TU059, TU060, DT9000 - Programming for Big Data - Spark

Due Before : (Week12)
Monday 24th April @23:00

Overview

Select a machine learning competition from Kaggle (https://www.kaggle.com/competitions) - it can be an old competition (excluding the Titanic competition) - and submit an entry. Then write a short report, **maximum 10 pages** plus code, outlining your approach.

Complete the typical life-cycle consisting of:

- Problem Definition, Kaggle Competition selected and why.
- Data Exploration, Feature Engineering and Data Preparation. Include explanation on all decisions made.
- Create models. You will create several models to determine which one works best for you.
- Model Evaluation and Explanations of Decision.
- Discussion of work completed, highlighting features of Spark used in your work.
- Kaggle Competition entry results/outcome, and evaluation of this.

Warning: All work **must** be completed using the Python Spark APIs (version 3). Other libraries/packages like Scikit-Learn should not be used.

Submission Details

The assignment is due by Monday 24th April @23:00.

You should create **one document/report** containing **all** the material for each item listed above and all code. **Convert this document into a PDF**. It is this PDF document that should be submitted.
All images should be imbedded in this document. This PDF document should contain all your code.

In addition to the report include:

- all code (to be included in PDF document).
- links to competition, data sets, etc

Submit your assignment on Brightspace before the deadline. Make sure that you submit your assignment to the correct assignment on BrightSpace. You should receive an acknowledgment from Brightspace when you submit your assignment. You should keep a copy of this acknowledgement.

Marking Scheme

The marking scheme for this assignment is:

- 5% Problem Definition, Kaggle Competition selected and why
- 25% Data Exploration, Feature Engineering and Data Preparation. Include explanation on all decisions made.
- 25% Create models. Create several models to determine which one works best for you.
- 25% Model Evaluation and Explanations of Decision
- 20% Discussion of work completed, competition outcomes, and highlighting features of Spark used in your work.

TU059, TU060, DT9000 - Programming for Big Data - Spark

Due Before : (Week12)
Monday 24th April @23:00

This assignment accounts for **40%** of your module mark for Programming for Big Data module.

The documentation for your assignment must clear contain the following details

- Full Name
- Student Number
- Programme Code (e.g. TU59, TU60, etc)
- Stream (ASD, DS, PhD, etc)
- Year of Programme (1st Year, 2nd Year)
- Module Name

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

Important Information

The assignment most be performed **individually.**

Each submission must be original work as **plagiarism** will result in a **zero** mark (0%).

Each submission should include a signed Plagiarism Cover Page (see Brightspace).

All code should be included. If not included your assignment mark will be reduced by 50%.

There will be a 10% penalty deduction applied for each day the assignment is late.
There is no penalty for submitting early.

TU Dublin **Plagiarism Policy** : https://tudublin.libguides.com/c.php?g=674049&p=4794713
https://www.tudublinsu.ie/advice/exams/breachesofregulations/

Ensure you are compliant with TU Dublin policy on usage of Large Language Models (LLMs), (e.g. ChatGPT, etc) and other similar tools and software.

Assignment Feedback

Feedback will be via Brightspace, where a mark for your assignment will be given and a short comment on the assignment. I hope to have this back to you approx. two weeks after the due date.