‘

PROJECT REPORT

Predicting Diabetes using Logical Regression

A PROJECT REPORT

submitted by

# Nishad Chaoji

College – MIT Manipal

Branch – Computer and communication engineering

ACKNOWLEDGEMENT

At the very outset, I would like to give the first honors to the Almighty, who gave me the wisdom and knowledge to complete this project. We also express our gratitude to Advisor and Project Guide for providing us with adequate facilities, ways and means by which I was able to complete the project.

ABSTRACT

Chronic metabolic disorder diabetes mellitus is a rapidly growing problem worldwide with serious social, health and economic consequences. It is estimated that in 2010 there were 285 million people (approximately 6.4% of adults) with the disease. This number is estimated to rise to 430 million if there is no better control or solution. The number of older people and obesity are two main reasons for the increase. In addition, it has been shown that about 50% of people with diabetes are diagnosed until ten years after the onset of the disease, which is why the true prevalence of diabetes worldwide should be very high. This chapter introduces the types of diabetes and diabetic complications such as immunosuppression, periodontal disease, retinopathy, nephropathy, somatic and autonomic neuropathy, cardiovascular disease and diabetes. Also included are current management and treatment, as well as emerging therapies. This project focuses on catching diabetes in its earliest stage and preventing it from further escalating into more severe type of diabetes. There are 9 parameters on which the outcome is obtained.

# Table of Contents

1. Abstract
2. Table of contents
3. Introduction
4. Existing Method
5. Proposed Method with architecture
6. Methodology
7. Implementation
8. Conclusion

INTRODUCTION

The data was collected and made available by “National Institute of Diabetes and Digestive and Kidney Diseases. Several constraints were placed on the selection of these instances from a larger database. All patients here belong to the ages 21 and above and are females.

We’ll be using Python and some of its popular data science related packages. First, we will import Pandas to read our data from a CSV file and manipulate it for further use. We will also use Numpy to convert out data into a format suitable to feed our classification model. We’ll use seaborn and matplotlib for visualizations. We will then import Logistic Regression algorithm from sckit-learn. This algorithm will help us build our classification model. Lastly, we will use joblib available in sckit-learn to save our model for future use.

## 

## EXISTING METHODS

In the context of machine learning and data science, regression specifically refers to the estimation of a continuous dependent variable or response from a list of input variables, or features. There are a variety of regression techniques, ranging from the simplest (linear regression), to complicated statistical classic regression models (Lasso, Elastic Net, etc.), to more complex techniques including gradient boosting and neural networks.

There are 2 types of regression

* Linear Regression
* Logistic Regression

Linear Regression - Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

Chart, scatter chart

Description automatically generated

Logistic regression - Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes. ... Mathematically, a logistic regression model predicts P(Y=1) as a function of X.

Diagram

Description automatically generated

## 

## Proposed Method with architecture

Here, for this project we are going to use logistical regression as the outcome here is either 1 or 0 or the patient has diabetes or does not have diabetes there is no in between.

There are 9 parameters in the given dataset which are Pregnancies, Glucose, Blood pressure, Skin thickness, Insulin, BMI, BPF, Age, Outcome

Using Logistic Regression, we are going to predict whether the patient has diabetes or not based on the above parameters

## METHODOLOGY

Libraries used –

1. Pandas version 1.3.1
2. Numpy version 1.21.1
3. Matplotlib version 3.4.2
4. Sckit-learn version 0.24.2
5. Seaborn version 0.11.1

Clustering Algorithm – K-means Algorithm

1. Specify number of clusters i.e., K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e., assignment of data points to clusters isn’t changing.

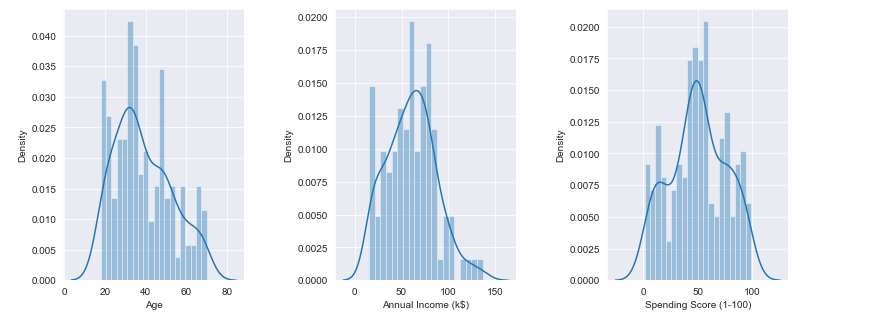
Dataset – Customer ID, Age, Gender, Annual Income and Spending Score.

## IMPLEMENTATION

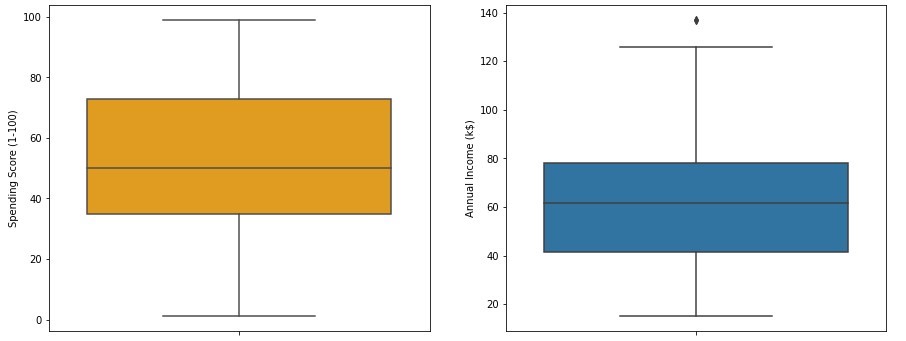
I started with loading all the libraries and dependencies and then imported the data. The columns in the dataset are customer id, gender, age, income and spending score.

I dropped the id column as that does not seem relevant to the context as we do not need the customer ID for segmentation.

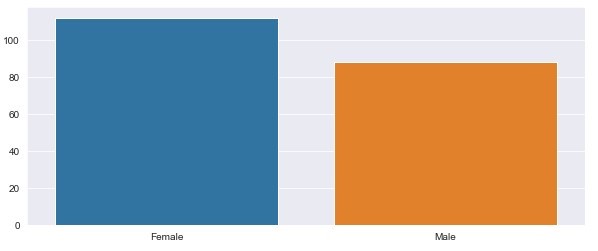
I plotted the Distplots for Age, Spending score and Annual income



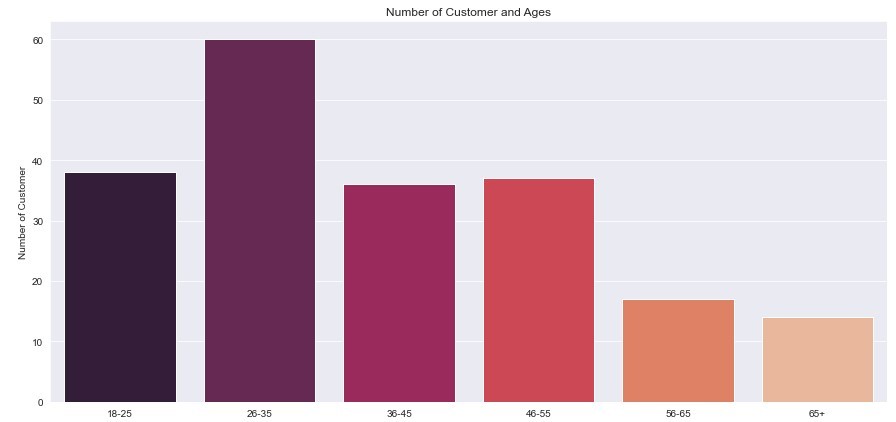
Next, I made a box plot of spending score and the annual income. This to better visualize the distribution range of the given dataset. The range of spending score is clearly more than the annual income range.



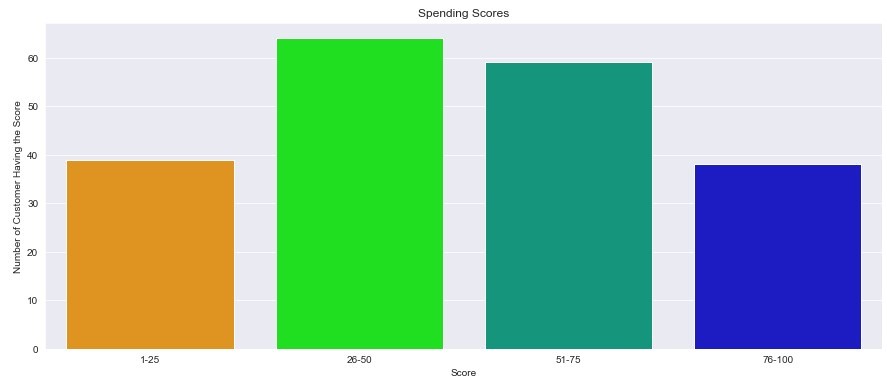
Next, I made a bar plot to check the distribution of male and female population. From the results the female population clearly outweighs the male counterpart.



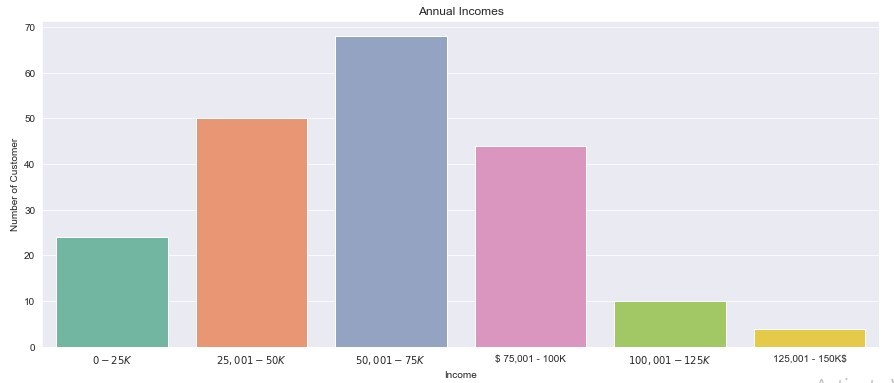
Next, I made a bar plot to check the distribution of number of customers in each age group. Clearly the age group of 26-35 has the highest no of customers.



Next we make a bar plot to visualize the number of customers according to their spending scores. Most of the customers have spending score in the range 26-50 closely followed by 51-75.



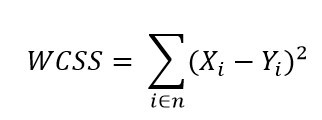
Next, I made a bar plot to visualize the number of customers according to their annual income. Many of the customers have annual income in the range 50,001 to 75,000.



Next, I plotted WCSS against K Value to find out the optimal value of K. WCSS measures sum of distances of observations from their cluster centroids which is given by the below formula.

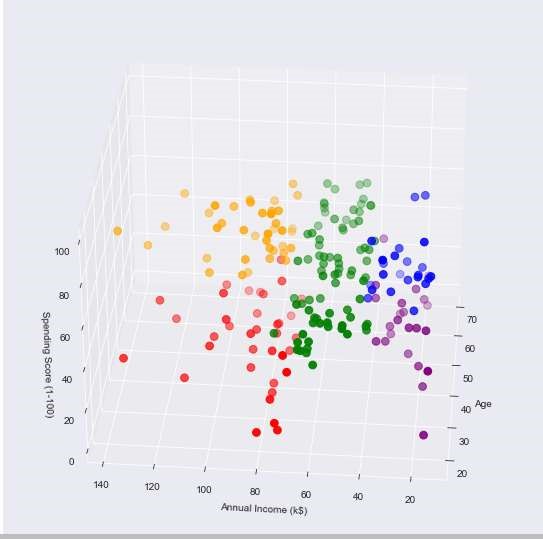
WCSS – Within clusters sum of squares

K- No of clusters



where Yi is centroid for observation Xi. The main goal here is to maximize number of clusters to the point where each clusters centroids doesn’t move anymore . We calculated the Value of K using the elbow method

Lastly, I made a 3D plot to visualize the spending score of the customers with their annual income. The data points are separated into 5 classes which are represented in different colors as shown in the 3D plot.



## CONCLUSION

Cluster 1 (Blue Colour) -> earning low but spending high

Cluster 2 (Red Colour) -> Earn more and spend less

Cluster 3 (Orange Colour) -> earning high and spending high [TARGET

Audience]

Cluster 4 (Purple Colour) -> earning less and spend less

Cluster 5 (Green Colour) -> Average in terms of spending and Earning

## Learning Outcomes

1. What is Data Analysis and why is it so Important

In today's world.

1. what is customer segmentation and why is it important
2. How machine learning helps with Customer

Segmentation

1. Practical Implementation of K-means Algorithm