Mechanical & Industrial Engineering
UNIVERSITY OF TORONTO

MIE 1624
Group Project Report

Group 9

Haozhe Yu
Zijie Zhao
Yichen Zhang
Yuan Shang Ding
Chin Ming Wong
Evan Yunze Lyu

April 3rd, 2022

Table of Contents

Introduction

The advancement of technology gives individuals open access to a wide range of information and brings the world together into a closer data-driven society. Currently, the utilization, quantification and analysis of data give companies a competitive advantage in the fast past global market, and an individual's ability to extract useful information from real-world datasets is becoming a vital skill in all professional positions.

With the demands for Data analysts, data scientists and other data professionals at a historical high, data science education is becoming the key to maintaining a sustainable future for this field. For this project, the team plans to propose multiple topics related to data science education, and career selection, and to provide tools for the student to grow their knowledge effectively and create viable successful career plans. In detail, the team aims to reconstruct the 'MIE1624 – Introduction to Data Science and Analytics' course to align with the latest skillset and technical requirements from popular job search sites and optimized the course's structure for a better education experience.

Going above and beyond, the team plans to launch a new "Master of Science and Artificial Intelligence" program at the University of Toronto. To enhance the students learning experience, the structure and contents of the new master's program are intelligently driven by analyzing industry data collected from multiple credited sources. This includes providing the students with core technical skills as well as relevant business and soft skills. Furthermore, the team also plans to build data science education support tools through the team's EdTech start-up company. The goal of the tools is to help the students with course evaluation and forecast the student's career potential.

Part 1. Redesign of MIE1624 Course

In this part, the team will redesign the course curriculum for "MIE1624: Introduction to Data Science and Analytics". In order to design an introductory course that is capable of helping students to find data-based jobs, the course design should be based on the data collected from the real-life data-related industry. Therefore, the team decided to apply web scraping to acquire data containing information about the most demanding data science topics, tools, as well as soft skills from job posting websites.

1.1 Data Collection

The team decided to acquire data from the Indeed job posting website. The team managed to design a web crawler that takes queries as input and returns the job title, the company, the location of the job, the job description, and the link for the job posting. Three different queries are made, which include data scientist, data analyst, and data engineer. In order to get an overview of data-related jobs, the data of three kinds of data-related jobs are concatenated and then analyzed.

1.2 Data Processing

The processing is basically conducted using the job description obtained. First, the team selected major data-related topics such as machine learning, deep learning, etc. Then, for each topic, the team designed a script that can detect the occurrence of the specific topic within the job description. Binary encoding is applied that if certain topics appear within the job description, it returns 1, otherwise 0. At last, the data is transformed into a data frame that contains not only the original job data, but also the topics information. In terms of the data science tools and soft skills, the data is processed in a similar way.

1.3 Results

The team managed to obtain the following results showing the top 15 most demanding data science topics, tools, and soft skills (Figure 1, Figure 2, Figure 3).

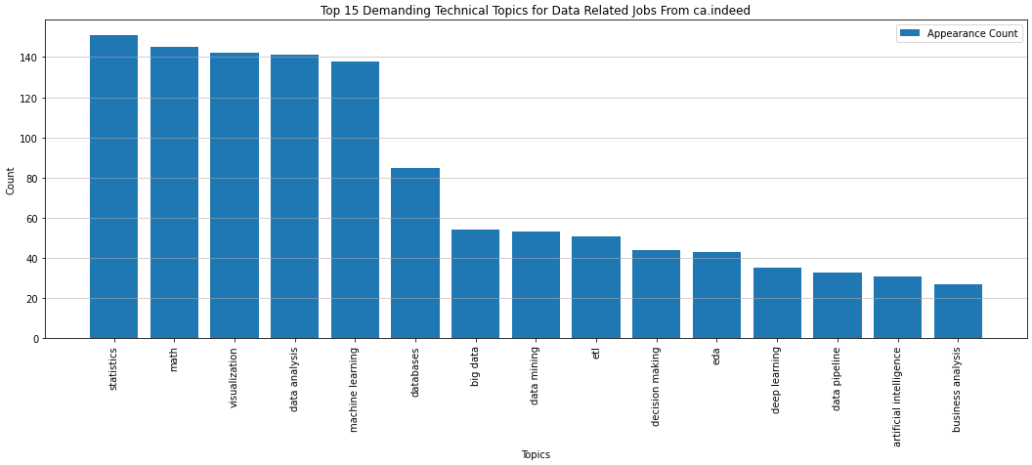

Figure 1. Top 15 Demanding Technical Topics for Data Related Jobs From ca.indeed

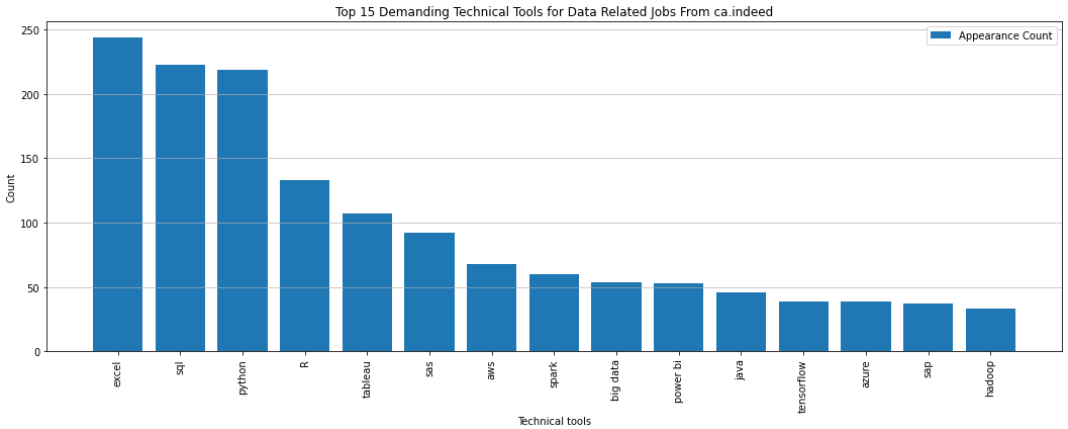

Figure 2. Top 15 Demanding Technical Tools for Data Related Jobs From ca.indeed

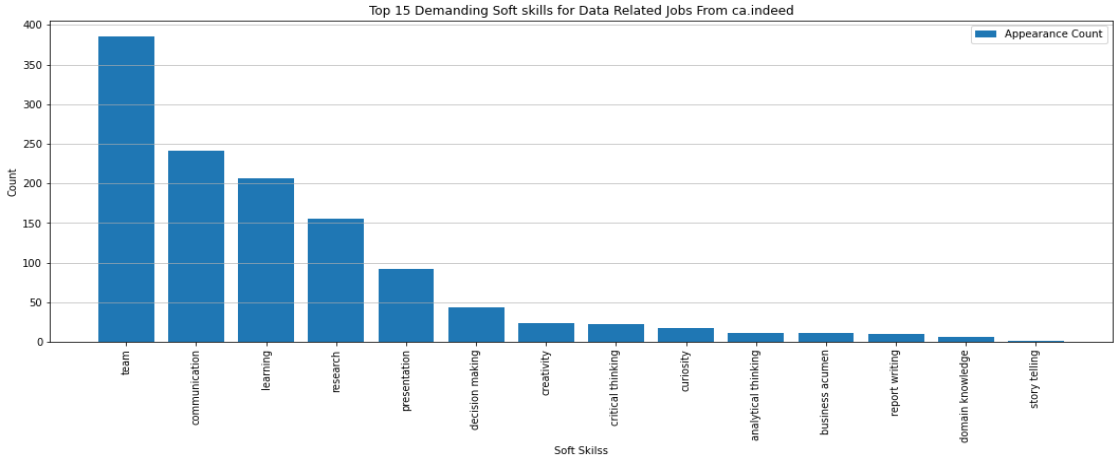

Figure 3. Top 15 Demanding Soft Skills for Data Related Jobs From ca.indeed

1.4 Findings

Based on the web scraping results from Indeed and Kaggle datasets, the team gains some insights in terms of top skills and popular fields required by data-related jobs. The top 3 most popular technical tools are Excel, Python, and SQL. The top 3 most popular fields required by data-related jobs are statistics, math, visualization, data analysis and machine learning. We need to take these into account when we re-design the MIE 1624 course.

To begin with, the course would introduce the definition of data science and its technological potential in today's world. In order to have a good understanding of the data, some prerequisites such as basic statistics and linear algebra should be introduced first. Since the majority of real-world problems in the data science field involve optimization, it is necessary to learn some optimization methods before being exposed to various models. When we deal with data science problems, it is always beneficial to understand the dataset, thus exploratory data analysis should be considered. Then it comes to the fundamental part of data science, which is machine learning. Starting from supervised machine learning algorithms, the course would introduce regressions, KNN, decision trees and ensembles, SVMs and Kernels, etc. For unsupervised machine learning algorithms, the course would introduce clustering, dimension reduction, etc. The neural network is another crucial part of machine learning, the course would cover basic principles of deep

learning. At last, since natural language processing (NLP) is getting more and more popular these days, the course would introduce basic ideas behind NLP, and one application called sentiment analysis would be discussed. The overall course schedule is illustrated in Table 1. Weekly course curriculum arrangement can be found in Figure 8 and Figure 9 in Appendix A.

1.5 Course Redesign

Table 1. Redesign of MIE1624

Week	Topic	Content
Week1	Introduction to Data Science and Python	- History and Future of Data Science - Python Basics - Numpy and Pandas Basics
Week2	Statistics in Data Science	- Important Statistics Concepts - Sampling & Distribution - Hypothesis testing - IPython Case Studies
Week3	Linear Algebra in Data Science	- Linear Algebra and Matrix Computations - Derivatives, Convexity **Assignment 1:** **Python Basics and Common Data Science Packages Implementation**
Week4	Optimization in Data Science	- Unconstrained Non-linear Optimization Algorithms - Overview of Constrained Optimization Algorithms - IPython Case Studies (Optimization)
Week5	Exploratory Data Analysis	- Dealing with Missing Data - Data Encoding - Data visualization
Week6	Supervised Machine Learning	- Linear Regression & Logistic Regression - Regularization - KNN, Decision Tree, Naive Bayes - Ensembles - Kernels and SVMs - IPython Case Studies (Supervised ML Algorithms) **Assignment 2 (Group):** **End-to-End Classification and Regression Model Implementation & Presentation**
Week7		
Week9	Unsupervised Machine Learning	- Clustering Algorithms - Dimension Reduction - IPython Case Studies (Clustering Algorithms)
Week10	Neural Network	- Feed forward - Back propagation - ANN - IPython Case Studies
Week11	Advanced Data Science Tasks - Natural Language Processing	- NLP Introduction - Sentiment Analysis - IPython Case Studies (Sentiment Analysis) **Final Project (Group):** **End-to-End Sentiment Analysis (Hotel Reviews) & Presentation**

Part 2. Curriculum Design for Master of Data Science and Artificial Intelligence

In the era of big data, data science and artificial intelligence are important research directions in the field of computer science, with high practical application value and a huge talent gap. Thus, the team aims to design a Master of Data Science and Artificial Intelligence program focusing on the latest trend and skills in the corresponding field.

Data science in the bigger picture involves the topics of analyzing, processing, extracting, visualizing, communicating, and rationalizing all forms and types of datasets using advanced data programming tools. On the other hand, the main research areas of artificial intelligence include machine learning, computer vision, natural language processing, image processing and scientific visualization. This program aims to provide comprehensive training for participants in a combination of two educational directions and also considers optional research projects to enhance their research ability and competitiveness in the workplace.

2.1 Prerequisite:

Participants should have a background in computer science, mathematics, engineering, etc. Participants are strongly recommended to have completed courses related to the following subjects: linear algebra, mathematical analysis, programming, probability, and statistics before joining the master's program.

2.2 Program composition:

Master of Data Science and Artificial Intelligence

To intelligently build the program for providing the most essential and in-demand skills and knowledge, the team utilizes the clustering study for all the popular skillsets from section 1 and conducted external research on similar programs from other education institutions(Figure 10 and Figure 11 in Appendix B).

The program consists of 3 groups of courses: compulsory, technical elective, and business. In the core compulsory courses group, there are 4 courses in total, covering the fundamentals and knowledge of this program. The compulsory courses include Introduction to artificial intelligence and algorithm, Introduction to machine learning, Data mining technology and big data, Introduction to file and database management (Figure 12 and Figure 13 in Appendix C).

The program also offers 7 elective courses with 4 technical electives and 3 business electives. The technical elective course covers Algorithms and Data Structures Specialization, Reinforcement Learning Specialization, Application of NLP, and Introduction to Computer Vision. On the business side, the program provides courses in Communication and Strategic Leadership, Data-based Management Decisions and Management Analytics Practicum. Additionally, students have the opportunity to work with faculty professors on cutting-edge data science and artificial intelligence projects in the field of biomedical, autonomous driving, speech

recognition etc. The program also offers an optional 4-month or 12-month co-op internship program for students' exposure to the industry.

For students to complete the program the is necessary to complete all 4 compulsory courses plus the student's selection of 4 elective courses, creating a total of 8 courses minimum graduation requirement.

2.2.1 Compulsory course:

Courses	Courses Description
Introduction to Artificial Intelligence and Algorithm [1]	Explores the foundation of computational intelligence. The overall goal is to understand and development of several theories, mathematical formalisms, and algorithms, that create some level of artificial intelligence. The course provides learning on latest ideas and tools that can be applied in the application of AI
Introduction to Machine Learning [1]	This is the first level course on the topic of ML, the course focuses on the introductory of various types of ML models such as the classification of image, text and data using ML. The course includes several project assignments, aiming to improve participant's coding ability and understanding different concepts of ML
Data Mining Technology and Big Data [1]	This course introduces the basic knowledge and common methods of data mining and data analysis, and guides participants to use distributed framework for feature mining of big data in practical cases
Introduction to File and Database Management [1]	This is an introductory course covering the fundamental principles and practices of database management and database design. Student will learn about relational database, normalization, and transformation. The course will also introduce the application of standard navigation language for relational database: SQL.

2.2.2 Technical Elective courses:

Courses	Courses Description
Algorithms and Data Structures Specialization [2]	This course is designed to provide students with a foundation in computer science research. The content of this course includes topics such as mathematics, permutations, and combinations, and basic to advanced data structures. Time and space complexity, sorting methods, red-black trees, hash functions will be discussed alone with other more advanced algorithms.
Reinforcement Learning Specialization [1]	This course explores the power of AI and adaptive learning systems. Reinforcement learning is a branch of machine learning that can autonomously explore the environment to gain

	relevant knowledge and use that knowledge to move in the direction of the optimal goal. The course covers topics including dynamic programming, Monte Carlo, and Temporal-Difference Learning.
Application of Natural Language Processing (NLP) [3]	Natural Language Processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence that uses algorithms to interpret and manipulate human language. This course focus on NLP applications such as information retrieval and extraction, intelligent web searching, speech recognition, and machine translation.
Introduction to Computer Vision [4]	This course focuses on using deep learning to solve problems in traditional computer vision, including cutting-edge method discussion, construction and optimization of YOLO framework, transformer application, etc.

2.2.3 Business elective courses:

Courses	Courses Description
Communication and Strategic Leadership [5]	This course provides a more conscious process for decision making, whether operational, tactical, or strategic, and for effective communications at the workplace, whether technical or non-technical.
Data-based Management Decisions [6]	This course introduces the key ideas about data-intensive decision making. The course covers various techniques in applying probabilistic concepts (distributions, measures of variability and co-variability and statistical hypothesis) to facilitate robust, data-driven management decision making.
Management Analytics Practicum [6] (Project based)	The practicum course is a practical project-based course that is designed to improve students' skills in all key steps of a data-driven management analytics project: understanding the managerial background, structuring the project, working with data, producing relevant results, presenting them effectively, while managing the project effectively along the way.

2.2.4 Optional Research Project:

The curriculum also offers optional research projects where students work under the supervision of faculty professors on challenging real-world data science and artificial intelligent projects. Topics include biomedical image scanning, speech recognition, autonomous driving etc.

Part 3. Visualization

This section gives the student visual reference to the courses and the overall structure of this new Master of Science and Artificial Intelligence" program. The team aims to utilize these visualizations to provide guidance for better course selection and easier career planning.

3.1 Radar Charts for All Courses

As mentioned in section 2.0, courses are separated into the core compulsory course, more advanced technical elected course, and business course. The radar chart for each course in Figure 4, shows the intensity of the top 6 skillsets or knowledge available for students at each course. The radar charts clearly illustrate the commonality and difference in content between each course. Additionally, based on the chart, a higher rating of a particular topic within a course represents a strong education focus on that topic during the course. Students can utilize this radar chart to align with their interests for a better course selections experience.

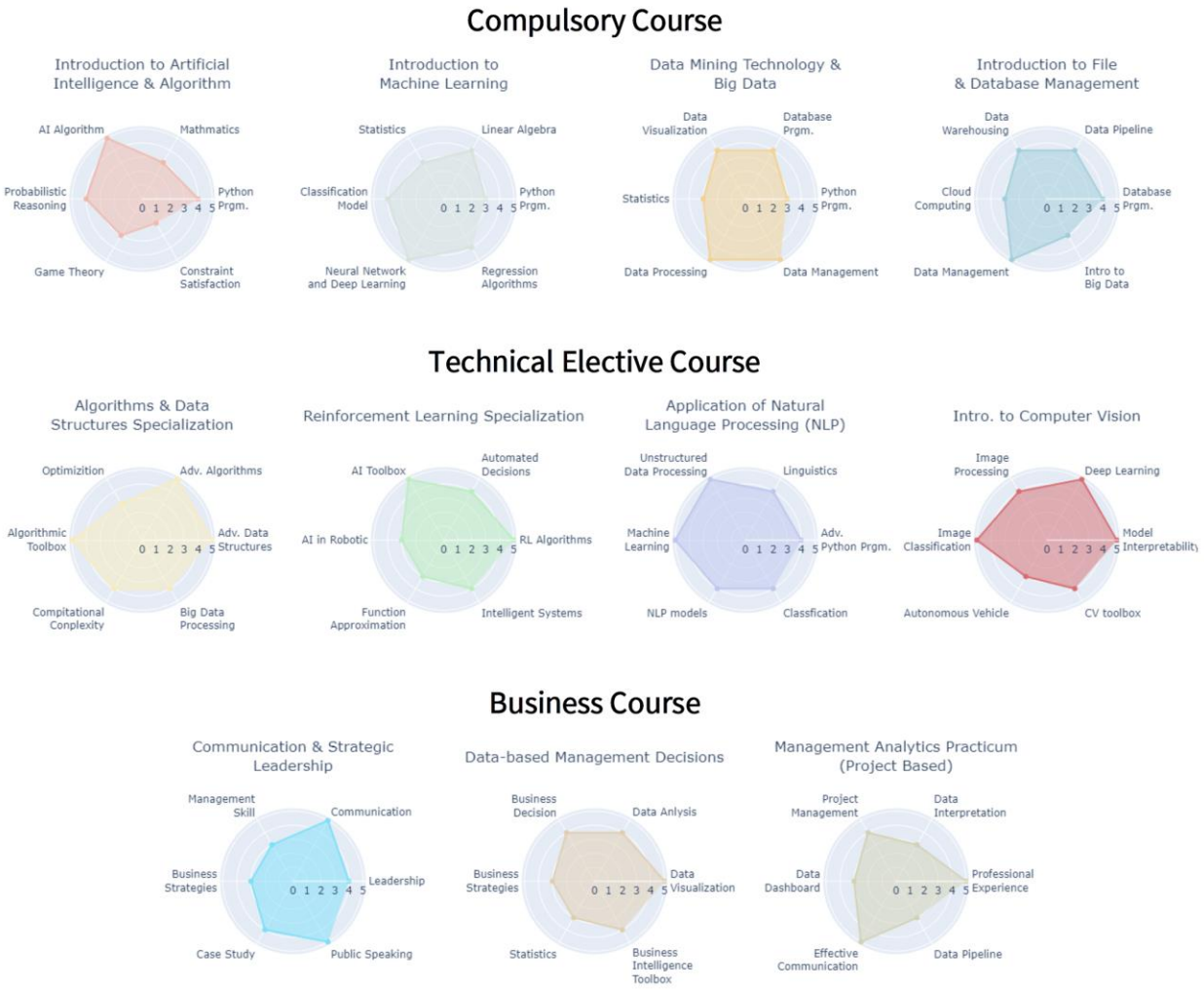

Figure 4. Radar chart for courses in Master of Science and Artificial Intelligence program

3.2 Career Planning Road Map

After completing the 4 core compulsory courses, students are faced with a selection of elective and business courses. From experience, a wrong choice in course selection may lead to uncertainty in a student's career goal. Therefore, from the knowledge obtained in section 1 and section 2, the team created the Figure 5 below indicating the preferred courses for a few popular professional positions. The technical courses are backgrounded in light orange and the business courses are backgrounded in blue. Students can utilize the course selection in this figure, to prepare the knowledge and skills necessary for landing their desired job.

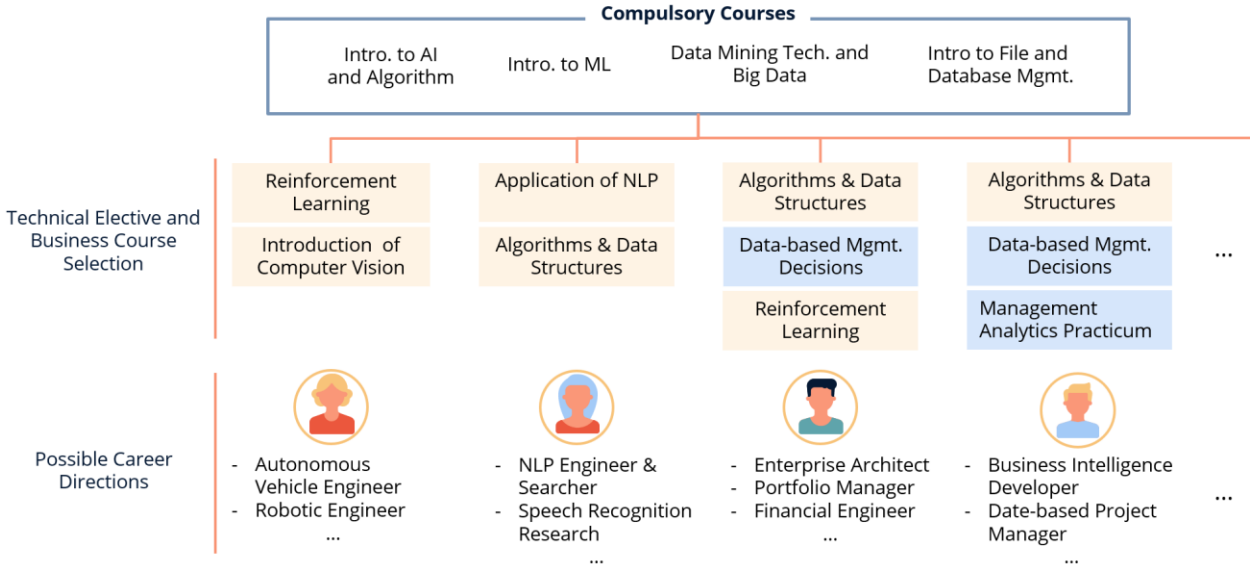

Figure 5. Career planning road map

Part 4. Data Science education EdTech effort

For the team's Edtech start-up company, the team's goal is to advance AI education and to give students advice on learning data-related skills. Our services contain two parts, the first part is the course recommender system which will recommend the top 5 courses for the specific student based on other similar students' ratings. The second part is the salary prediction model which will output a range of the annual salary based on the student's information.

4.1 Course Recommendation System

To assist students in selecting courses, the team managed to develop a demo collaborative filtering course recommendation system. Once the student rates at least one course, the recommendation system can make predictions about the student's interest in different courses by collecting information from other students with similar tastes.

To build the recommendation system, a dataset is generated to simulate the course rating data. The dataset contains 5000 students, 30 courses, and about 15000 ratings given by students. The dataset is then split into training data and test data using a 70-30 split.

The team then applied three methods to construct the course recommendation systems, which are user-user based, item-item based, and SVD decomposition-based recommendation systems. The user-user-based and the item-item-based recommendation systems calculate similarity to make recommendations while the SVD decomposition-based recommendation system uses a simplified version of the matrix factorization method. The three recommendation systems are evaluated using RMSE and MAE metrics, and the results are shown in Figure 6 below.

	Recommendation System Type	RMSE	MAE
0	User-User Based	1.439556	1.242251
1	Item-Item Based	1.818040	1.486338
2	SVD Decomposition Based	2.875992	2.493127

Figure 6. Recommendation System Performance

According to the results, the user-user-based course recommendation system has the best performance and the team decided to select it as the final course recommendation system demo. Figure 7 shows a demonstration of making course recommendations for a student whose ID is 2100. The top-5 recommendations are courses No. 23, No.26, No.2, No.28, and No.22. When implemented in real life, the actual course name along with the course ID will be displayed.

```
Top 5 Course Recommendation for Student ID2100

   course_id  Predicted Rating
0         23          3.610132
1         26          3.587247
2          2          3.574792
3         28          3.376375
4         22          3.372853
```

Figure 7. Demonstration of Making Course Recommendations for Students

There are a few points worth noticing for the practical use of the course recommendation system. In practice, the course recommendation system requires students' course rating data. The data can be collected from the University of Toronto's current course evaluation survey. In addition, although the user-user collaborative-based course recommendation system generated good performance in the team's demonstration, incorporating a more advanced matrix factorization method may further improves the model performance.

In conclusion, the course recommendation would help students select their courses based on their other students' ratings, which contributes to enhancing the learning experience for students enrolled in the program.

4.2 EdTech: Annual Salary Prediction

Implemented using machine learning tools, the team created salary prediction model based on the 2021 Kaggle Machine Learning and Data Science Survey. The goal of this model is to provide the student or any one with interest the data related field with an overview of the salary potential for this career path.

4.2.1 Data: 2021 Kaggle Machine Learning and Data Science Survey

This survey was done annually by Kaggle since 2017. The purpose was to collect a comprehensive view of data science and machine learning from different data-related jobs in different fields. This survey contains many useful data for this project such as participant's most used programming language, participant's most used integrated development environments and so on.

4.2.2 Annual salary prediction scenario

For the model to generate a prediction for an annual salary range, it requires information such as job direction, work field, abilities, and others. The prediction output of the model will be one of three salary range scenarios. Demonstration for this model is illustrated in Figure 14, Figure 15 and Figure 16 in Appendix D.

Part 5. Conclusion

In this project, the team-initiated changes to the current Introduction to Data Science and Analytics course to align it with the latest knowledge and skillset requirements from the industry. The changes are based on data collected and analyzed from popular job search sites as well as surveys conditioned in this professional field. In addition, from the clustering group of the collected dataset, the team builds a brand-new Master of Science and Artificial Intelligence program for the University of Toronto. The program offers three groups of courses with a total of 11 carefully designed courses combined. For providing guidance to course selection and career planning, the team visualized the skills in each individual course and generated a roadmap for more competitive career planning. Lastly, the team designed a course recommendation system and a salary prediction model to further advise the student on data-related skills and career potential.

Part 6. References

[1] “Emphasis in analytics,” University of Tronto Engineering Graduate Studies. [Online]. Available: https://gradstudies.engineering.utoronto.ca/professional-degrees/emphasis-in-analytics/. [Accessed: 03-Apr-2022].

[2] “Computer Engineering Course Catalogue,” University of Toronto Electrical & Computer Engineering. [Online]. Available: https://www.ece.utoronto.ca/research/computer/grad-courses/catalogue/. [Accessed: 03-Apr-2022].

[3] “Computational Linguistics Courses,” Computational linguistics. [Online]. Available: https://www.cs.toronto.edu/compling/courses.html. [Accessed: 03-Apr-2022].

[4] “Prospective: Visual computing and Computer Graphics,” Berkeley EECS. [Online]. Available: https://eecs.berkeley.edu/academics/graduate/industry-programs/meng/requirements/vccg. [Accessed: 03-Apr-2022].

[5] “Master of Science in Strategic Communication Courses,” Columbia University School of Professional Studies, 12-Apr-1970. [Online]. Available: https://sps.columbia.edu/academics/masters/strategic-communication. [Accessed: 03-Apr-2022].

[6] “Master of Management Analytics Courses,” Rotman School of Management. [Online]. Available: https://www.rotman.utoronto.ca/Degrees/MastersPrograms/MMA/TheProgram/CourseDescriptions. [Accessed: 03-Apr-2022].

Appendix

Appendix A

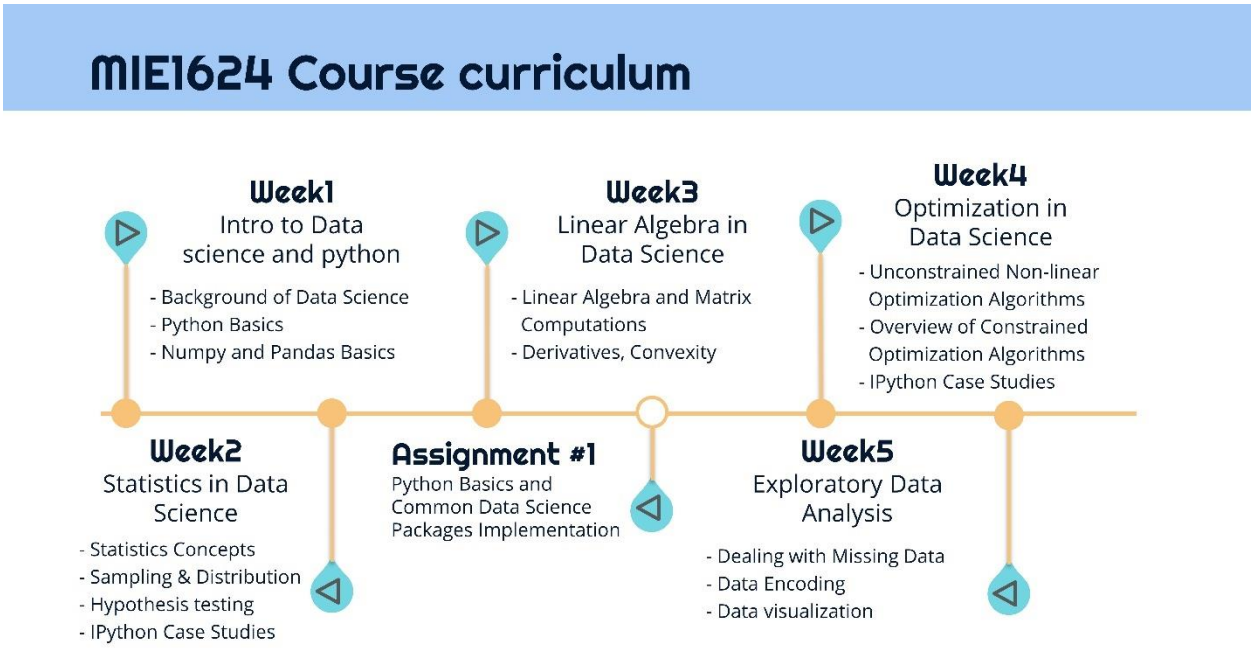

Figure 88. Resigned Course Curriculum for MIE1624

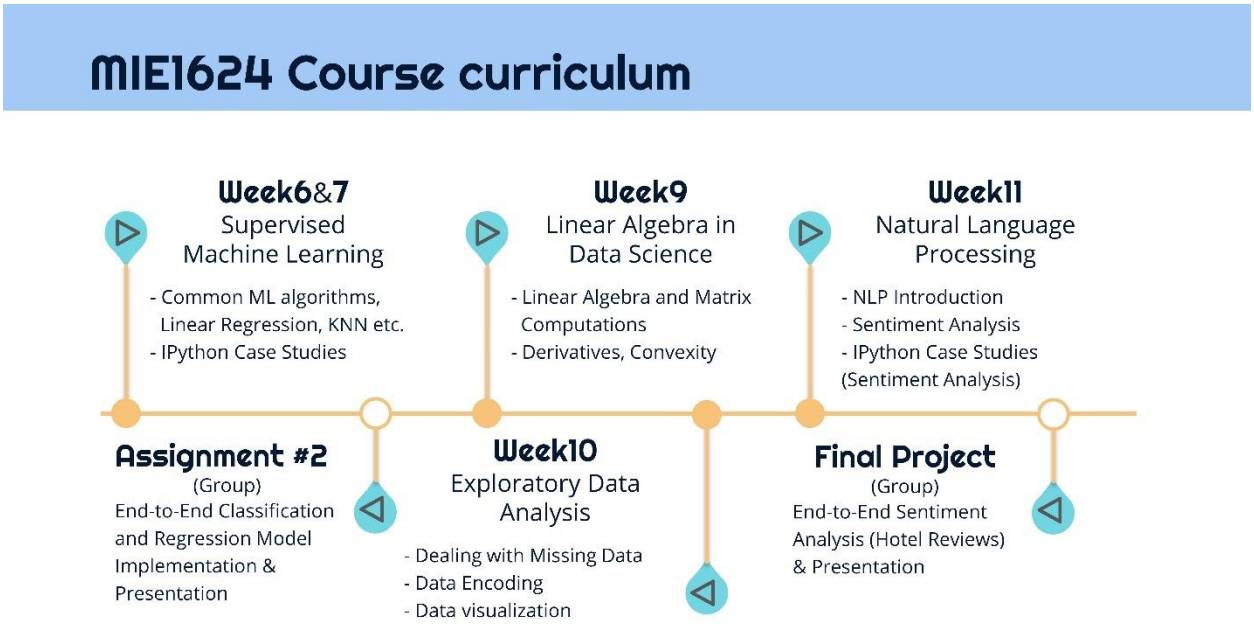

Figure 99. Resigned Course Curriculum for MIE1624

Appendix B

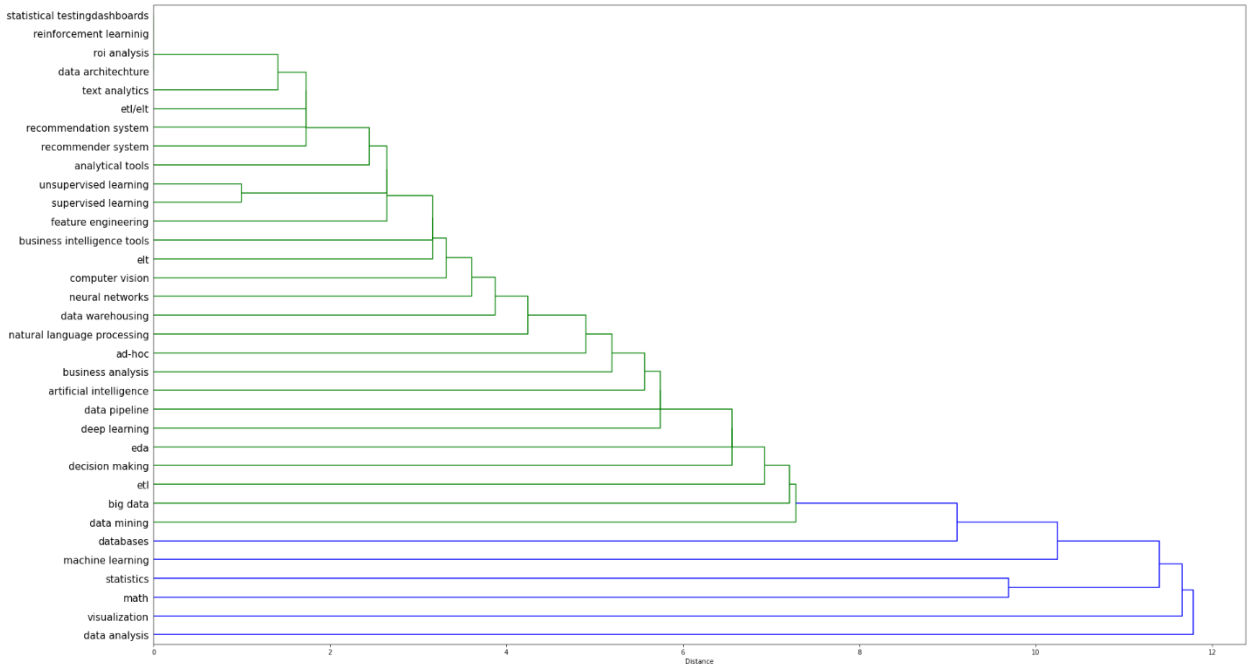

Figure 1010. Top Demanding Technical Topics Clustering for Data Related Jobs From ca.indeed

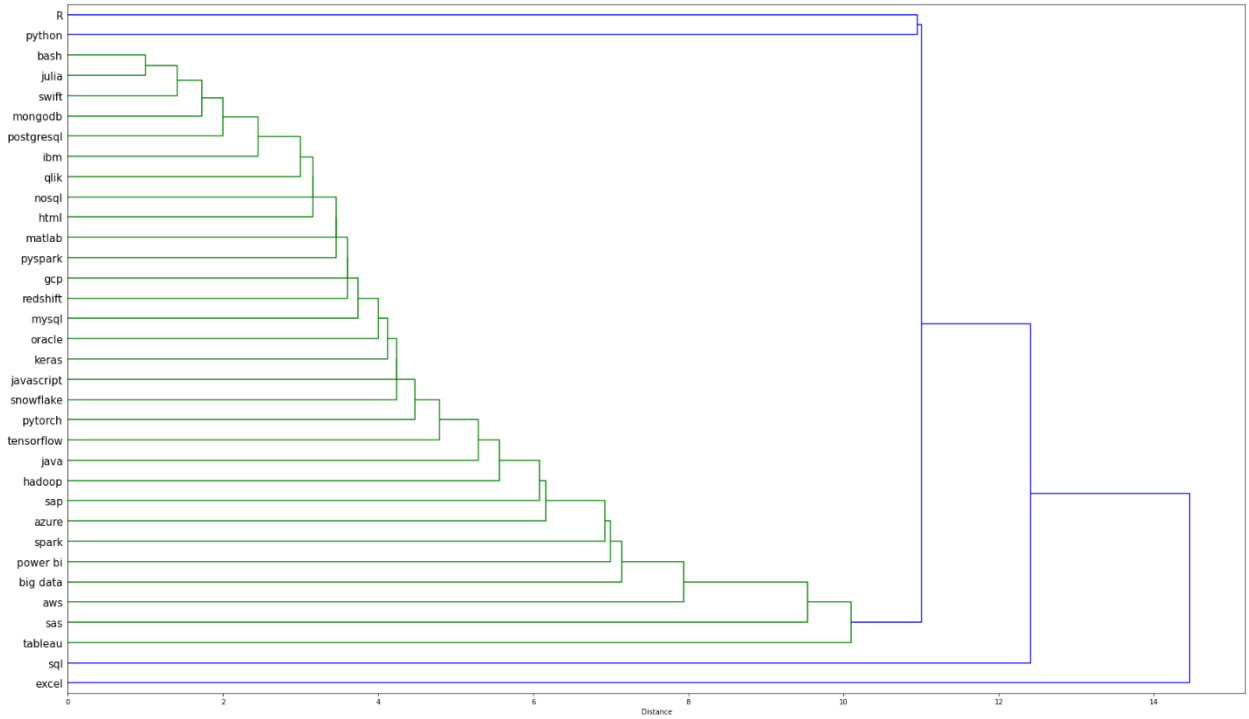

Figure 1111. Top Demanding Technical Tools Clustering for Data Related Jobs From ca.indeed

Appendix C

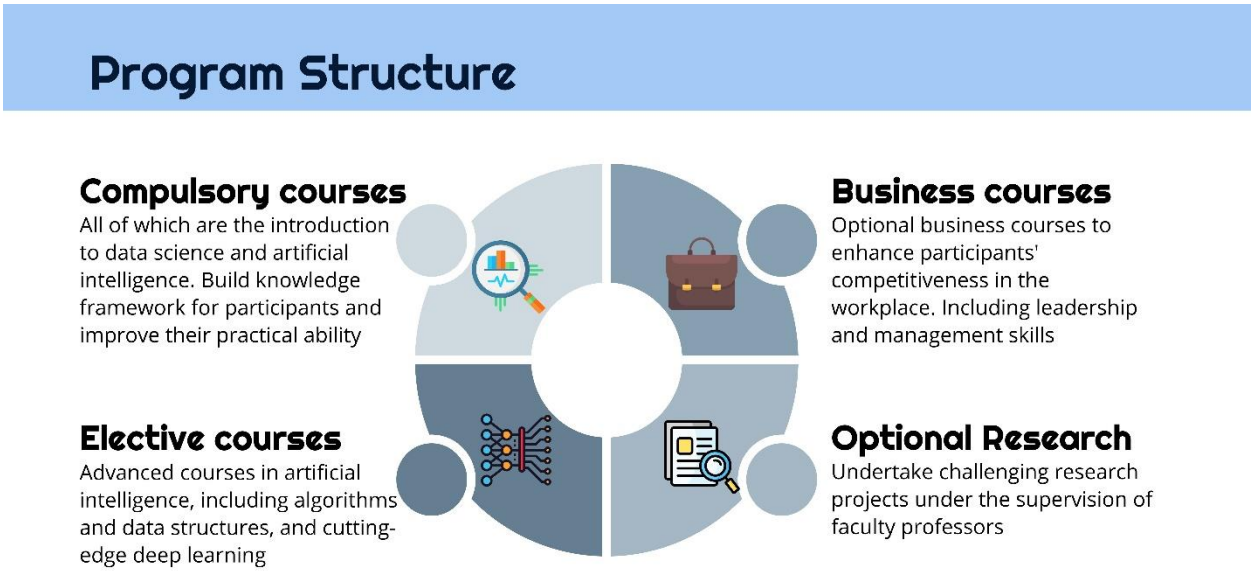

Figure 12. Program Structure for Master of Data Science and Artificial Intelligence

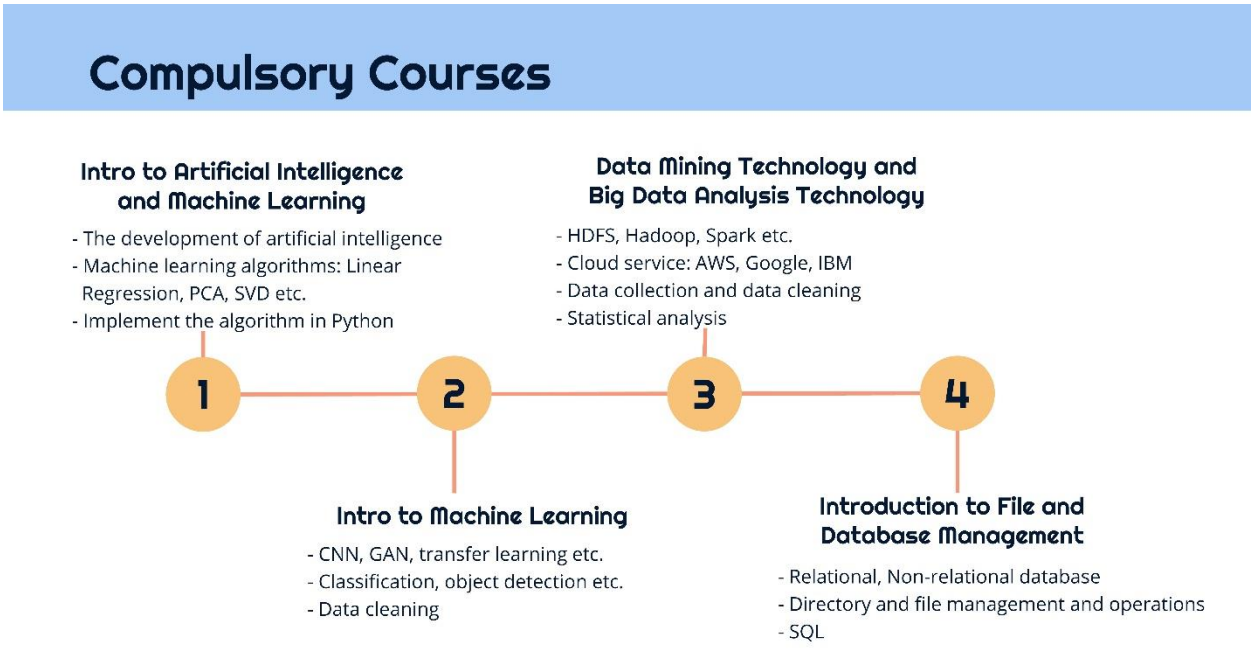

Figure 13. Timeline for Compulsory Courses

Appendix D

▼ Student1

- Age:
  - 25-29
- Country of work:
  - India
- Highest level of formal education attained or plan to attain within the next 2 years:
  - Master's degree
- Job direction:
  - Data Engineer
- Coding and/or programming experience:
  - 5-10 years
- Work field:
  - Computers/Technology
- Programming language:
  - Python
  - SQL
  - Java
- Integrated development environments:
  - PyCharm
  - Jupyter Notebook
- Hosted notebook products:
  - Databricks Collaborative Notebooks
- Data visualization libraries or tools:
  - Matplotlib
  - Bokeh
- Used machine learning methods:
  - I do not use machine learning methods

```
Salary class prediction: 2
Annual salary between $20,000 - $29,999
```

Figure 14. Salary prediction scenario 1

▾ Student2

```
Salary class prediction: 1
Annual salary between $10,000 – $19,999
```

- Age:
  - 22-24
- Country of work:
  - India
- Highest level of formal education attained or plan to attain within the next 2 years:
  - Master's degree
- Job direction:
  - Data Analyst
- Coding and/or programming experience:
  - 1-3 years
- Work field:
  - Computers/Technology
- Programming language:
  - Python
  - SQL
- Integrated development environments:
  - Visual Studio
  - Visual Studio Code (VSCode)
- Data visualization libraries or tools:
  - Matplotlib
  - Seaborn
- Platforms begun or completed data science courses:
  - Coursera
  - edX
  - Kaggle Learn Courses
  - Udacity
- Years using machine learning methods:
  - I do not use machine learning methods
- big data products:
  - Oracle Database)

Figure 15. Salary prediction scenario 2

▼ Student3

- Age:
  - 30-34
- Country of work:
  - United States of America
- Highest level of formal education attained or plan to attain within the next 2 years:
  - Master's degree
- Job direction:
  - Data Scientist
- Coding and/or programming experience:
  - 5-10 years
- Work field:
  - Computers/Technology
- Programming language:
  - Python
  - SQL
- Integrated development environments:
  - Spyder
  - Jupyter Notebook
- Hosted notebook products:
  - Google Cloud Notebooks (AI Platform / Vertex AI)
- Data visualization libraries or tools:
  - Matplotlib
  - Seaborn
- Machine learning frameworks:
  - Scikit-learn
  - Xgboost
  - Prophet
- Machine learning algorithms:
  - Linear or Logistic Regression
  - Decision Trees or Random Forests
  - Gradient Boosting Machines (xgboost, lightgbm, etc)
- Cloud computing platforms:
  - Google Cloud Platform (GCP)
- Automated machine learning tools:
  - Automated model selection (e.g. auto-sklearn, xcessiv)
  - Automation of full ML pipelines (e.g. Google AutoML, H20 Driverless AI)
- Platforms begun or completed data science courses:
  - Coursera
  - Cloud-certification programs (direct from AWS, Azure, GCP, or similar)
  - University Courses (resulting in a university degree)
- Years using machine learning methods:
  - 4-5 years

```
Salary class prediction: 10
Annual salary between $100,000 - $124,999
```

Figure 16. Salary prediction scenario