IT3051 – Fundamentals of Data Mining

Data Science

Faculty of Computing

Sri Lanka Institute of Information Technology

Practical 10

(01) Create the following data frame in R

Name	Age	Gender	Marks
Sam	23	Male	78
Kane	21	Male	58
Jane	24	Female	30
Anne	25	Female	85
Sammie	20	Female	90

a) Remove the Primary Key column in the dataset.
b) Change the categorical variables to factors.
c) Find the average mark for a student.
d) Add 5 marks for each student and find the average mark again.
e) Change the Age of Jane (3rd Row) to 22.
f) Create a new column for results such that,
    a. If mark ≥ 50 ---> Pass
    b. If mark < 50 ---> Fail
g) Get the overall summary of the data frame.
h) Separate the data frame to two data frames based on Gender (df_Male & df_Female)
i) Get statistical summary for each df_Female & df_Male data frames.

(02) Import the Boston inbuilt dataset to the R environment and do the following tasks.

a) Fit a simple linear regression model for the response variable **medv** using **lstat** independent varaiable and save as **fit1**.

b) Fit a ~~simple~~ Multi linear regression model for the response variable **medv** using **lstat** & **black** independent varaiables and save as **fit2**.

c) Fit a multiple linear regression model for the response variable **medv** using all other independent variables and save as **fit3**.

d) Fit a multiple linear regression model for the response variable **medv** using all other independent variables except **indus** variable and save as **fit4**.

e) Consider the **fit3** and get the following charts

    a. Residuals VS Fitted Value

    b. Normality plot of Standard Residuals

    c. Standardized Residuals VS Fitted Value

    d. Residuals VS Leverage

f) Check the Variance Inflation Factor (VIF) and discuss the multicollinearity among the independent variables.

g) Split the data into train & test such that 80% will be the training data.

    a. Fit the full model to the train data and discuss the significance of variables.

    b. Get the predictions to the test data.

    c. Find the Mean Squared Error and then find the Root Mean Squared Error