CSCI316 – Big Data Mining Techniques and Implementation
Group Assignments 1 and 2
2024 Session 1 (SIM)

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

One task 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: Credit score data set
(https://www.kaggle.com/datasets/parisrohan/credit-score-classification)

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 stratified sampling to select 80% data for training and 20% for testing.
(3) Main steps of the project are (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.

The assessment is based on the correctness and quality of your project. You must not copy any code from any public source directly.

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)
Dataset: Credit score data set
(The same as in Assignment 1.)

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 *classification* problem.
(2) Use stratified sampling to select 80% data for training and 20% for testing.
(3) Main steps of the project are (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).

The assessment is based on the correctness and quality of your project. You must not copy any code from any public source directly.

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.