UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

School of Information Technology

Department of Computer Science

COS781-2025

Course Final Project

Lecturers: Prof. Vukosi Marivate, Dr Abiodun Modupe, Will van Heerden

Last: 08 September, 2025 (Version 1.0)

Deadline:

Abstract submission:	25 **September, 2025, at 23:59PM**
Final Project Report:	**14 November,** 2025 at 23:59 PM
Project Presentations & Documentation:	**21 November 2025** at 23:59 PM

1. Project Expectations

The goal of **the COS 781** project is to help students become comfortable with **data mining** by moving toward using **state-of-the-art** (SOTA) approaches to solve real-world problems. In this module, the student has the opportunity to explore the broader applications of data mining methods by leveraging a published paper that aligns with the syllabus covered in the provided **study guide.**

You have been assigned to groups, and each group has been allocated a research paper to read. Each group will utilise the methodology proposed in their assigned research paper, along with any datasets provided in Section 3, to replicate the study presented in the paper. You are required to submit an abstract of your idea about the paper by September 20, 2025, at 23:59 on ClickUp. Then **replicate** the research paper using the selected dataset and submit 4 pages of your implementation, including the abstract, introduction, literature review, methodology, dataset, result and conclusion of your **idea** in the KDD format by the submission deadline.

The key requirements for this project are summarised below:

- Read the research paper that has been assigned to your group along with the appropriate dataset in Section 3
- Submit a 1-page abstract of your idea based on the research paper and the dataset by September 20, 2025, at 23:59 on ClickUP
- Replicate the methodology in the research paper with the provided dataset and conduct an exploratory data analysis (EDA) to illustrate the dataset's patterns and distributions. Reproduce the experimental setup of the paper and report the evaluation results of the model. You may also propose an alternative approach to address the research question in the research paper, provided that your results still align with the paper's objectives and contributions."

2. Deliverables

Overview

Deliverable	**Score/Points**
Choose a Data Set that You Will work on	-
Initial Submission of Abstract of your Approach **[25 September, 2025]**	10
Project Report **[14 November 2025]** **4 pages in KDD format.**	50

Project Presentations **[21 November 2025]**	30
Project Files with Documentation **[21 November 2025]**	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?

3. Datasets

Lacuna Fund Datasets

The Lacuna Fund has a number of African datasets covering agriculture, languages, and health. You can look at the datasets and choose one you would like to work on

Agriculture

Lacuna Fund agriculture datasets unlock the power of machine learning to alleviate food security challenges, spur economic opportunities, and give researchers, farmers, communities, and policymakers access to superior agricultural datasets. Learn more and download released datasets below.

- Dataset(s): https://lacunafund.org/datasets/agriculture/

Example of a dataset included that would be good fit for data mining

- Sensor-Based Aquaponics Fish Pond Datasets https://www.kaggle.com/datasets/ogbuokiriblessing/sensor-based-aquaponics-fish-pond-datasets

Language [Use NLP + other parts of the course]

Lacuna Fund language datasets provide free text and speech resources that support natural language processing technologies in various languages, especially in low- and middle-income countries around the world. Explore and download released datasets below.

- Dataset(s) https://lacunafund.org/datasets/language/

Health

Lacuna Fund health datasets reduce health disparities by helping providers and patients make decisions that lead to more equitable healthcare outcomes. These datasets can be used to train chatbots, provide reliable medical information to the public, assist with disease screening and diagnosis, and assess the health and treatment of large populations over time (e.g., maternal health or HIV data). Learn more and download released datasets below.

- Dataset(s) https://lacunafund.org/datasets/health/

Datasets For Recommender Systems

This is a repository of topic-centric public data sources in high quality for recommender systems (RS). They are collected and tidied from Stack Overflow, articles, recommender sites and academic experiments. Most of the datasets presented here are free, having open-source licences; however, some are not, and you need to ask permission to use or cite the authors' work.

- Dataset(s) https://github.com/caserec/Datasets-for-Recommender-Systems

Data Science for Social Impact Language Datasets

- HuggingFace Dataset(s): https://huggingface.co/dsfsi
- Zenodo Dataset(s) https://zenodo.org/communities/africanlp/records?q=&l=list&p=1&s=10&sort=newest [DSFSI and other community members of AfricaNLP on Zenodo]

eCommerce Data

- Instacart Market Basket Analysis https://www.kaggle.com/c/instacart-market-basket-analysis https://www.kaggle.com/retailrocket/ecommerce-dataset
- https://gist.github.com/entaroadun/1653794
- https://github.com/RUCAIBox/RecSysDatasets
- 30music / impresions / tv audience - https://recsys.deib.polimi.it/datasets/

- http://archive.ics.uci.edu/ml/datasets/KASANDR

From https://gist.github.com/entaroadun/1653794

Movies Recommendation:

- MovieLens - Movie Recommendation Data Sets http://www.grouplens.org/node/73
- Yahoo! - Movie, Music, and Images Ratings Data Sets http://webscope.sandbox.yahoo.com/catalog.php?datatype=r
- Jester - Movie Ratings Data Sets (Collaborative Filtering Dataset) http://www.ieor.berkeley.edu/~goldberg/jester-data/
- Cornell University - Movie-review data for use in sentiment-analysis experiments http://www.cs.cornell.edu/people/pabo/movie-review-data/

Music Recommendation:

- Last.fm - Music Recommendation Data Sets http://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/index.html
- Yahoo! - Movie, Music, and Images Ratings Data Sets http://webscope.sandbox.yahoo.com/catalog.php?datatype=r
- Audioscrobbler - Music Recommendation Data Sets http://www-etud.iro.umontreal.ca/~bergstrj/audioscrobbler_data.html
- Amazon - Audio CD recommendations http://131.193.40.52/data/

Books Recommendation:

- Institut für Informatik, Universität Freiburg - Book Ratings Data Sets http://www.informatik.uni-freiburg.de/~cziegler/BX/

Food Recommendation:

- Chicago Entree - Food Ratings Data Sets http://archive.ics.uci.edu/ml/datasets/Entree+Chicago+Recommendation+Data

Merchandise Recommendation:

- Amazon - Product Recommendation Data Sets http://131.193.40.52/data/

Healthcare Recommendation:

- Nursing Home - Provider Ratings Data Set http://data.medicare.gov/dataset/Nursing-Home-Compare-Provider-Ratings/mufm-vy8d
- Hospital Ratings - Survey of Patients Hospital Experiences http://data.medicare.gov/dataset/Survey-of-Patients-Hospital-Experiences-HCAHPS-/rj76-22dk

Dating Recommendation:

- www.libimseti.cz - Dating website recommendation (collaborative filtering) http://www.occamslab.com/petricek/data/

Scholarly Paper Recommendation:

- National University of Singapore - Scholarly Paper Recommendation http://www.comp.nus.edu.sg/~sugiyama/SchPaperRecData.html

Recommender systems Datasets

Link: https://github.com/RUCAIBox/RecSysDatasets

Shopping

- Amazon: Amazon Review Data includes reviews (ratings, text, helpfulness votes) and product metadata (descriptions, category information, price, brand, and image features), which includes a previous version in 2014 and an updated version in 2018. Our processed datasets are detailed here.
  - Amazon 2014: This dataset contains product reviews and metadata from Amazon, including 24 categories and 142.8 million reviews spanning May 1996 - July 2014.
  - Amazon 2018: This Dataset is an updated version of the Amazon review dataset released in 2014. The total number of reviews is 233.1 million and the number of categories is 29 (142.8 million and 24 in 2014) and current data includes reviews in the range May 1996 - Oct 2018.
  - Amazon 2023: This Dataset is the latest version of the Amazon review dataset released in 2014. The total number of reviews is 571.54 million and the number of categories is 33 and current data includes reviews in the range May 1996 - Sep 2023.
- Amazon_M2: This dataset is a collection of anonymized customer sessions containing products from six different locales: English, German, Japanese, French, Italian, and Spanish.
- Alibaba-iFashion: This dataset is a fashion outfit dataset collected from Alibaba online shopping systems in the paper POG. The items from each outfit are viewed as the items being recommended to users, where each item consists of attributes such as category and title.
- Epinions: This dataset was collected from Epinions.com, a popular online consumer review website. It contains trust relationships amongst users and spans more than a decade, from January 2001 to November 2013.
- Yelp: This dataset was collected from Yelp. The Yelp dataset is a subset of our businesses, reviews, and user data for use in personal, educational, and academic purposes. Starting from Yelp Challenge 2018 (the original link to this competition is not found and there will not be another round of Yelp Dataset Challenge), there are four versions of Yelp datasets in total and Yelp has also posted the dataset on

Kaggle, where you can also download a few earlier versions. Our processed 5 datasets are detailed here.

  - Yelp 2018: It is the first version of Yelp dataset released in Yelp Challenge 2018 including 5,261,669 reviews.
  - Yelp 2020: It is the second version of Yelp dataset released in 2020, including 8,021,122 reviews.
  - Yelp 2021: It is the first version of Yelp dataset released in 2021, including 8,635,403 reviews.
  - Yelp 2022: It is the latest version of Yelp dataset, which contains 908,915 tips by 1,987,897 users over 1.2 million business attributes like hours, parking, availability, and ambience aggregated check-ins over time for each of the 131,930 businesses.
  - Yelp-full: This is a combination dataset including four versions of yelp datasets mentioned above, where the duplicates are dropped and the number of total reviews is 28,908,240.
- Tmall: This dataset is provided by Ant Financial Services, using in the IJCAI16 contest.
- DIGINETICA: The dataset includes user sessions extracted from an e-commerce search engine logs, with anonymized user ids, hashed queries, hashed query terms, hashed product descriptions and meta-data, log-scaled prices, clicks, and purchases.
- YOOCHOOSE: This dataset was constructed by YOOCHOOSE GmbH to support participants in the RecSys Challenge 2015.
- Retailrocket: The data has been collected from a real-world ecommerce website. It is raw data, i.e. without any content transformations, however, all values are hashed due to confidential issues.
- Ta Feng: The dataset contains a Chinese grocery store transaction data from November 2000 to February 2001.

Advertising

- Criteo: This dataset was collected from Criteo, which consists of a portion of Criteo's traffic over a period of several days.

- Avazu: This dataset is used in Avazu CTR prediction contest.

- iPinYou: This dataset was provided by iPinYou, which contains all training datasets and leaderboard testing datasets of the three seasons iPinYou Global RTB(Real-Time Bidding) Bidding Algorithm Competition.

- AliEC: Ali_Display_Ad_Click is a dataset of click rate prediction about display Ad, which is displayed on the website of Taobao. The dataset is offered by the company of Alibaba.

Check-in

- Foursquare: This dataset contains check-ins in NYC and Tokyo collected for about 10 month. Each check-in is associated with its time stamp, its GPS coordinates and its semantic meaning.

- Gowalla: This dataset is from a location-based social networking website where users share their locations by checking-in, and contains a total of 6,442,890 check-ins of these users over the period of Feb. 2009 - Oct. 2010.

Movies

- MovieLens: GroupLens Research has collected and made available rating datasets from their movie website.
- Netflix: This is the official data set used in the Netflix Prize competition.
- Douban: Douban Movie is a Chinese website that allows Internet users to share their comments and viewpoints about movies. This dataset contains more than 2 million short comments of 28 movies in Douban Movie website.
- Twitch: This is a dataset of users consuming streaming content on Twitch. We retrieved all streamers, and all users connected in their respective chats, every 10 minutes during 43 days.
    - Twitch-100k: Twitch-100k is a subset of 100k users for benchmark purposes. The code is available in this Github repository.
    - Twitch-full: See the Google Drive folder containing all Twitch files. Twitch-full contains the full dataset while Twitch-100k is a subset.

Music

- Last.FM: This dataset contains social networking, tagging, and music artist listening information from a set of 2K users from Last.fm online music system.
- LFM-1b: This dataset contains more than one billion music listening events created by more than 120,000 users of Last.FM. Each listening event is characterized by artist, album, and track name, and includes a timestamp.
- Yahoo Music: This dataset represents a snapshot of the Yahoo! Music community's preferences for various musical artists.
- KGRec: Music and Sound Recommendation with Knowledge Graphs are two different datasets with users, items, implicit feedback interactions between users and items, item tags, and item text descriptions are provided, one for Music Recommendation (KGRec-music), and other for Sound Recommendation (KGRec-sound).
    - KGRec-music: All the data comes from `songfacts.com` and `last.fm` websites. Items are songs, which are described in terms of textual description extracted from `songfacts.com`, and tags from `last.fm`.
    - KGRec-sound: All the data comes from `Freesound.org`. Items are sounds, which are described in terms of textual description and tags created by the sound creator at uploading time.
- Music4All-Onion : The dataset expands the Music4All dataset by including 26 additional audio, video, and metadata characteristics for 109,269 music pieces.

Books

- Book-Crossing: This dataset was collected by Cai-Nicolas Ziegler in a 4-week crawl (August / September 2004) from the Book-Crossing community with kind permission from Ron Hornbaker, CTO of Humankind Systems. It contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books.

- GoodReads: This dataset contain reviews from the Goodreads book review website, and a variety of attributes describing the items. Critically, datasets have multiple levels of user interaction, raging from adding to a shelf, rating, and reading.

Games

- Steam: This dataset is reviews and game information from Steam, which contains 7,793,069 reviews, 2,567,538 users, and 32,135 games. In addition to the review text, the data also includes the users' play hours in each review.

Anime

- Anime: This dataset contains information on user preference data from MyAnimeList.net - Anime and Manga Database and Community. Each user is able to add anime to their completed list and give it a rating and this dataset is a compilation of those ratings.

Pictures

- Pinterest: This dataset is originally constructed by paper Learning image and user features for recommendations in social networks for evaluating content-based image recommendation, and processed by paper Neural Collaborative Filtering.

Jokes

- Jester: This dataset contains anonymous ratings of jokes by users of the Jester Joke Recommender System.

Exercises

- KDD2010: This dataset was released in KDD Cup 2010 Educational Data Mining Challenge, which contains the situations of students submitting exercises on the systems.

- EndoMondo: This is a collection of workout logs from users of EndoMondo. Data includes multiple sources of sequential sensor data such as heart rate logs, speed, GPS, as well as sport type, gender and weather conditions.

Websites

- Phishing Websites: This dataset contains 30 kinds of features of 11,055 websites and labels of whether they are phishing websites or not. The websites' features includes 12 address-bar based features, 6 abnormal based features, 5 HTML-and-JavaScript based features and 7 domain based features.

- Behance: This is a small, anonymized, version of a larger proprietary dataset about likes and image data from the community art website Behance.

Adult

- Adult: This dataset is extracted by Barry Becker from the 1994 Census database, which consists of a list of people's attributes and whether they make over 50k a year.

News

- MIND: This dataset is a large-scale dataset for news recommendation research. It was collected from anonymized behavior logs of Microsoft News website. MIND contains about 160k English news articles and more than 15 million impression logs generated by 1 million users.

Food

- DianPing: This dataset contains the user reviews as well as the detailed business meta data information crawled from a famous Chinese online review webset DianPing.com, including the 3,605,300 reviews of 510,071 users towards 209,132 businesses.

- Food: These datasets contain recipe details and reviews from Food.com (formerly GeniusKitchen). Data includes cooking recipes and review texts.

Beverages

- BeerAdvocate: This dataset includes beer reviews with multiple rated dimensions, covering sensory aspects such as taste, look, feel, and smell.
- RateBeer: This dataset contains beer reviews with multiple rated dimensions, including item attributes with sensory aspects such as taste, look, feel, and smell.

Clothes

- ModCloth: These datasets contain measurements of clothing fit from ModCloth.
- RentTheRunway: These datasets contain measurements of clothing fit from RentTheRunway.

4. Project Report

Format

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

Expected Sections

1. Exploratory Data Analysis (EDA)

- **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

- **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

- **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.

Rubrics

The rubrics will be published on ClickUp.

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.

5. 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.

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.