NYU SCHOOL OF PROFESSIONAL STUDIES

Data Mining and Data Warehousing

MASY1-GC 3510 | 200| Spring 2024 | 01/22/2024 -05/06/2024 | 3 Credit
Modality: Online (Sy)
Course Site URL: https://brightspace.nyu.edu/

General Course Information
Name/Title: Dr. Andres Fortino, Clinical Associate Professor, (He/Him/His)
NYU Email: agf249@nyu.edu
Class Meeting Schedule: 01/22/2024 -05/06/2024 | Mondays | 07:00pm - 09:35pm
Class Location: Distance Learning, Online Synchronous
Office Hours: By Appointment, method: NYU Zoom.

Description
In an increasingly competitive information age, data mining and data warehousing are essential in business decision-making. This course teaches students concepts, methods and skills for working with data warehouses and mining data from these warehouses to optimize competitive business strategy. In this course, students develop analytical thinking skills required to identify effective data warehousing strategies such as when to use outsource or in-source data services. Students also learn to Extract, Transform and Load data into data warehouses (the ETL process) and use the CRISP approach to data mining to extract vital information for data warehouses. The course also teaches students how to secure data and covers the ethical issues associated with the uses of data and data models for business decisions.

Prerequisites
1210 - Quantitative Models for Decision Makers

Learning Outcomes
At the conclusion of this course, students will be able to:

- Translate business requirements into well-constructed, normalized conceptual and logical data models
- Apply logical database design and the relational model to build a STAR data warehouse
- Establish a successful ETL process to load a data warehouse
- Write basic SQL statements, including some advanced SQL features
- Apply the CRISP model to conduct successful data mining
- Employ data mining tools to build data models to answer business questions

Communication Methods
Be sure to turn on your NYU Brightspace notifications and frequently check the "Announcements" section of the course site. This will be the primary method I use to communicate information critical to your success in the course. To contact me, send me an email. I will respond within 24 hours.

Structure | Method | Modality

There are 14 session topics in this course.

Active learning experiences and small group projects are key components of the course. Assignments, papers, and exams will be based on course materials (e.g., readings, and videos), lectures, and class discussions. Course sessions will be conducted synchronously on NYU Zoom, which you can access from the course site in NYU Brightspace. Sessions will be recorded for the benefit of students to review the class material.

Expectations

Learning Environment
You play an important role in creating and sustaining an intellectually rigorous and inclusive classroom culture. Respectful engagement, diverse thinking, and our lived experiences are central to this course and enrich our learning community.

Participation
You are integral to the learning experience in this class. Be prepared to actively contribute to class activities, group discussions, and work outside of class.

Course Technology Use
We will utilize multiple technologies to achieve the course goals. I expect you to use technology in ways that enhance the learning environment for all students. All class sessions require the use of Zoom.

The Use of AI
You are expected to use AI (ChatGPT and code generation tools) appropriately in this class. In fact, some assignments will require it. Learning to use AI is an important skill, and we will provide tutorials on how to use it. Be aware of the limits of ChatGPT:

- If you provide minimum effort prompts, you will get low-quality results. You will need to refine your prompts in order to get good outcomes. This will take work.
- Don't trust anything it says. If it gives you a number or fact, assume it is wrong unless you either know the answer or can check in with another source. You will be responsible for any errors or omissions provided by the tool. It works best for topics you understand.
- AI is a tool, but one that you need to acknowledge using. *Please include a paragraph at the end of any assignment that uses AI explaining what you used the AI for and what prompts you used to get the results as given below. Please do so in compliance with academic honesty policies.*
- Each assignment should have properly annotated version of the following statement at the end of the assignment, which you, in good faith, are pledging is true:

  "I/We have used AI and AI-assisted technologies as per the course policy in preparing this assignment. Specifically, I/we employed ***[Insert Name of AI Tool/Service]*** *for* ***[Insert Purpose]****. This tool/service assisted in* ***[elaborating specific contributions - e.g., drafting initial ideas, improving language]****. Following the use of the AI tool, I/we have thoroughly reviewed, edited, and critically*

analyzed the content to ensure it aligns with the learning objectives and academic integrity standards of the course. The final submission represents my/our original thought and analysis, with AI tools serving only as an aid in the process. I/We take full responsibility for the content of this assignment."

- Be thoughtful about when this tool is useful. Only use it if it is appropriate for the case or circumstance.

Assignments and Deadlines

Please submit all assignments to the appropriate section of the course site in NYU Brightspace. If you require assistance, please contact me BEFORE the due date. Students must complete and submit all assigned homework on time. Late homework or assignment submission will result in a 25% loss of credit for one-week lateness and a 50% loss for submission by the last day of class and no credit after that. Students are also expected to develop and present a team project with other students and take and pass a final exam.

Feedback and Viewing Grades

Timely, meaningful feedback will be provided to you on all your work via our course site in NYU Brightspace. You can access your grades on the course site Gradebook.

Attendance

You are expected to attend all class sessions. Attendance will be taken into consideration when determining your final grade. Refer to the SPS Policies and Procedures page for additional information about attendance.

Textbooks And Course Materials

- Data Mining and Predictive Analytics for Business Decision Making
  - **Author** – Dr. Andres Fortino
  - **Publisher** – Mercury Publishers, 2023
  - **ISBN-13:** 978-1683926757**, ISBN-10:** 1683926757

- The Data Warehouse Lifecycle Toolkit (2nd Edition) – Available through Amazon
  - **Authors** - Kimball, Ross, Thornthwaite, Mundy & Becker
  - **Publisher** – Wiley, 2008
  - **ISBN** - 978-0470149775

- Harvard Business School Publishers Cases
  - You must obtain the two cases used in class
  - Obtain a copy from https://hbsp.harvard.edu/import/1022647
- AI Access to Course materials
  - You may access all course materials via a ChatGPT chatbot
  - https://chat.openai.com/g/g-oNQXFBcZf-data-mining-course-companion

Software:

1. **Required JASP**, https://jasp-stats.org/download/

2. **Required Orange3,** https://orangedatamining.com/
3. **Required ChatGPT v4 4Plus,** https://openai.com/blog/chatgpt/

Grading | Assessment
Your grade in this course is based on your performance on multiple activities and assignments. Since all graded assignments are related directly to course objectives and learning outcomes, failure to complete any assignment will result in an unsatisfactory course grade. All written assignments are to be completed using APA format and must be typed and double-spaced. Grammar, punctuation, and spelling will be considered in grading. Please carefully proof-read your written assignments before submitting them for a grade. I will update the grades on the course site each time a grading session has been completed— typically three (3) days following the completion of an activity. Late homework or assignment submission will result in a 25% loss of credit for one-week lateness and a 50% loss for submission by the last day of class and no credit after that.

Assignments – Module Assignments – 2 Assignments (30% of final grade)
Each of the major modules of the class will be concluded with an assignment in the form of an exercise to assure the student have mastered the material presented. Instructions for the assignments are posted to the class website. Late homework or assignment submission will result in a 25% loss of credit for one-week lateness and a 50% loss for submission by the last day of class and no credit after that.

- Assignment 1 – Case Study 1 Data Warehouse (Individual) - 10%
  - Data Warehouse Design

- Assignment 2 – Case Study 2 Data Mining (Team) – 20%
  - Call Center Performance Data Mining

Labs – Seven graded Labs (40% of final grade). The top six grades will be used in the final course grade, In the data mining second half of the class there will be a lab every other week. The answers to the labs will be entered in the appropriate Quiz in the Brightspace class website. They are due one week after the class.

- Lab 1 - Designing a Data Warehouse Data Schema (8%)
- Lab 2 - Using SQL for ETL (8%)
- Lab 3 - Framing Analytical Questions (8%)
- Lab 4 - Data Preparation (8%)
- Lab 5 - Descriptive Data Mining (8%)
- Lab 6 - Predictive Data Mining (8%)
- Lab 7 - Practice Final Exam (8%)

Team Class Workshops (TCW) – (10% total, 1% each). Ten required team workshop deliverables. An additional optional workshop on visual programming is provided for practice and additional topic coverage. There is a team workshop due every week. The top 10 out of 12

TCW grades will be retained to contribute to the final grade; the lowest team workshop grade will be dropped. Student answers to the team workshops will be entered in the appropriate Assignment on the Brightspace class website. They are due by the end of the next day. The team works on the assignments at the end of each class, so there is no need for extra time to complete assignments. No credit will be given for team class workshop assignments delivered after that.

TCW 1 -	What are Data Warehousing and Data Mining?	(1%)
TCW 2 -	Planning the Data Warehouse	(1%)
TCW 3 -	Data Warehouse Design	(1%)
TCW 4 -	Components of Data Warehouse	(1%)
TCW 5 -	Loading the Data Warehouse - ETL	(1%)
TCW 6 -	Data Mining, CRISP-DM and Framing Questions	(1%)
TCW 7 -	Data Preparation	(1%)
TCW 8 -	Descriptive Data Mining	(1%)
TCW 9 -	Predictive Analytics and Linear Regression	(1%)
TCW 10 -	Analytic Models – Logistic Regression	(1%)
TCW 11 -	Analytic Models – Decision Trees	(1%)
TCW 12 -	Analytic Models – Clustering	(1%)

Review Quizzes - (10%). There are 10 out of 12 required REs (Reflection Exercises), 1% each for taking them; credit is not based on the score. This is not an exercise to measure what you know but to help you transfer knowledge from short-term memory to long-term memory. Students who use these exercises may improve their final exam grades by as much as 30%. The quizzes are open online for a whole week after the class meeting on that subject and are timed to maximize knowledge transfer. Students are advised to take each Quiz when it is available.

Final – There will be a 60-minute in-class final. (10%). A final practice exam will be made available for you to practice taking the final exam: Lab 7.

See the "Grades" section of Academic Policies" for the complete grading policy, including the letter grade conversion, and the criteria for a grade of incomplete, taking a course on a pass/fail basis, and withdrawing from a course.

NYU SCHOOL OF PROFESSIONAL STUDIES

Course Outline
Start/End Dates: **01/22/2024 -05/06/2024 / Mondays**
Time: **07:00pm -- 09:35pm**
No Class Date(s): **M - 2/19/2024 and 03/18/2024**
Special Notes: **Spring Break 03/18/24 - 03/24/24**

Session 1, 01/22/24
Introduction to Data Warehousing
We answer the question: What is a data warehouse?

- Introduction to Data Warehousing
- Relationship of Data Mining and Data Warehousing
- What is a Data Warehouse?
- Data Warehousing ROI
- DSS - Decision Support Systems
- Operational vs. Analytical Systems
- Evolution of DSS and Data Warehousing
- OLTP - Online Transaction Processing
- Characteristics of a Data Warehouse
- What is a Data Mart? Creating a Data Mart
- Data Comparison Chart
- OLAP - Online Analytical Processing

Assignments (due in one week):
Reading: Chapter 1 from Data Warehouse Lifecycle Toolkit
Team workshop 1 - due in 24 hrs.
Reflection Exercise 1 - due in 24 hrs.

Session 2, 01/29/24
Module 2: Data Warehousing 1 - Planning the Data Warehouse
We answer the question: Why do we build and use a data warehouse?

- Planning & Building the Data Warehouse
- Sponsorship and Cost Justification
- Project Prerequisites
- Barriers, Challenges and Risks
- Preparing for Implementation
- Developing the Data Warehouse
- SDLC Methodologies - Waterfall vs. RUP Approach
- Planning & Project Management
- Analysis
- Logical & Physical Design
- Implementation and Deployment
- Operations

Assignments due:
Reading: Chapter 1, 2 (The Data Warehouse Lifecycle Toolkit)

Team workshop 2 - due in 24 hrs.
Reflection Exercise 2 - due in 24 hrs.

Session 3, 02/05/24
Module 3: Data Warehousing 2- Data Warehouse Design
We answer the question: How are data warehouses designed?

- Data Warehouse Design
- Drivers for Multi-Dimensional Analysis
- Limitations of Relational Models
- The Data Cube
- What is dimensional modeling?
- Advantages of Dimensional Models
- Logical and Physical Design
- Data Normalization
- Benefits and Drawbacks of Data Normalization
- De-Normalizing of Data
- Characteristics of a Data Warehouse
- Subject Oriented, Integrated, Time Variant, Non-Volatile
- The Star Schema

Assignments due:

Assignment 1 - Data Warehouse Case Due
Reading: Chapter 6 - The Data Warehouse Lifecycle Toolkit
Team workshop 3 - due in 24 hrs.
Reflection Exercise 3 - due in 24 hrs.

Session 4, 02/12/24
Module 4: Data Warehousing 3 - Components of a Data Warehouse
We answer the question: How are data warehouses built?

- Data Warehouse Schemas
- Dimensions and Dimension Tables
- Facts and Fact Tables
- The Star Schema
- The Snowflake Schema
- The Data Warehouse Bus Architecture
- Conformed Dimensions and Standard Facts
- Data Granularity
- Changing Dimensions
- Components of a Data Warehouse
- Source Systems, Staging Area, Presentation, Access Tools
- Building the Data Matrix
- The Four Steps Process
- Multiple Fact Tables in a single Data Mart
- Chain, Heterogeneous, Transaction/Snapshot & Aggregate Facts
- Fact and Dimension Table Detail

- Identifying Source for each Fact & Dimension

Assignments due:
Reading: Chapters 4, 6 and 7 (The Data Warehouse Lifecycle Toolkit)
Team workshop 4 - due in 24 hrs.
Reflection Exercise 4 - due in 24 hrs.

Session 5, 02/26/247
Module 5: Data Warehousing 4 - Loading the Data Warehouse - ETL
We answer the question: How are data warehouses loaded with data?

- The ETL Process
- Extracting the Data into the Staging Area
- The Challenge of Extracting from Disparate Platforms
- Full vs. Incremental Extracts
- Detecting Changes to Data
- Transforming the Data
- Complexity of Data Integration
- Dealing with Missing & Dirty Data
- Data Transformation Tasks
- Loading the Data
- Timing and Job Control of Data Loads

- Assignments due:
  Reading: Chapter 9 (The Data Warehouse Lifecycle Toolkit)
  Assignments - Lab 1 - Create a logical database model using data normalization rules
  Team workshop 5 - due in 24 hrs.
  Reflection Exercise 5 - due in 24 hrs.

Session 6, 03/04/24
Module 6: Data Mining 1 - Data Mining, CRISP-DM and Framing Questions
We answer the question: What is Data Mining?

- Data Mining and Concepts
- What is Data Mining Good For?
- Statistics, Artificial Intelligence & Machine Learning
- Data Mining Examples and Tools
- Connection between Data Mining and Data Warehousing
- Retrospective Reporting vs. Predictive
- Data Mining Applications
- Data Mining vs. Statistics vs. OLAP
- Keys to commercial success of Data Mining

Assignments due:
Reading: Chapter 1,2 (Data Mining and Predictive Analytics)
Assignment – Lab 2 - Using SQL

Team workshop 6 - due in 24 hrs.
Reflection Exercise 6 - due in 24 hrs.

Session 7, 03/11/24
Module 7: Data Mining 2 - Data Preparation
We answer the question: How do we prepare data for data mining?

- Framing Analytical Questions
- Data Preparation

Assignments due:
Reading: Chapter 3,4 (Data Mining and Predictive Analytics)
Assignment: Lab 3 Framing Questions
Team workshop 7 - due in 24 hrs.
Reflection Exercise 7 - due in 24 hrs.

Session 8, 03/25/24
Module 8: Data Mining 3 – Descriptive Data Mining
We answer the question: What is Descriptive Data Mining?

- Data Mining Statistical Techniques (Sampling, Regression & Decision Trees)
- Data Mining Techniques
- Data Models
- Terminology

Assignments due:
Reading: Chapter 5,6 (Data Mining and Predictive Analytics)
Assignments: Lab 4 Data Preparation
Team workshop 8 - due in 24 hrs.
Reflection Exercise 8 - due in 24 hrs.

Session 9, 04/01/24
Module 9: Data Mining 4 – Predictive Analytics and Linear Regression
We answer the question: What is Predictive Data Mining?

- Linear Regression
- Linear Regression with multiple independent variables

Assignments due:
Reading: Chapter 7 (Data Mining and Predictive Analytics)
Assignment: Lab 5 - Descriptive Data Mining
Team workshop 9 - due in 24 hrs.
Reflection Exercise 9 - due in 24 hrs.

Session 10, 04/08/24
Module 11: Data Mining 5 – Analytic Models – Logistic Regression
We answer the question: What is Supervised Machine Learning?

- Logistic Regression

Assignments due:
Reading: Chapter 7 (Data Mining and Predictive Analytics)
Team workshop 10 - due in 24 hrs.
Reflection Exercise 10 - due in 24 hrs.

Session 11, 04/15/24
Module 11: Data Mining 6 – Analytic Models – Decision Trees
We answer the question: What is Supervised Machine Learning?

- Supervised vs. Unsupervised Data Mining
- Classification and Segmentation
- Decision Trees

Assignments due:
Reading: Chapter 8 (Data Mining and Predictive Analytics)
Team workshop 11 - due in 24 hrs.
Reflection Exercise 11 - due in 24 hrs.

Session 12, 04/22/24
Module 12: Data Mining 7 – Clustering
We answer the question: *What is unsupervised machine learning?*

- Clustering

Assignments due:
Reading: Chapter 8 (Data Mining and Predictive Analytics)
Team workshop 12 - due in 24 hrs.
Reflection Exercise 12 - due in 24 hrs.

Session 13, 04/29/24
Module 13: Data Mining 8 – Model Building and Maintenance
We answer the question: *How do we build and maintain data models?*

- Data Models
- Model Maintenance

Assignments due:
Reading: Chapter 6 (Data Mining and Predictive Analytics)
Assignment: Lab 6 – Linear Regression
Reflection Exercise 13 - due in 24 hrs.

Session 14, 05/06/24
Module 14: Exam

- Final Exam

Assignments due:
Data Warehouse Team Project: Assignment 2 - Team Call Center Case

NYU SCHOOL OF PROFESSIONAL STUDIES

Assignment: Lab 7 – Final Exam Practice

Dates	Topics	Reference	Lab Homeworks	Question Asnwered
Session 1 Jan 22	Module 1: Introduction What are Data Warehousing and Data Mining?	Kimball Ch 1		What is a data warehouse?
Session 2 Jan 29	Module 2: Data Warehousing 1 Planning the Data Warehouse	Kimball Ch 1,2		Why do we build and use a data warehouse?
Session 3 Feb 5	Module 3: Data Warehousing 2 Data Warehouse Design	Kimball Ch 6	**Assignment 1** Data Warehouse Case	How are data warehouses designed?
Session 4 Feb 12	Module 4: Data Warehousing 3 Components of Data Warehouse	Kimball Ch 4, 6, 7		How are data warehouses built and loaded with data?
20-Feb	NO CLASS - HOLIDAY			
Session 5 Feb 26	Module 5: Data Warehousing 4 Loading the Data Warehouse - ETL	Kimball Ch 9	**Lab 1:** DW Design - Schema **Quiz**	How do we extract data from warehouses for data mining?
Session 6 Mar 4	Module 6: Data Mining 1 Data Mining, CRISP-DM and Framing Questions	Fortino Ch 1, 2	**Lab 2** Using SQL for ETL **Quiz**	What is Data Mining?
Session 7 Mar 11	Module 7: Data Mining 2 Data Preparation	Fortino Ch 3, 4	**Lab 3** Framing Questions **Quiz**	How do we prepare data for data mining?
18-Mar	NO CLASS - SPRING BREAK			
Session 8 Mar 25	Module 8: Data Mining 3 Descriptive Data Mining	Fortino Ch 5, 6	**Lab 4** Data Preparation **Quiz**	What is Descriptive Data Mining?
Session 9 Apr 1	Module 9: Data Mining 4 Predictive Analytics and Linear Regression	Fortino Ch 7	**Lab 5** Descriptive Data Mining **Quiz**	What is Predictive Data Mining?
Session 10 Apr 8	Module 10: Data Mining 5 Analytic Models – Logistic Regression	Fortino Ch 7		What is Supervised Machine Learning?
Session 11 Apr 15	Module 11: Data Mining 6 Analytic Models – Decision Trees	Fortino Ch 8		What is Supervised Machine Learning?
Session 12 Apr 22	Module 12: Data Mining 7 Analytic Models – Clustering	Fortino Ch 6		What is Model Maintenance?
Session 13 April 29	Module 13: Data Mining 6 Analytic Models – Clustering	Fortino Ch 9	**Lab 6** Predictive Data Mining **Quiz**	
Session 14 May 6	Module 14: Exam **FINAL EXAM – ONLINE 7 PM - 8 PM EST**		**Assignment 2** Team Call Center Case due **Lab 7 Final Exam Pratice (optional)**	

NOTES:

The syllabus may be modified to better meet the needs of students and to achieve the learning outcomes.

The School of Professional Studies (SPS) and its faculty celebrate and are committed to inclusion, diversity, belonging, equity, and accessibility (IDBEA), and seek to embody the IDBEA values. The School of Professional Studies (SPS), its faculty, staff, and students are committed to creating a mutually respectful and safe environment (*from the SPS IDBEA Committee*).

NYU SCHOOL OF PROFESSIONAL STUDIES

New York University School of Professional Studies Policies

1. Policies - You are responsible for reading, understanding, and complying with University Policies and Guidelines, NYU SPS Policies and Procedures, and Student Affairs and Reporting.

2. Learning/Academic Accommodations - New York University is committed to providing equal educational opportunity and participation for students who disclose their dis/ability to the Moses Center for Student Accessibility. If you are interested in applying for academic accommodations, contact the Moses Center as early as possible in the semester. If you already receive accommodations through the Moses Center, request your accommodation letters through the Moses Center Portal as soon as possible (mosescsa@nyu.edu | 212-998-4980).

3. Health and Wellness - To access the University's extensive health and mental health resources, contact the NYU Wellness Exchange. You can call its private hotline (212-443-9999), available 24 hours a day, seven days a week, to reach out to a professional who can help to address day-to-day challenges as well as other health-related concerns.

4. Student Support Resources - There are a range of resources at SPS and NYU to support your learning and professional growth. For a complete list of resources and services available to SPS students, visit the NYU SPS Office of Student Affairs site.

5. Religious Observance - As a nonsectarian, inclusive institution, NYU policy permits members of any religious group to absent themselves from classes without penalty when required for compliance with their religious obligations. Refer to the University Calendar Policy on Religious Holidays for the complete policy.

6. Academic Integrity and Plagiarism - You are expected to be honest and ethical in all academic work. Moreover, you are expected to demonstrate how what you have learned incorporates an understanding of the research and expertise of scholars and other appropriate experts; and thus recognizing others' published work or teachings—whether that of authors, lecturers, or one's peers—is a required practice in all academic projects.

Plagiarism involves borrowing or using information from other sources without proper and full credit. You are subject to disciplinary actions for the following offenses which include but are not limited to cheating, plagiarism, forgery or unauthorized use of documents, and false form of identification

Turnitin, an originality detection service in NYU Brightspace, may be used in this course to check your work for plagiarism.

Read more about academic integrity policies at the NYU School of Professional Studies on the Academic Policies for NYU SPS Students page.

7. Use of Third-Party Tools - During this class, you may be required to use non-NYU apps/platforms/software as a part of course studies, and thus, will be required to agree to the “Terms of Use” (TOU) associated with such apps/platforms/software.

These services may require you to create an account but you can use a pseudonym (which may not identify you to the public community, but which may still identify you by IP address to the company and companies with whom it shares data).

You should carefully read those terms of use regarding the impact on your privacy rights and intellectual property rights. If you have any questions regarding those terms of use or the impact on the class, you are encouraged to ask the instructor prior to the add/drop deadline.