भारतीय सूचना प्रौद्योगिकी संस्थान गुवाहाटी

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY GUWAHATI

Data Analytics Lab, M.Tech 3rd Semester

Instructions

1. Upload all your codes to Github.
2. You will be called randomly to explain the code based on which marks/grade will be assigned.

Assignment -2

1. Perform the following pre-processing tasks.
   (a) Get the dataset and read it using **read.csv** function. Dataset contains four columns: Country, Age, Salary, Purchased and has total 10 observations. Dataset contains information of customers of some company and first three columns are information of a customer like country, age, salary and fourth column tells if the customer has purchased the product of the company or not. The attribute "Purchase" is the target variable.
   (b) The dataset contains missing data. Handle the missing data by taking mean/median/mode
   (c) Encode categorical data. Encode "France", "Spain" and "Germany" with 1, 2, 3 values. Also encode the target variable with 0 and 1 i.e. Encode Yes with 0 and No with 1. **factor** function or **if-else** statement can be used to convert the categorical features into numeric features.
   (d) Scale the dataset using any of the scaling technique. Use **scale** function to scale the values.
2. Perform the following aggregation tasks.
   (a) Get the dataset from UCI ML Repository: https://archive.ics.uci.edu/ml/datasets/iris.
   (b) Find the mean of all the metrics (Sepal.Length Sepal.Width Petal.Length Petal.Width)
   (c) Compute the sum of all the metrics across species and group by species.
   (d) Compute the count of all the metrics across species and group by species.
   (e) Compute the maximum of all the metrics across species and group by species.
   Note: The aggregate function has the following syntax aggregate(x, by, FUN, ..., simplify = TRUE, drop = TRUE)
   X is an R object, Mostly a dataframe
   by :- a list of grouping elements, by which the subsets are grouped by
   FUN :- a function to compute the summary statistics
   simplify:- a logical indicating whether results should be simplified to a vector or matrix if possible
   drop:- a logical indicating whether to drop unused combinations of grouping values.