Working With Data : Database Assignment

Due Date
18th December @11pm

Part A – Data Analytics [30 marks]

Select a '*interesting*' data set from Kaggle or from some other data set repository. You will have used many different data sets throughout your various courses. One of these can also be used.

Load this data set into a table(s) in your Oracle Schema. Document the steps you take and provide the code.

After loading the data set into an Oracle table, use a variety of SQL statistical functions to analyse this data set.

Provide ten SQL queries that use statistical SQL functions, along with a description explaining why you used that function, what you wanted to achieve by using this statistical function, and an explanation of results and what they mean.

Include the outputs for these queries, limited to the first 20 records (built into your SQL query).

You can use any of the 350+ statistical functions available in SQL, except for the following,

AVG(): returns the mean
COUNT(): returns the population (or sample, depending on the row source)
SQUARE(n): returns the square of the value specified
POWER(a,n): returns the value of a to the nth power
SQRT(n): returns the square root of n
SUM(): returns the sum of the values in a set
STDDEV(): returns the standard deviation of a sample
STDDEVP(): returns the standard deviation of a population
VAR(): returns the variance of a sample
VARP(): returns the variance of a population
CORR(): correlation analysis
MIN(): minimum value
MAX(): maximum value

No marks will be awarded if the above SQL functions are used.

Part B – Data Audit Report using PL/SQL and SQL Statistical Functions (40%)

For this part of the assignment you will need to use the Portuguese Bank data set (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing). You will need to download this data set.

Load this data set into a table in your Oracle Schema. Document the steps taken to load the data set into your Oracle schema and provide the code. Your code should be documented.

Write a PL/SQL Procedure to perform a data audit of this analytic record. You will need to use a variety of statistical functions, and may require slightly different SQL statistical functions to be used based on the data type, for example VARCHAR2, NUMBER, DATE.

The output from the analysis from the data audit can be stored in a table.

Deliverables for this part of assignment.

- Details of how the data set was loaded into the tables.
- Design for the code
- PL/SQL code used to perform the data audit.
- Output from the data audit table formatted to make it readable

IMPORTANT: you are **not** allowed to use the DBMS_STAT_FUNCS.SUMMARY() function in PL/SQL

Part C – Machine Learning using SQL (30%)

For this part of the assignment you will create three machine learning models using the in-database machine learning features. Use the Portuguese Bank data set (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing) for this part of the assignment This is a Classification data mining problem.

NOTE : Decision Trees machine learning algorithm should not be used. If used, no marks will be allocated.

You may need to create additional tables, views, and use sampling to prepare the training and test data sets.
You should create machine learning models using at least four algorithms.
You may need to create a view that contains the predicted values for the testing data set.

Complete this process by evaluating the accuracy of the models.

Marking Scheme

The marking scheme for this assignment is:

- 30% Part A – Data Analysis
- 40% Part B – PL/SQL Procedure
- 30% Part D – Machine Learning Using SQL

IMPORTANT:

- Your report should not consist solely of code, images and/or screen shots. Explanations should be given in addition to documented code.
- All code for a section/part of the assignment should be in that section and not in a separate appendix in a different part of the report.
- Everything should be in one PDF document, organised to have sections for each part of the assignment.

Your assignment report should contain all the workings for each component of the assignment. This report should be converted into PDF format. This file, along with any additional files requested, should be ZIPPED (**not** RAR, TAR or any other compressing formats.

The ZIPPED file should be in ZIP format, and **not** in RAR, TAR or any other compression formats.

The documentation for your assignment must contain

- Full name
- Student number
- Class code (TU??)
- Stream (DS?ASD/etc)
- Assignment Title/Description

Failure to give this information will incur a 10% penalty.

The assignment most be performed **individually.**

Assignments to be submitted on BrightSpace.

- **NO email submissions will be accepted**
-

Each submission must be original work as **plagiarism** will result in a **zero** mark (0%). You should make yourself familiar with the plagiarism policy of Technological University Dublin.

There will be a 10% penalty for each day the assignment is late.

There is no penalty for submitting early.