DIABETES DETECTION

BY

GROUP A

RECESS PROJECT-CONCEPT PAPER

DEPARTMENT OF NETWORKS

SCHOOL OF COMPUTING AND INFORTION TECHNOLOGY

CORDINATOR

DR.GRACE KAMULEGEYA

DEPARTMENT OF NETWORKS

DATE: 12th – JULY – 2019

GROUP MEMBERS				
#	F.NAME	L.NAME	REGNO	SIGNATURE
1	HASSAN	KIJJAMBU	17/U/44437	
2	ISAAC	OKWE	17/U/18975	
3	HILLARY MOSES	LUGALA	17/U/4141/ps	
4	BENARD	BYAKATONDA	17/U/3802/ps	

Introduction

This documentation provides the overall background of our idea behind the diabetes detection system. It examines the background, description and the source our dataset and the overall step by step procedure toward the implementation of this project (i.e. data pipeline).

1.0 Background of the Data set

In this project, we are using the 'pima Indians diabetes' data set which is a CSV file downloaded from kaggle's website.
The attitude to look, analyze and make predictions on this data emerged due to rapid increase in suffering of different diseases in this century that used not to exist even before. To a greater extent, people used not to suffer from diabetes but it has now been found out that it is at a rampant increase and yet there no sure reasons for the contributions and more so in pregnant mothers.
Diabetes is yet another serious disease in this century that is as result of low or high sugar concentration levels as compared to those recommended levels the human body must have, therefore, our research and analysis revolves around women and more so pregnant mothers by looking at the different test features and how they contribute to be either positive or negative after the diabetes test.

1.1 Description of the data

The data set is a traditional structured data that is inform of table with rows as the instances of the diabetes test and columns as the features (attributes).
The data set consists of 9 columns in total and 769 entries. The first 8 columns are results from various medical tests performed on each patient and the last column (outcome) is the dependent variable which determines whether the patient is diabetes positive or negative.

These columns are explained below;

a) Pregnancies: This is the number of pregnancies the patient has had.
b) Glucose: This is the level of blood sugar in the patient's body.
c) Blood Pressure: The rate of blood flow in the patient's body.
d) Skin. This is the triceps skin fold thickness in millimeter (mm), in other words, it is the length of skin folds or wrinkles on the patient's skin
e) Insulin. This is the hormone that regulates the amount of glucose in the body.
f) BMI (Body Mass Index). The weight of the patient measured in kilograms (kg)
g) DiabetesPedigreeF. Diabetes pedigree function is a function used to provide data on diabetes mellitus history in relatives and the patient's genetic relationship with them
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 which may be a zero (0) or one (1), signifying absence and presence of diabetes respectively.

1.2 Source of data

According to Kaggle's website, the data set is a traditional structured data inform of table with rows as the instances and columns as the features (attributes) and is originally from the National institute of Diabetes and kidney diseases.
The main thesis of this data set is to diagnostically predict the presence or absence of diabetes in a patient's body given certain diagnostic medical tests as prescribed in the data set.

2.0 Data pipeline

This is the overall step by step process towards obtaining, cleaning, visualizing, modeling, and interpreting data within a business and in a management information system. Below are the steps in the flow diagram and later explained in the next section

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

2.1 Data loading

This step involves importing the required packages and libraries 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“, though “numpy” only deals with arrays.

```
In [9]:
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 %matplotlib inline
```

```
In [13]:
1 data=pd.read_csv('diabetes.csv')
2 data.head(10)
```

Out[13]:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

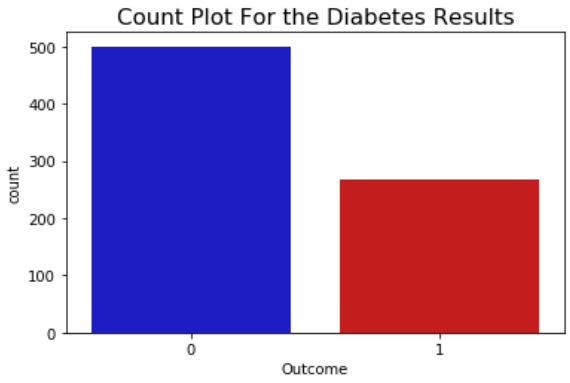
Figure 2: Packages imported and a sample of the dataset printout

2.2 Data wrangling

The process involves cleaning the data as a way of transforming and mapping it from one "raw" data form into another format thus making it appropriate and valuable for a variety of intended purposes such as visualization and modeling. The sub-steps involved in this process include; label encoding, filling missing values, feature scaling and removing the outliers and many more.

2.3 Visualization

This is the graphical representation of data by using visual elements such as charts, graphs, and maps. For the purpose of this project, a count plot will be used to show the total of those with and without diabetes, other graphs will be used to show the relationship between the different features in the dataset. Such visual representations may be a way to argue to those responsible so as to take immediate response if need be. They also provide an accessible way to see and understand the relationships between features and identify outliers and patterns in the data.
Data visualization also helps in intuitive understanding of the data and observes certain patterns in data.



2.4 Transformation of machine learning algorithm

Since we are dealing with categorical data, Logistic regression is used as a Machine Learning classification algorithm to detect the probability of a dependent variable. In this algorithm, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.), hence the algorithm will detect if the patient has diabetes or not given the conditions and test results.

2.5 Evaluation

This stage involves the evaluation of the above mentioned algorithm by using evaluation algorithm to determine the level of accuracy of our model. In this project we are using confusion matrix to evaluate our model since it is categorical problem.

2.6 Presentation/Deployment

Having accomplished the project, it will be deployed and presented as a web application where the success will be measured given its outcome.