\*\*Data\*\*

- Diabetes.csv : The dataset

\*\*Results\*\*

This directory has the codes and the outputs, presented in a neat format.

- Prediction.html : Presents the project as html file, with wonderful designs (Download is recommended)

- Prediction.pdf : Presents the project as simple document.

# Data-Science-Capstone-Healthcare

Data Science Capstone Project Using Python and Tableau 10

**DESCRIPTION**

Problem Statement

NIDDK (National Institute of Diabetes and Digestive and Kidney Diseases) research creates knowledge about and treatments for the most chronic, costly, and consequential diseases.

The dataset used in this project is originally from NIDDK. The objective is to predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset.

Build a model to accurately predict whether the patients in the dataset have diabetes or not.

**Dataset Description**

The datasets consists of several medical predictor variables and one target variable (Outcome). Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, and more.

Variables Description

Pregnancies Number of times pregnant

Glucose Plasma glucose concentration in an oral glucose tolerance test

BloodPressure Diastolic blood pressure (mm Hg)

SkinThickness Triceps skinfold thickness (mm)

Insulin Two hour serum insulin

BMI Body Mass Index

DiabetesPedigreeFunction Diabetes pedigree function

Age Age in years

Outcome Class variable (either 0 or 1). 268 of 768 values are 1, and the others are 0

**Project Task: Week 1**

**Data Exploration:**

1. Perform descriptive analysis. Understand the variables and their corresponding values. On the columns below, a value of zero does not make sense and thus indicates missing value:

• Glucose

• BloodPressure

• SkinThickness

• Insulin

• BMI

2. Visually explore these variables using histograms. Treat the missing values accordingly.

3. There are integer and float data type variables in this dataset. Create a count (frequency) plot describing the data types and the count of variables.

**Project Task: Week 2**

**Data Exploration:**

1. Check the balance of the data by plotting the count of outcomes by their value. Describe your findings and plan future course of action.

2. Create scatter charts between the pair of variables to understand the relationships. Describe your findings.

3. Perform correlation analysis. Visually explore it using a heat map.

**Project Task: Week 3**

**Data Modeling:**

1. Devise strategies for model building. It is important to decide the right validation framework. Express your thought process.

2. Apply an appropriate classification algorithm to build a model. Compare various models with the results from KNN algorithm.

**Project Task: Week 4**

**Data Modeling:**

1. Create a classification report by analyzing sensitivity, specificity, AUC (ROC curve), etc. Please be descriptive to explain what values of these parameter you have used.

Data Reporting:

2. Create a dashboard in tableau by choosing appropriate chart types and metrics useful for the business. The dashboard must entail the following:

a. Pie chart to describe the diabetic or non-diabetic population

b. Scatter charts between relevant variables to analyze the relationships

c. Histogram or frequency charts to analyze the distribution of the data

d. Heatmap of correlation analysis among the relevant variables

e. Create bins of these age values: 20-25, 25-30, 30-35, etc. Analyze different variables for these age brackets using a bubble chart.

Week 5 final conclusion

Tableau Screen Shot-

![Tableau](Tableau dashbord.PNG)