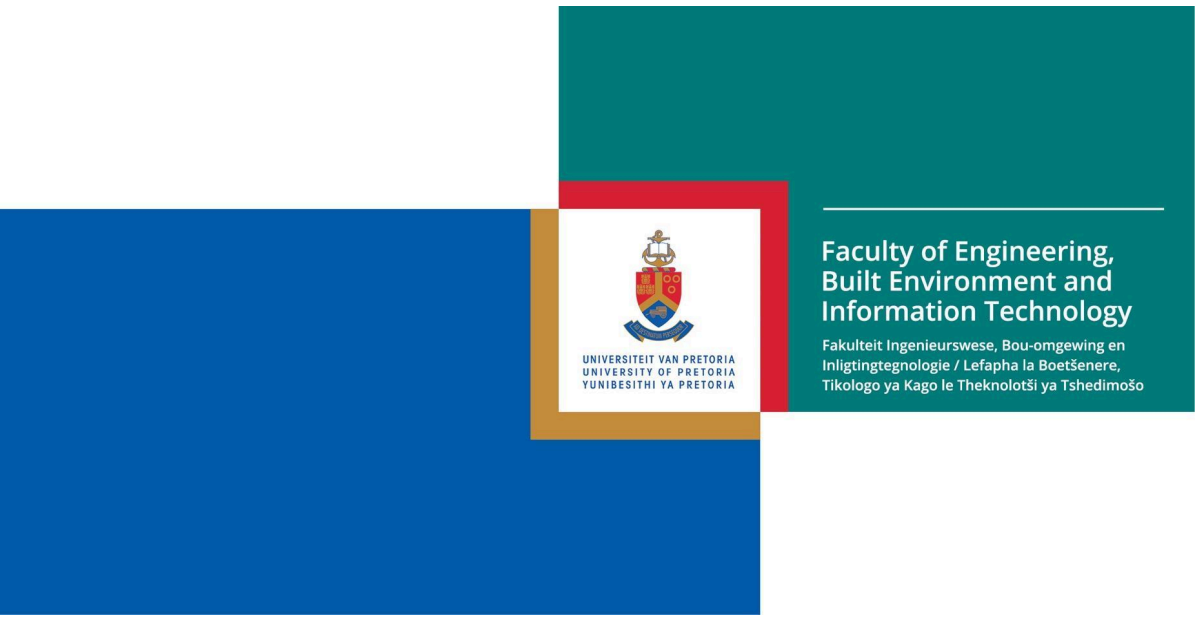

School of Information Technology

Department of Computer Science

COS781-2024

Course Final Project

Lecturer Prof. Vukosi Marivate, Dr Abiodun Modupe, Seani Rananga

Last Revision: 19 September2024 (Version 1.0)

Deadline: 7 October and 6 November 2024

1. Project Expectations

One of the main goals of COS 781 is to get you comfortable with Data Mining and moving towards using state of the art approaches in Data Mining to solve real-world problems. For 2024 this is looking at wider applications of methods for Data Mining. If you have an interest in the future in doing research in this area, COS 781 also prepares you for this. The project is the main output of this class for you and as such do take time working on it. The project will be done **individually**.

Your first task will be choosing a project topic. For 2024, you will first bid on a dataset and or problem combo. Datasets have been made available on the course ClickUp. It is encouraged you use this to get your topics.

You can have 2 types of approaches to a project :

- Experimental evaluation of algorithms and models on an interesting dataset;
- A theoretical project that considers a model, an algorithm or a property (measure) and derives a rigorous result about it;

A typical project will have a mix of the above.

What makes an interesting project?

- Given the topic, you ask questions that may lead to answers that get us to better understand the problem, problem area or the world.
- Provides an extension or possible extension to the current state of the art approaches in the area.

2. Deliverables

Overview

Deliverable	Score/Points
Choose a Data Set that You Will work on	-
Project Proposal - Topic and Short Description(with research questions) ***[7 October 2024]***	10
Project Report ***[11 November 2024]*** **4 pages KDD Format.**	50
Project Presentations ***[6 November 2024]***	30
Project Files with Documentation ***[11 Nov 2024]***	10
Total	**100**

Project Proposal [1 page, Arial size 11]

The proposal should answer the following questions:

- **Research Questions:**
  - What is the problem you are solving?
  - Why is it interesting?
- **Data:**
  - What data will you be using?
  - How big is the data? Attributes?
  - **Note:** You can use your own data, but make sure you have it by the time the proposal is submitted.
- **Approach:**
  - What methods, algorithms, techniques will you be using? [Be specific]
  - What do you expect from them?
- **Evaluation:**
  - How will you measure success?
  - Are there baselines?
- **Expected Outputs:**
  - What do you expect your outputs to be at the end of the semester?

Project Report

Format

Template - KDD Explorations https://www.kdd.org/author-instructions
Length: 4 pages.

Expected Sections

1. Exploratory Data Analysis (EDA) (20 Marks)

- **Data Inspection:** Provide an initial overview of the dataset. This should include information about the features (variables), their types (categorical or numerical), missing values, and basic statistics.
- **Visualisations:** Include appropriate visualisations (e.g., histograms, box plots, pair plots) to summarise the data and highlight any interesting patterns or insights.
- **Insights:** Discuss any patterns or trends you observe during the EDA phase. Identify any relationships between the variables that may be relevant for further analysis.

2. Data Preprocessing (20 Marks)

- **Handling Missing Data:** Apply appropriate methods to handle any missing data in the dataset.
- **Feature Engineering:** Create new features or modify existing features that may be useful for clustering or dimensionality reduction. This could include encoding categorical variables or creating new interaction terms.
- **Standardisation/Normalisation:** Apply scaling to numerical variables if necessary (for example, when using distance-based algorithms).

3. Data Mining Methods and Analysis (40 Marks)

- **Choose appropriate methods for your problem and apply them to the data.**
- **Discussion**

4. Conclusion and Reflection

Project Presentation [7 slides, 3-4 minutes]

This is an opportunity to share your work with your other classmates.

3. Resources

See the Dataset Document

4. Rubrics

Available on Clickup.

5. Submission Instructions and Deadline

Submit your report and declaration of originality, in **pdf format**, on the course ClickUP by the due date. No submissions will be allowed after the due date.

6. Plagiarism

This department considers plagiarism to be a serious offence. Disciplinary action will be taken against students who commit plagiarism. For more information on plagiarism, please refer to http://www.library.up.ac.za/plagiarism/index.htm.

Plagiarism is a serious form of academic misconduct. It involves both appropriating someone else's work and passing it off as one's own work afterwards. Thus, you commit plagiarism when you present someone else's written or creative work (words, images, ideas, opinions, discoveries, artwork, music, recordings, computer-generated work, etc.) as your own.

7. Generative AI Statement

This assignment has been designed to promote your learning, critical thinking, skills, and intellectual development without reliance on unauthorised technology including chatbots and other forms of "artificial intelligence" (AI). Although you may use search engines, spell-check, and simple grammar-check in crafting your report, you will be asked to submit your written work with the following statement. **"I certify that this assignment represents my own work. I have not used any unauthorised or unacknowledged assistance or sources in completing it including free or commercial systems or services offered on the internet or text generating systems embedded into software."** Please consult with the lecturer if you have any questions about the permissible use of technology in this class.