DIABETES DETECTION

BY

GROUP A

RECESS PROJECT-SYSTEM DESIGN SPECIFICATION DOCUMENT

DEPARTMENT OF NETWORKS

SCHOOL OF COMPUTING AND INFORTION TECHNOLOGY

CORDINATOR

DR.GRACE KAMULEGEYA

DEPARTMENT OF NETWORKS

DATE: 26th – JULY – 2019

Group Members			
#	F.NAME	L.NAME	REGNO
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

System Design Documentation

This design document explains the different data pipelines that are to be used in the implementation of a diabetes detection system, basing on the following flow diagram.

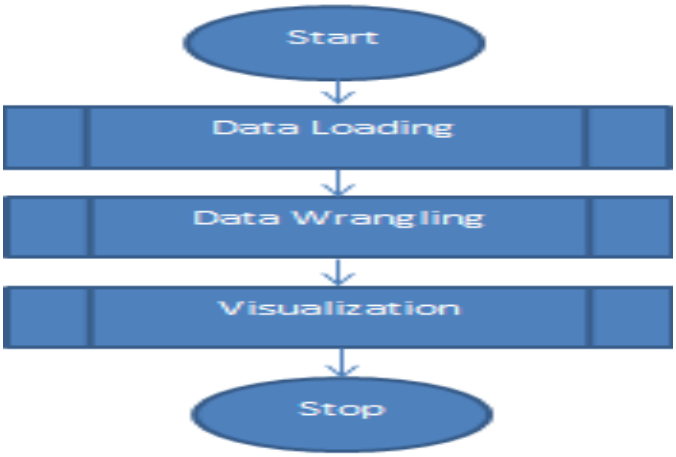

Figure 1: High level view of pipe

The above, is a high level view of the data pipe with three modules, as Data Loading, Data Wrangling and visualization. These steps are further narrowed down and explained as follows.

Data Loading Module:

Most often in structured data, different formats are dealt with in data analysis, which may include csv files, excel files, database (sql) or even web services and all these must be put in tabular formats organized in rows and columns and hence the process of data loading. It is a primary step used to transform different structured data formats into row-column formats.

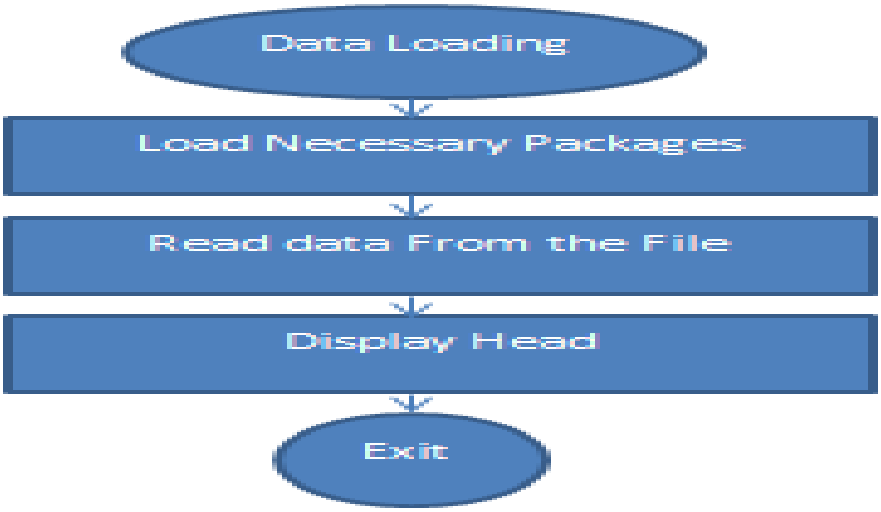

Figure 2: Data Loading Module

In this section, we present the ways how the intended structured data will be transformed into the intended formats for further exploration analysis steps and visualization.

Algorithm: DATA LOADING

1. Load Necessary Packages
2. Read data from file
3. Print data sample

Load packages: The import term is used to load the packages and several of these include; pandas as a basic package for analysis, numpy for array manipulation, matplotlib and seaborn for plotting graphs, sklearn for data modeling, validation and evaluation purposes.

Read data: Pandas package will be used because of its rich and efficient object (data frame) for data manipulation, and this will purposely be used to read the csv file using its alias pd together with read_csv method and transform it into a data frame. The numpy package with its alias np capable of manipulating arrays is used together with pandas.

Print data sample: The head () function from pandas package is used to specify how many instance observations to return and this is to help have a feel of the successfulness of the above operations.

Data Wrangling module

The step involves transformation of data from its raw form by cleaning, restructuring, reshaping and possibly converting it into a useful form upon which decisions can be based on.

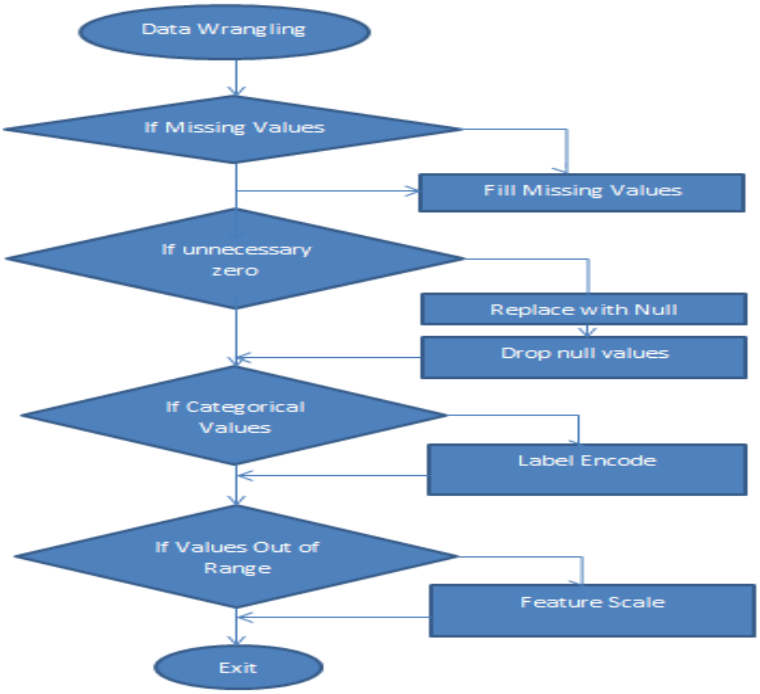

Figure 3: Data wrangling module

Algorithm: DATA WRANGLING

1. If data has missing values:
   Fill missing values
2. If data has unconvincing zero:
   Replace with null values
   Drop affected row values
3. If data is categorical:
   Perform label encoding
4. If data is out of range:
   Perform feature scaling

Following the algorithm and according to the data set which is under study, there are no missing values that we refer to as null values and this is done using the pandas is-null method, but rather, a lot of unconvincing zeros are to be handled by replace methods from numpy package and the dropna method from pandas package is to drop the affected rows.

Feature scaling is used to handle data out of range, this is one of the data pre-processing steps that is applied on to data features, possibly independent variables as a way of normalizing the data based on the intended range which helps in speeding up the calculations in a given algorithm or model. It is possible with sk-learn preprocessing package, which allows the use of standardizer and the fit methods as well.

Data Visualization

This is a way of understanding the nature of the data by representing it in a graphical format so that the hidden relationships, trend and elements in the data which cannot be easily seen is depicted in the visual context.

It is used to analyze the data so as to understand the nature of the data, trend, and the relationship between the features in the data.

The section below examines how the Diabetes detection system performs visualization in three different graphs/plots in a flow diagram.

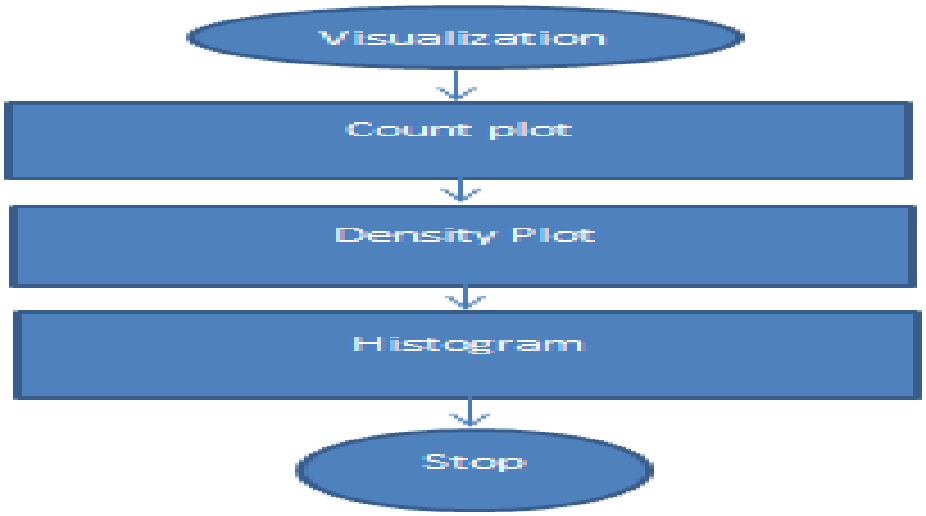

Figure 4. Visualization module

Algorithm: Data Visualization

1. Draw Count plot
2. Draw Density plot
3. Draw Histogram

Count plot

The visual uses the outcome column of the diabetes data set to graphically show the two target categories, that is; those with and without diabetes, a count plot () method of the seaborn package to show the count of the observation in each categorical group(bin) using a bar.

Density plot

This shows the distribution of the values in a given column for all the continuous variables so as to decide whether to scale a given column or not. The distplot () method from seaborn package is used to show the distribution of data over a continuous interval.

Histogram

This is similar to a density plot but it graphical show the shape of the distribution and it gives us a deeper understanding of the distribution by analyzing a variable at a time. The hist () method of the matplotlib package is used to show the distribution of numerical data over a discrete interval.