TU59/TU60 - Data Mining Assignment **Due Before : Wednesday 16th December @23:00**

Group Assignment: This assignment can be completed individually or in groups of 2 people. **Students are responsible for the formation of groups.**

Introduction & Overview

This data mining assignment requires you to analyse a data set, identify data insights, build and evaluate a number of data mining models and to compare the performance of these models with previously published work on the same data set.

The data set and problem is from a well know (and discussed) problem and the data set is available for download from the UCI ML Repository

https://archive.ics.uci.edu/ml/datasets/Bank+Marketing

The data set is related to a direct marketing campaign for a Portuguese banking institution. The bank conducts marketing campaigns and uses their call center to contact their customers using phone calls.

The purpose of this data mining project is to identify customers who are most likely to subscribe to a term deposit account based on previous marketing campaigns.

The data sets provided contain the results of the previous marketing campaigns. Use **bank-additional.zip**

The website (given above) contains the data files and attribute descriptions of the data set.

NOTE : There are two versions of the data set given/ One of the data sets contain 17 input attributes. The second data set contains 20 input attributes. It is this data set (20 attributes) that you should use for the assignment. The files for this data set are called 'bank-additional' (see bank-additional-full.zip).

The data sets contain two files. the 'bank-additional.csv' contains a 10% sample (4,119 records) of the entire data set. the entire data set is included in the file called 'bank-additional-full.csv' and contains **41,188 records**. Use this data set.

A published paper is available that describes the project, the data set and what data mining methods were applied. Here is the full citation for the paper and a link to where you can find it. You should study and read this paper carefully.

S. Moro, P. Cortez and P. Rita. *A Data-Driven Approach to Predict the Success of Bank Telemarketing*, Decision Support Systems, Elsevier, 62:22-31, June 2014 https://core.ac.uk/download/pdf/55631291.pdf [also see paper attached to assignment in VLE]

You will be required to build a number of data mining models, using SAS Enterprise Miner, and compare the results that you generate with those that were generated by the original researchers.

Required Tasks

You are required to produce a report detailing your work investigating the data, building classification models, analysing the results, and comparing your results with the original findings.

The first task you should complete is a data investigation exercise, where you will document the characteristics and other information that you can determine about each Feature. Identify any data insights discovered and detail all data preparation tasks and any decisions made. This work can be completed using SAS software, and any additional data preparation and exploration can be completed using SAS Software, R, Python and Go Lang.

You will need to work through/develop a number of classification models. To do this you need to use the data mining tool used in class (SAS Enterprise Miner). In this tool you can have a number of different classification techniques and within each of these you can modify the various parameter settings.

You will need to evaluate the results from each of the models to determine which of the models gives the best results for you. You can then compare your results with the original research and discuss the outcomes.

TU59/TU60 - Data Mining Assignment **Due Before : Wednesday 16th December @23:00**

The original research project used a certain number of data mining algorithms. For your assignment you should use all the algorithms that are available to you in SAS Enterprise Miner.

Deliverables

You will be required to document your approach to solving and evaluating this classification problem, based on the CRISP-DM process and documentation template guide.
Your report will probably be between 16-20 pages long. The maximum length of the report should be 20 pages.

The report should clearly show your work in the following areas (similar to CRISP-DM):

- Definition of problem
- Data Exploration and Descriptive Analytics
- Identification of data insights from previous step
- Details of any additional data preparation (cleaning, transformations, etc), data enrichment, feature engineering, feature reduction, etc
- Details of each data mining algorithm used, the configuration settings used, etc
- Details of the evlautation and performance measures from your data mining models. Examine which one performed best, why this might have been the case and how the results compare across all the models
- Discussion of of how your results compared to the results from the original research and any conculusions that you can draw from this comparison

Submission Details

The assignment is due by Wednesday 16th December @23:00

You should create one document/report containing all the material for each part of the assignment. Convert this document into a PDF. It is this PDF document that should be submitted.
All images should be imbedded in this document.

Maximum page limit of 20 pages for your report.

You will need to submit your assignment on **BrightSpace VLE**. You cannot submit your assignment via email.

Marking Scheme

The marking scheme for this assignment is:

- 25% Problem Definition, Descriptive Analytics, Data Insights, etc & summary of initial findings/insights
- 15% Details of any additional data preparation, data enrichment, feature engineering, feature reduction, etc
- 15% Details of each data mining algorithm used, the configuration settings used, etc
- 20% Details & Discussion of the evaluation and performance measures from your data mining models.
- 25% Discussion of how your results compared to the results from the original research and any conclusions that you can draw from this comparison

The documentation for your assignment must contain the name, student number, class, course (**TU??**) and year information for each student in the group. **Failure to give this information will incur a 10% penalty.**

This assignment can be completed individually or in groups of 2 people. **Students are responsible for the formation of groups.**

Each submission must be original work as plagiarism will result in a **zero** mark (0%).
There will be a 10% penalty deduction will be applied for each day the assignment is late.
There is no penalty for submitting early.
DIT Plagiarism Policy : https://tudublin.libguides.com/c.php?g=674049&p=4794713
https://www.tudublinsu.ie/advice/exams/breachesofregulations/

Assignment Feedback

TU59/TU60 - Data Mining Assignment **Due Before :**
Wednesday 16th December @23:00

I will endeavor to mark the assignments before Christmas. Feedback will be via Brightspace VLE, where a mark for your assignment will be given and a short comment on the assignment. If feedback isn't available before Christmas, it will be available in early-mid January.