## 17MDC36 - BUSINESS STATISTICS LABORATORY USING R

## PROJECT

## DIABETES PREDICTION

## SUDHARSHINI . R – 71762133046

## HANSA . A. S – 71762133015

## ABSTRACT:

The purpose of this project is to predict the rate of the diabetes of the women who are above the age of 21. The diabetes dataset consists of 100 data points with 9 features. In particular all patients here are females atleast 21 years old of Pima Indian heritage. The objective is to predict if the patient is diabetic or not and it is also to diagnostically predict whether a patient has diabetes based on certain diagnostic measurements which includes number of pregnancies, glucose level in the blood, blood pressure, thickness of the skin, insulin level in blood, body mass index, diabetes percentage, age. The outcome for this project is to identify the patients who are affected from diabetes. It is predicted by comparing with 4 parameters respectively (i.e) age, blood pressure, insulin and diabetespedigree function and the final result will be termed as 0 and 1. The term 0 denotes the patient is not affected from diabetes and the term 1 denotes the patient is affected from diabetes. Several constraints were placed on the selection of these instance from a large dataset.

**OBJECTIVE:**

1. To predict based on diagnostic measurements whether a patient has diabetes.
2. To predict the relationship between the glucose, insulin, skinthickness, diabetes pedigree function, bmi and blood pressure which is the major reason for the cause of diabetes.
3. To analyse the factor which causes diabetes by using one way anova method.

**DATA SOURCE:**

The data is downloaded from the website kaggle

Link**:** [**https://www.kaggle.com/datasets/mathchi/diabetes-data-set**](https://www.kaggle.com/datasets/mathchi/diabetes-data-set)

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases.

**METHODS USED:**

1. **ONE WAY ANOVA –** itis used to determine whether there are any statistically significant differences between the means of three or more independent (unrelated) groups.
2. **Time series –** it isto any group of statistical information collected at regular intervals of time. Time series analysis is used to detect the changes in patterns in these collected data.
3. **Mean -** The mean is calculated by adding the value of each individual item in a group and dividing it by the total number of items in the group
4. **Correlation** - itto denote association between two quantitative variables.
5. **Regression -**  itdescribes the average relationship existing between twovariables.
6. **Kruskal wallis test -** This test determines whether the [medians](https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/mean-median-mode/#median) of two or more groups are different.
7. **Spearman’s rank correlation coefficient –** it is astatistical measure that determines how closely two variables fluctuate.
8. **P chart –** itis a type of [control chart](https://en.wikipedia.org/wiki/Control_chart) used to monitor the proportion of [nonconforming units](https://en.wikipedia.org/wiki/Nonconformity_(quality)) in a [sample](https://en.wikipedia.org/wiki/Sample_(statistics)), where the sample proportion nonconforming is defined as the ratio of the number of nonconforming units to the sample size.
9. **Np – chart -** It is generally used to monitor the number of non-conforming or defective items in the measurement process.
10. **Barplot -** A bar chart is used when you want to show a distribution of data points or perform a comparison of metric values across different subgroups of your data.

**CODE:**

library(qcc)

library(ggplot2)

library(plotly)

library(dygraphs)

df<-read.csv(file="D:\\3rd sem\\business statistics using R\\diabetes.csv")

print(df)

y1<-c(148,85,183,89,137,116,78,115,197,125,110,168,139,189,166,100,118,107,103,115)

y2<-c(72,66,64,66,40,74,50,70,96,95,74,80,60,72,84,74,30,70,88,84)

name<-c("glucose","blood pressure")

f<-c(y1,y2)

n<-rep(2,20)

group = rep(1:20, n)

data = data.frame(y = f, group = factor(group))

fit = lm(y ~ group, data)

print(anova(fit))

print("Mean of glucose")

print(mean(df$Glucose))

print("Mean of Insulin")

print(mean(df$Insulin))

print("Mean of blood pressure")

print(mean(df$BloodPressure))

print("Mean of BMI")

print(mean(df$BMI))

print("Mean of diabetes pedigree function")

print(mean(df$DiabetesPedigreeFunction))

print("Mean of age")

print(mean(df$Age))

correlation=cor(df$Insulin,df$Glucose, method="pearson")

print(correlation)

regression=lm(df$Glucose~df$Insulin)

print(regression)

t=ts(df)

plot(t,xlab="Diabetic rate",ylab="age")

k=kruskal.test(Insulin~BMI,data=df)

print(k)

r<-cor.test(x=df$Glucose,y=df$Insulin,method="spearman")

print(r)

sample\_size <- as.integer(rnorm(50, 20, 3))

p1<-c(148,85,183,89,137,116,78,115,197,125,110,168,139,189,166,100,118,107,103,115)

p2<-c(72,66,64,66,40,74,50,70,96,95,74,80,60,72,84,74,30,70,88,84)

p\_chart <- with(df, qcc( sample\_size,sample\_size,type = "p", data.name = "p\_chart"))

summary(p\_chart)

np\_chart <- with(df, qcc( sample\_size,sample\_size,type = "np", data.name = "np\_chart"))

summary(np\_chart)

barplot(df$Glucose,

main = "Glucose level",

col = "darkred",

horiz = FALSE)

barplot(df$Insulin,

main = "Insulin level",

col = "green",

horiz = TRUE)

**OUTPUT:**

**REFERENCE LINK:**

1. [**https://chartio.com/learn/charts/bar-chart-complete-guide/**](https://chartio.com/learn/charts/bar-chart-complete-guide/)
2. [**https://www.kaggle.com/datasets/mathchi/diabetes-data-set**](https://www.kaggle.com/datasets/mathchi/diabetes-data-set)
3. [**https://www.simplilearn.com/tutorials/statistics-tutorial/spearmans-rank-correlation#:~:text=Spearman's%20rank%20correlation%20measures%20the,represented%20using%20a%20monotonic%20function**](https://www.simplilearn.com/tutorials/statistics-tutorial/spearmans-rank-correlation#:~:text=Spearman's%20rank%20correlation%20measures%20the,represented%20using%20a%20monotonic%20function)**.**
4. [**https://sciencing.com/uses-mean-median-mode-6323388.html**](https://sciencing.com/uses-mean-median-mode-6323388.html)