DIABETES DETECTION

BY

GROUP A

RECESS PROJECT-SYSTEM IMPLEMENTATION REPORT

DEPARTMENT OF NETWORKS

SCHOOL OF COMPUTING AND INFORTION TECHNOLOGY

CORDINATOR

DR.GRACE KAMULEGEYA

DEPARTMENT OF NETWORKS

DATE: 05th – AUGUST– 2019

Contents

Introduction

This documentation examines the implementation of the diabetes detection system. It explains each visual element as to the implementation of the system, the insight in the dataset and what they deduce. These visual elements includes count plot, histogram and density plot as explain in the visualization section below.

Visualization

Count plot
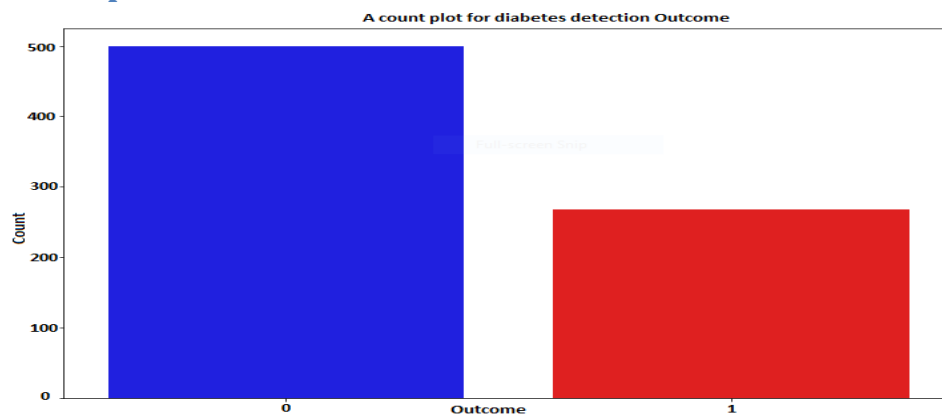
This is a plot of Outcome against number of patients (count), where the number of patients without diabetes is more as compared to those with diabetes hence an imbalanced dataset. Analysts generally use this kind of plot to identify whether there some imbalances in the data set for which they are dealing with as it can be viewed from the plot above.

Histogram

In analyzing the dataset we have explored each feature in the dataset. Here we have used a histogram to show how each feature is distributed. The distribution of each variable as a feature in the data set can clearly be viewed using a histogram and thus, analysts use it to tell the variable's distribution with respect to the feature count.
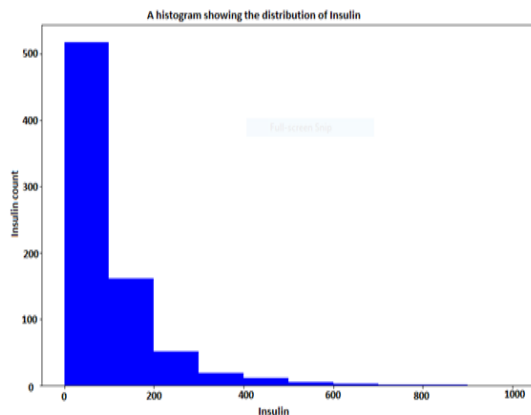
Glucose Distribution
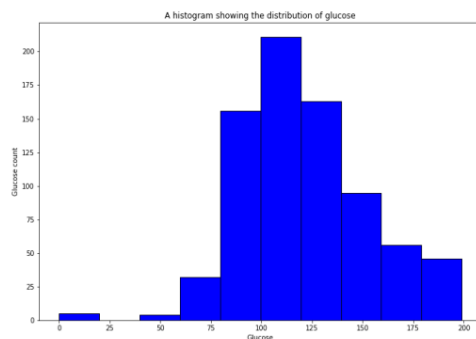
Pregnancy Distribution
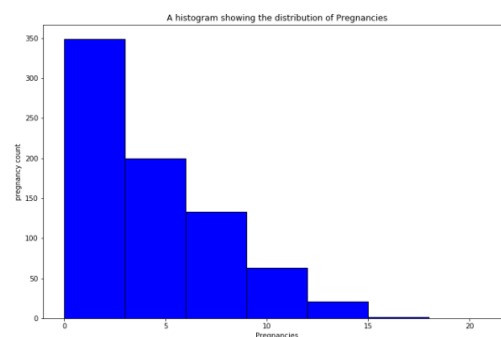
Under pregnancy distribution, it is viewed that, the majority of patients had pregnancies utmost 3 times, that is, 0-3 , followed by those between 3-6, then 6-9, a few patients with 9-12 and very few with 15-18 times.
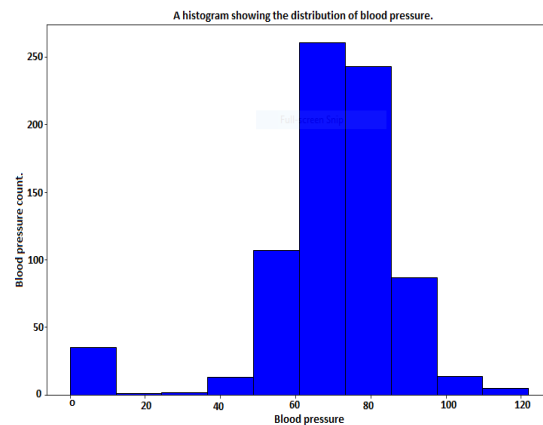
The distribution in glucose, shows that majority of the patients had glucose concentration ranging from 80-140mM, then those ranging from 160-200mM, then 40-80mM, 0-20mM and those without ranging between 20-40mM.

Insulin Distribution

A histogram showing the distribution of Insulin

Blood pressure distribution

A histogram showing the distribution of blood pressure.

From Insulin histogram, high levels of insulin in patients are between 0-100U/ml, followed by 100-200 U/ml, then 200-300U/ml, hence tending to zero with the increase in the insulin-values.

The distribution of blood pressure shows that, high blood pressure in patients was ranging from 60-85mmHg and fewer patients range 0-60mmHg and 85-120mmHg, thus, patients with very high and very low blood pressure are few. Majority have blood pressure in between the central and upper quartile range.

BMI Distribution

A histogram showing the distribution of BMI
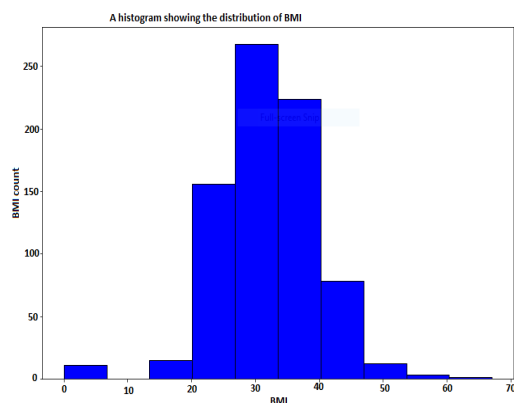
Sick thickness Distribution

A histogram showing the distrubution of skin thickness.
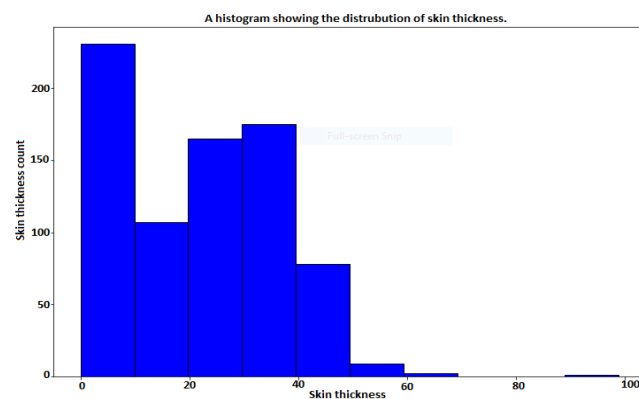
From BMI distribution, it can be noted that many patients have BMI ranging from 20-40kg/m^2 and fewer patients with BMI ranging between 0-20kg/m^2 and 40-70kg/m^2.

From skin thickness histogram, the distribution shows that many patients had the skin thickness ranging between 0-10mm, followed by 20-40mm, 10-20mm, 40-50mm with very few in the range 50-70mm and 90-100mm and non in the range 70-90mm.

Diabetes Pre-degree Function Distribution

A histogram showing the distribution of diabetes pedigree function.

Age Distribution

A histogram showing the distribution of ages
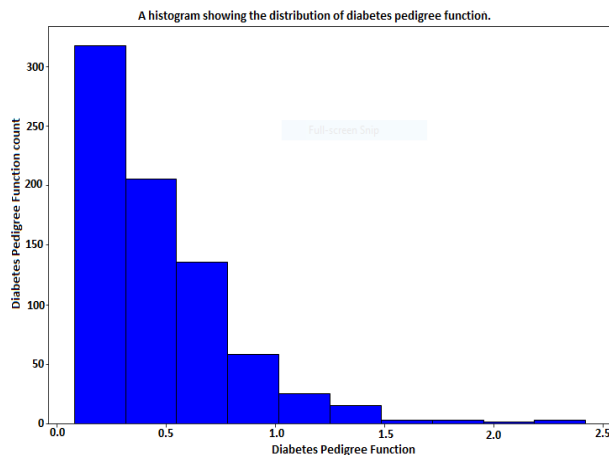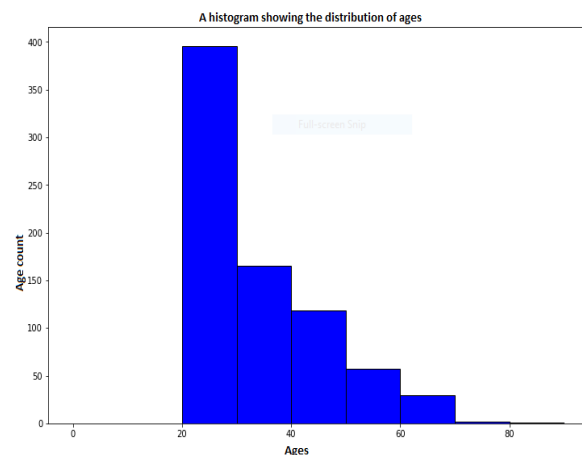
Concerning diabetes predegree function, the distribution shows that many patients have little pedigree function.

From Age distribution, shows that many patients are in the age range of 20-30, followed by 30-40 and then followed by those in 70-90 years in the order of decreasing number of patients.

Density plot

To begin with, a density plot shows the distribution of data over a continuous variable's interval in which the peaks depict where the values are concentrated for a given interval.

The section examines the distribution of multiple features (variables) in the same plot to compare their variations in terms of their mean values.

The distribution of pregnancies, Blood pressure, Skin thickness, and Age.

A density plot showing the distribution of Pregnancies, Blood Pressure, Skin Thickness and Age
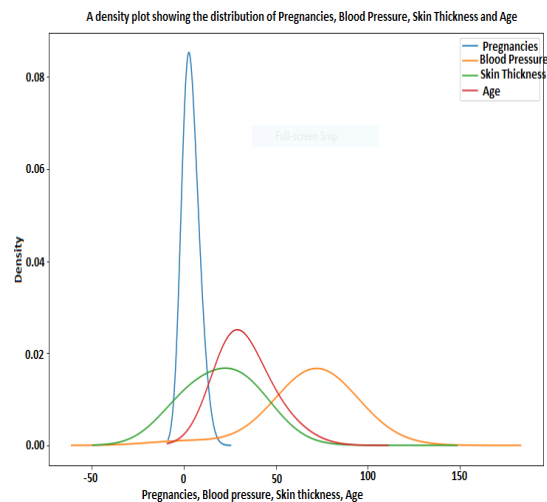
The distribution of Insulin, Glucose and BMI

A density plot showing the distribution of Glucose, Insulin, and BMI
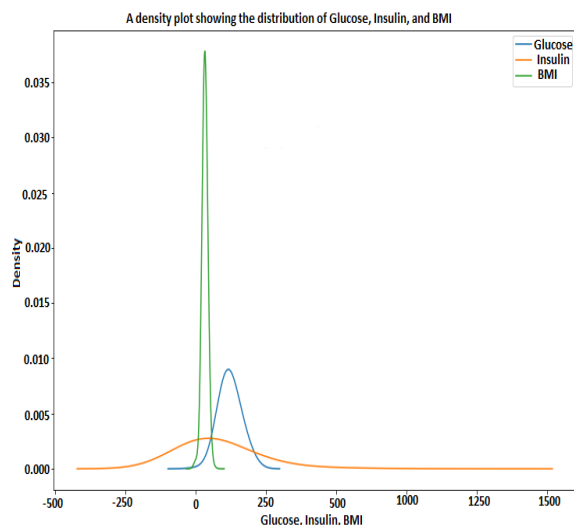
The distribution shows that the mean number of pregnancy is around 1 time, the mean blood pressure is around 75mmHg, the mean Skin thickness is around 25mm and that of age is in around 30 years.

The distribution shows that the mean amount of glucose concentration in patients is around 120mM, the mean amount of Insulin is around 100U/ml and the mean value of the BMI in the patients is around 30kg/m^2.

Recommendations

For count plots, it is always a good idea to use them to find and tell whether there are imbalances in the data set and possibly identify the different ways of handling imbalances most specifically, when dealing with a categorical problem, thus we argue to analysts and data engineers to always find possibilities of handling imbalances by using a count plot so sooner as imbalances contribute to a greater error bound and lower the accuracy levels of an algorithm if not handled critically.

For purposes of exploring a single variable, it would be a brilliant idea for an analyst to always refer to histogram, therefore, we used a histogram to explore the distribution of a single variable in our data set and this helps to clearly understand the data variable thoroughly.

For comparison of different features and their variations over a continuous interval, it is highly recommended for analysts to use a density plot as it gives a good differentiation for two or more variable distributions by using their mean and peak values; hence this helps to identify the greatest contributor to problem under study.