CSCI316 – Big Data Mining Techniques and Implementation
Group Assignments
2023 Session 3 (SIM)

10 Marks
Deadline: Refer to the submission link of assignments on Moodle

<u>One task</u> is included in each assignment. The specification of the task(s) starts in a separate page.

You must implement and run all your Python code in Jupyter Notebook. ***The deliverables are project presentation slides and source code.***

All results of your implementation must be reproducible from your submitted Jupyter notebook source files. In addition, the submission must include all execution outputs as well as clear explanation of your implementation algorithms (e.g., in the Markdown format or as comments in your Python codes).

Submission must be done online by using the correct submission link for this subject on MOODLE.

This is a group assignment. Only one submission per group. State the names and student numbers of group members at the beginning of each submitted file.

Marking guidelines:

Correctness of source code, and completeness and clearness of the project presentation.

Assignment 1

(10 marks)
Dataset: Loan data set for credit risk analysis
(https://www.kaggle.com/datasets/rameshmehta/credit-risk-analysis)
This data set has different types of features such as categorical, numeric & date. The target variable is the default (index). In financing, a default can occur when a borrower is unable to make timely payments, misses payments, avoids or stops making payments. An explanation of the features in the appendix of this document.

Objective
The objective of this task is to develop an end-to-end data mining project by using the Python machine learning library ***Scikit-Learn***. Only the Scikit-Learn library can be used in this task. However, all non-ML libraries (e.g., SciPy) are allowed.

Requirements
(1) This is a *classification* problem.
(2) Use 80% data for training and 20% for testing. Stratified sampling *must* be used.
(3) Main steps of the project should be (a) "discover and visualise the data", (b) "prepare the data for machine learning algorithms", (c) "select and train models", (d) "fine-tune the model" and (e) "evaluate the outcomes". You can structure the project in your own way. Some steps can be performed more than once.
(4) In the steps (c) and (d) above, you must work with at least three machine learning algorithms.
(5) In step (b), define at least one new feature by using the User-Defined Transformer. This transformer includes a parameter indicating whether use the new feature(s) or not. In step (d), fine-tuning step must use this parameter (as a hyper parameter).
(6) Explanation of each step together with the Python codes must be included.
(7) A comparison of the models' performance must be included.

Deliverables
Deliverables include (1) a project presentation* and (2) a submission including the following files:

- the Jupiter Notebook source code,
- a PDF document generated from your Jupiter Notebook source code, and
- the presentation slides.

*Note: The project presentation is announced by your tutorial teacher.

Assignment 2

(10 marks)
UNSW Network Intrusion Dataset (UNSW_NB15_training-set.csv, UNSW_NB15_testing-set.csv)
https://research.unsw.edu.au/projects/unsw-nb15-dataset

Several datasets are available for model development and model testing for IDS. This project will utilize the UNSW-NB15 dataset. The UNSW-NB15 dataset is published by Cyber Range Lab of the Australian Centre for Cyber Security. The data was collected over 15 hours by an IXIA traffic generator in 2014, then pre-processed and labelled as "normal" and various types of "attack". Download the *training dataset* and the *test dataset* from the above link. The task is to predict whether a record represents "normal" or "attack" (a binary classification problem). Note that the last two columns represent the targe variables, which should not be used as training features.

Objective
The objective of this task is to develop an end-to-end data mining project by using the Python machine learning library ***Spark MLlib***. Only the Spark MLlib can be used in this task. However, all non-ML libraries (e.g., SciPy) are allowed.

Requirements
(1) This is a *multi-classification* problem.
(2) Use a data in UNSW_NB15_training-set.csv for training and data in UNSW_NB15_testing-set.csv for testing.
(3) Main steps of the project should be (a) "discover and visualise the data", (b) "prepare the data for machine learning algorithms", (c) "select and train models", (d) "fine-tune the models" and (e) "evaluate the outcomes". You can structure the project in your own way. Some steps can be performed more than once.
(4) In the steps (c) and (d) above, you must work with at least three machine learning algorithms.
(5) Explanation of each step together with the Python codes must be included.
(6) A comparison of the models' performance must be included.
(7) Based on your experience in the assignments, write a brief report that compares Spark MLlib and Scikit-Learn (e.g., their pros/cons or similarity/difference).

Deliverables
Deliverables include (1) a project presentation* and (2) a submission including the following files:
- the Jupiter Notebook source code,
- a PDF document generated from your Jupiter Notebook source code, and
- the presentation slides.

*Note: The project presentation is announced by your tutorial teacher.