Welcome to CIND820

As the final capstone course of the Data Analytics, Big Data, and Predictive Analytics certificate program, CIND820 addresses different aspects of data analytics. It may tackle problems from various perspective and use traditional machine learning techniques, including text mining, classification, or clustering analysis.

The course aims to provide a practical foundation in the applied data analytics domain. An individually designed and implemented project to (re)solve a real-world problem will be delivered, to demonstrate the skills and knowledge accumulated throughout the program.

As data scientists need to possess superior communication skills, you will be documenting your project, and you will be provided with instructions for the documentation at each stage.

In general, you will be developing a machine learning based solution, and for that you need to provide the overall applied approach and interpret the achieved results based on a medium-sized structured, semi-structured, or unstructured dataset.

Your project may be used as part of a research paper in future. In that case, you will be one of the co-authors of the article.

Notably, you must have passed all the certificate program courses (CIND123, CIND830, CIND110, CIND119 and CMTH642), to enroll in this course.

Please Note

CIND820 is a 14-week course, like all other courses in the program. As it is a capstone project course, the five modules are tied with the project milestones and not indicative of the weeks in the course.

Course Outcomes and Capstone Project Milestones

This course provides the opportunity to apply various algorithms, methods, and techniques to real-world data with an individual-based advanced data analytics project. By the end of this project course, you will be able to:

- Apply practical machine learning algorithms to frame a research problem in a way that could be addressed using AI-based approaches.
- Use relevant tools and visualization software to address the problem under study.
- Replicate state-of-the-art studies to highlight similarities as well as differences.
- Build a solid foundation on “how” and “when” to apply an AI-based approach.
- Communicate the findings of the research effectively in both written and oral presentations.

Below is an overview of the milestones of this course. Details for each milestone are in subsequent modules.

Assessment	Course Weight (in %)	Week Due

Assessment	Course Weight (in %)	Week Due
Project Abstract	5	3
Literature Review, Data Description, and Approach	25	6
Initial Results and the Code	10	9
Final Results and Project Report	35	12
Final Presentation	25	13
Total	**100%**	

Technical Requirements

This checklist will ensure you have the necessary computer and software requirements to successfully complete this course:

- I have access to a reliable computer with the minimum recommended requirements (dual-core 1.6 GHz or faster processor, 200GB of disk, 8GB of RAM) to use the software in this course (e.g., Python).
- I have internet access and adequate data allowance.
- I can access Google Meet and/or Zoom and know how to use it.

Remote Platform

Each student will be granted access to a Google Cloud Platform with all the tools to complete their project. These platforms will be available until the end of the term and will be archived by then. The lead instructor will announce all the instructions at the beginning of the course.

Please Note

While efforts have been made to make this course accessible to students using assistive technology, the software used in this course is not fully accessible, and there is no comparable software application to date. Software delivered through the remote platforms will not be accessible for screen reader users.

Topics and Learning Objectives

Topics

- Identifying a theme and research questions
- Selecting a dataset
- Writing the project abstract

Learning Objectives

By the end of the module, you will be able to:

- Identify a theme to run your analyses from themes that have been taught in the other certificate program courses (Text Mining, Classification, Clustering, Sentiment Analysis, Recommender Systems, Clickstream Analysis)
- Conduct research to select a publicly available dataset for your project (either from a list of repositories suggested by the course lead or on your own)
- Formulate at least three research questions recognizing the main problems of the selected theme (Notably, the research questions should be connected, relevant, and justifying the research effort.)

Choosing a Theme

To get you started, identify the theme on which you would like to work. Here is a list of tentative themes that you can choose for your project. You can either choose one or combine two themes. If you would like to work on a different theme, please contact the lead instructor with your proposal and the tentative dataset.

- Text Mining and Sentiment Analysis
- Classification and Regression (non-textual dataset)
- Predictive Analytics (Pattern mining, Time-series, Causality, etc.)
- Recommender systems (Collaborative; Content-based filtering, etc.)
- Anomaly Detection (outliers detection)
- Data Mining and knowledge discovery
- Click Stream Analysis

You will be assigned to a supervisor based on the project theme, and you will be working with them remotely. It is also possible that your supervisor asks you to use another dataset if the required task is not feasible or difficult to achieve.

Selecting a Dataset

Here is a list of public repositories of data that you can use for your project. Please select one of these datasets or any other public dataset(s) that match at least one of the themes mentioned above:

1. The Yelp Dataset

   "The Yelp dataset is a subset of our businesses, reviews, and user data for use in personal, educational, and academic purposes. Available as JSON files."

2. The Medicare & Medicaid Services Datasets

   "This site is dedicated to making high value health data more accessible to entrepreneurs, researchers, and policy makers in the hopes of better health outcomes for all."

3. The HealthData Datasets

"This site gives you direct access to the Centers for Medicare & Medicaid Services' (CMS) official data that are used on the Medicare Care Compare website and directories."

4. The Ontario Government Datasets

   "This site shares government datasets online to increase transparency and accountability. The aim is to help encourage innovation and solve problems through new ideas and applications."

5. The Canada Government Datasets

   "This site makes government of Canada data more accessible to everyone. You can browse a collection of more than 26,000 open data and information assets"

6. UC Irvine Machine Learning Repository

   "This site maintains 588 datasets as a service to the machine learning community. "

You may use Google Dataset Search, a search engine that Google launched to help scientists find the datasets they need.

Please Note

No private or proprietary datasets are allowed after the course start date, due to the delay in the legal process of filing a non-disclosure agreement with the university. If you wish to use any organization's dataset, the dataset must be publicly available on their website.

Working with a Supervisor

After submitting the abstract that demonstrates your ideas regarding the problem domain, project theme, research questions, and selected datasets, the lead instructor will contact you with the supervisor's information and availability. It would be best to regularly schedule a meeting with your supervisor to share your ideas and verify your project's workflow. Supervisors will not be coding for you; they provide technical and theoretical guidance to thrive in your project. You shall be building the product yourself. Notably, you need to create a GitHub repository, share the link with the designated supervisor, and upload all the interim and technical reports as you advance in the course.

Assessments

Project Abstract (5% of your final grade)

An abstract is a summary of your project. The abstract should contain:

- A brief context about the problem and the theme(s) you have chosen
- The problem that you are solving (e.g., the research questions or the summary of research questions)
- The data you are using
- The techniques and the tools that you are proposing to solve the stated problem

The abstract should:

- Include a cover page that includes your name, student number, supervisor's name, and date of submission
- Be approximately 500 words (excluding the cover page and references), double-spaced, Times New Roman font 12, using APA writing conventions where appropriate

This assignment is due by 11:59 p.m. EST on Monday of Week 3. Submit your assignment to the D2L "Project Abstract" assignment link. Late submissions will be penalized with 10% loss per day, up to a maximum of five business days. After five business days, it is up to the designated supervisor to accept or reject your submission. In the case of acceptance, the submission should be marked out of 50%.

References

- N/A