DIABETES DETECTION

By

GROUP A
RECESS PROJECT
DEPARTMENT OF NETWORKS
SCHOOL OF COMPUTING AND INFORMATICS TECHNOLOGY

CORDINATOR

DR GRACE KAMULEGEYA
DEPARTMENT OF NETWORKS

12TH-July 2019

GROUP MEMBERS :

#	Names	Registration Number
1	Kijjambu Hassan	17/U/44437
2	Okwe Isaac	17/U/18975
3	Lugala Hillary	17/U/4141/ps
4	Byakatonda Benard	17/U/3802/ps

1.0 Background of the Data set

In this project we are using the pima indians diabetes data set.. This data set was downloaded as a CSV file from kaggle.com.

1.1 Description of the data

The data sets consists of 8 medical predictor (independent) variables and one target (dependent) variable (Outcome) and 769 entries.

The columns in the dataset are Pregnancies, Glucose, BloodPressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeF, Age and Outcome.

1.2 Sources of the data

The data set is a traditional structured data inform of table with rows as the instances of the diabetes dataset and columns as the features (attributes)

This data set is originally from the National institute of Diabetes and kidney diseases. The objective of the data set is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the data set. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years of pima indian heritage.

1.3 Features

a) Pregnancies. This is the number of pregnancies the patient has had.

b) Glucose. This is plasma glucose concentration in an oral glucose tolerance test.

c) BloodPressure. This is the diastolic blood pressure of the patient.

d) Skin. This is the triceps skin fold thickness in millimeter (mm).

e) Insulin. This is a 2-hour serum insulin.

f) BMI. This is the body mass index of the patient

g) DiabetesPedigreeF. Diabetes pedigree function

h) Age. This is the age of the patient who underwent the diabetes test.

i) Outcome. The outcome refers to the result of the diabetes test.

2.0 Data pipeline

This is the overall step by step process towards obtaining, cleaning, visualizing, modeling, and interpreting data within a business or group. These steps are shown in the flow diagram below;

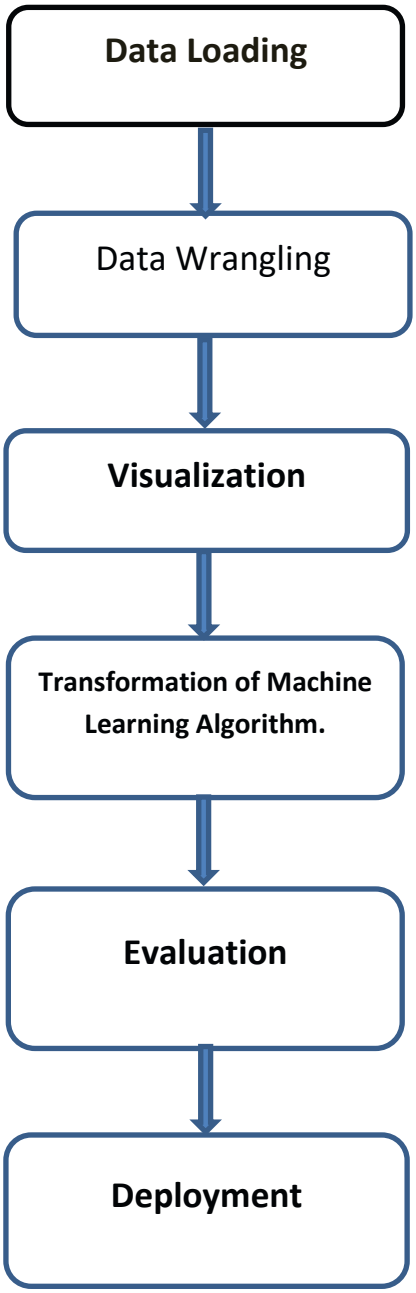

Figure1: Flow diagram showing the data analysis approach (Data pipeline).

Data loading

This is the first step in the data pipeline. It involves importing the required packages like pandas, “numpy” which are used for loading and manipulation of the file. After the file is loaded using “pandas” library, manipulation can be done using both “pandas” and “numpy “, but “numpy” only deal with arrays.

Data wrangling

This is process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as visualization and modeling. The sub-steps involve in this process include; label encoding, filling missing values, feature scaling and scaling the outliers.

Visualization

This is the graphical representation of data by using visual elements like charts, graphs, and maps. These data visualization tools shall provide an accessible way to see and understand the relationships between features and identify outliers and patterns in our data.

Transformation of machine learning algorithm

Logistic Regression
Logistic regression is a Machine Learning classification algorithm that is used to detect the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). The algorithm will detect if there is diabetes or not.

Evaluation of logistic regression algorithm

At this stage we shall apply logistic evaluation algorithm that is confusion matrix to our model to evaluate its performance. This algorithm will help us identify the level of accuracy of the model

Presentation

After accomplishing this project we shall use web application to present the outcome of this project.