CSCI316 – Big Data Mining Techniques and Implementation
Group Assignment
2025 Session 3 (SIM)

20 Marks
Deadline: Refer to the submission link on Moodle

Two tasks are included. 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.

Task 1

(10 marks)
Dataset: Apartment Rent Dataset
Source: https://www.kaggle.com/datasets/shashanks1202/apartment-rent-data
The target variable is "price". (Important note: Another attribute named "price_display" also contains the values of the target variable. You must remove this attribute during training.)

Objective
The objective of this task is implementing an end-to-end data mining project by using the Python machine learning library ***Scikit-Learn*** to predict the rental price. Refer to the above link for more information about the dataset.

Requirements
(1) 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. (Note. some steps can be performed more than once.)
(2) In the steps (c) and (d) above, you must work with at least three machine learning algorithms.
(3) 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).
(4) Use 80% data for training and 20% for test.
(5) Explanation of each step together with the Python codes must be included.
(6) 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.

This task must be completed by you and your group mates. You must not copy any code from existing sources from the internet.

*The project presentation is announced separately.

Task 2

(10 marks)
Dataset: Apartment Rent Dataset
Same as Task 1.

Objective
The objective of this task is to implement a 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 are allowed. The task is to predict the rental price.

Requirements
(1) Sample 80% data for training and 20% for testing.
(2) 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. (Note. some steps can be performed more than once.)
(3) In step (b), define at least one new feature by using the RFormula.
(4) In the steps (b) and (c) above, you must work with at least three machine learning algorithms.
(5) Use 80% data for training and 20% for test.
(6) Explanation of each step together with the Python codes must be included.
(7) A comparison of the models' performance must be included.
(8) 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.

This task must be completed by you and your group mates. You must not copy any code from existing sources from the internet.

*The project presentation is announced by separately.