**Bachelor of Science in Data Science and Business Administration**

**BSDSBA 2027**

**School Year 2023-2024**

**Course Outline**

|  |  |
| --- | --- |
| **School** | Aboitiz School of Information, Technology, and Entrepreneurship |
| **Course Code** | DMW 2301 |
| **Course Title** | Data Mining and Wrangling |
| **Units** | 3 |
| **Term** | Year 2 – Semester 2 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Faculty Name** | Christian Alis, Ph.D. |  |  |
| **Email Address** | calis@aim.edu |  |  |
| **Consultation Hours** | T 5-6pm |  |  |

|  |  |
| --- | --- |
| **Program Staff** | Full Name |
| **Email Address** | None |
| **Extension No.** |  |

**A. Course Description**

In this course, students will learn the foundational concepts of data wrangling (collecting, extracting, transforming, and cleaning data into a form usable for further analysis) and data mining (extracting patterns from data). They will be taught and exposed to how these foundational concepts can be applied in various real-world applications. Moreover, students will acquire the basic skills to identify specific data needs of clients on top of learning the more technical/computational data mining skills.

In this course, students will also be trained on how to extract data from various sources: the web, APIs, and databases, in addition to dealing with datasets in various formats. *Note: Content of the course outline is subject to change.*

**B. Pre-requisites**

1. General Intro to Programming (PROG 1301)

2. Business Statistics (BUSI 2305)

3. Calculus (CALCUL 1301)

4. Data Structures and Algorithms (PROG 2301)

5. Finite Math (MATH 1324)

**C. Course Learning Outcomes  
At the end of the course, students will be able to:**

1. Explain the different procedures in data wrangling and mining for various data types
2. Collect and mine data from various data sources using various techniques
3. Generate hypotheses and derive insights from different forms of datasets by having operational know-how in data mining and wrangling
4. Write and present technical reports on data analysis for a specialized audience.

**D. Course Contribution to Program Learning Goals and Objectives**

|  |  |
| --- | --- |
| **Course Learning Outcomes** | **Program Learning Goals/Objectives** |
|  | 1. BSDSBA students will be proficient in communicating business value.​ |
| 3, 4 | 1.1 Visualize data with a focus on extracting actionable insights |
| 1, 3, 4 | 1.2 Convey compelling narratives and stories related to data and insights |
|  | 2. BSDSBA students will be adept in data science tools and platforms. |
| 1, 2 | 2.1 Develop an in-depth understanding of data science and business analytics: data mining, machine learning, applied statistics, and predictive modeling. |
| 1, 3 | 2.2 Apply the right data science principles in the analysis of business problems. |
| 1, 2, 3 | 2.3 Use cutting-edge tools and data mining software to analyze data and solve real-world problems. |
|  | 3. BSDSBA students will be critical and highly skilled in programming languages and big data analysis. |
| 1, 2 | 3.1 Gain hands-on experience with programming languages and big data tools through coursework, research, and engagements with actual field practitioners. |
| 3 | 3.2 Recognize and analyze ethical issues in business situations and make well-reasoned ethical management decisions |
|  | 4. BSDSBA students will be informed decision-makers and implementors who can work with teams. |
| 1, 3, 4 | 4.1 Practice real-world problem analysis to guide pitching and decision-making. |
| 1, 2, 3, 4 | 4.2 Demonstrate knowledge of algorithms to build machine intelligence utilized in business decision-making. ​ |
| 3 | 4.3 Demonstrate use of teamwork, leadership skills, and decision-making. |

**E. Learning Methodology**

* Lectures
* Exercises and Assignments
* Project and Lab Reports
* Exam
* Online sessions if conditions require
* Course deliverables will be submitted over ALICE and nbgrader (AIM supercomputer system, Jojie)
* Generative AI technology is considered a tool or reference similar to Wikipedia and Stack Overflow. You may use them in the same manner as you use these tools, and just like them, should be cited and acknowledged. Your submissions should be your intellectual output and not lifted directly from their output. Passing off their output as your own will be considered as an academic misconduct and will be reported to the Office of the Dean for appropriate action.
* Group submissions are done and submitted by group but graded individually.

**F. Grading Criteria**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | |  | **Weight** |
| Assignments |  | | (Individual) | **20%** |
| Exams |  | | (Individual) | **25%** |
| Project and Lab Reports | (Project: 20%, Lab reports: 15%) | | (Group) | **35%** |
| Exercises |  | | (Individual) | **10%** |
| Class Participation |  | | (Individual) | **10%** |
| **Total** |  |  | | **100%** |

*\*Final Grades automatically calculated in the ALICE Grade Center are not conclusive and are subject to Program Deliberations of the Faculty at the end of the Program Term*

**G. Student Responsibilities and Conduct**Students are expected to conduct themselves with the utmost professionalism in all classes. Information and policies on student responsibilities and conduct, including dysfunctional behavior (such as attendance, plagiarism, cheating, etc.) and grievance procedures are in the Student Handbook.

**H. Course Schedule**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Session Number | Session Topic (in-session) | Pre-session Activity | *Faculty/Resource Speaker* | Learning or Case Materials | Supplementary Readings | Assessment activities | Requirements or Submissions | Post-session Activity |
| Session 1  January 13, 2025  3:30PM - 5:00PM | Class management; course intro  Working with different data formats: JSON, XML, XLS |  | Christian Alis, Ph.D. |  |  | Class Participation |  |  |
| Session 2  January 16, 2025  3:30PM - 5:00PM | Working with different data formats: JSON, XML, XLS |  | Christian Alis, Ph.D. |  |  | Class Participation | Exercise 1 (Due: 23 Jan 2025) |  |
| Session 3 January 20, 2025  3:30PM - 5:00PM | Regular expressions |  | Christian Alis, Ph.D. |  | Corey Schafer, “Python Tutorial: re Module – How to Write and Match Regular Expressions (Regex)”, YouTube | Class Participation |  |  |
| Session 4  January 23, 2025  3:30PM - 5:00PM | Regular expressions |  | Christian Alis, Ph.D. |  | Corey Schafer, “Python Tutorial: re Module – How to Write and Match Regular Expressions (Regex)”, YouTube | Class Participation | Exercise 2 (Due: 30 Jan 2025) |  |
| Session 5  January 27, 2025  3:30PM - 5:00PM | Working with databases: SQL |  | Christian Alis, Ph.D. |  | Foo Barstein, “SQL CRUD with SQLite,” YouTube | Class Participation |  |  |
| Session 6  January 30, 2025  3:30PM - 5:00PM | Working with databases: SQL |  | Christian Alis, Ph.D. |  | Foo Barstein, “SQL CRUD with SQLite,” YouTube | Class Participation | Exercise 3 (Due: 6 Feb 2025)  Assignment 1 (Due: 14 Feb 2025) |  |
| Session 7 February 3, 2025  3:30PM - 5:00PM | Lab 1 |  | Christian Alis, Ph.D. |  |  |  | Lab report (Due: 25 Feb 2025) |  |
| Session 8  February 6, 2025  3:30PM - 5:00PM | Web scraping: HTML, Beautiful Soup, requests |  | Christian Alis, Ph.D. |  | Skills Factory, “HTML - Tutorial for Beginners in 12 MINUTES! - [ 2022 UPDATED ],” YouTube   freeCodeCamp.org, “Web Scraping with Python - Beautiful Soup Crash Course”, YouTube | Class Participation |  |  |
| Session 9  February 10, 2025  3:30PM - 5:00PM | Web scraping: HTML, Beautiful Soup, requests |  | Christian Alis, Ph.D. |  | Skills Factory, “HTML - Tutorial for Beginners in 12 MINUTES! - [ 2022 UPDATED ],” YouTube   freeCodeCamp.org, “Web Scraping with Python - Beautiful Soup Crash Course”, YouTube | Class Participation | Exercise 4 (Due: 17 Feb 2025) |  |
| Session 10  February 13, 2025  3:30PM - 5:00PM | Working with APIs: Wikipedia, REST, Mastodon/Bluesky, Google |  | Christian Alis, Ph.D. |  | CertBros, "APIs Explained | Real World Examples", YouTube  WebConcepts, “REST API concepts and examples,” YouTube | Class Participation |  |  |
| Session 11  February 17, 2025  3:30PM - 5:00PM | Working with APIs: Wikipedia, REST, Mastodon/Bluesky, Google |  | Christian Alis, Ph.D. |  | CertBros, "APIs Explained | Real World Examples", YouTube  WebConcepts, “REST API concepts and examples,” YouTube | Class Participation | Exercise 5 (Due: 24 Feb 2025)  Assignment 2 (Due: 3 Mar 2025) |  |
| Session 12 February 20, 2025  3:30PM - 5:00PM | Midterm Exam |  | Christian Alis, Ph.D. |  |  |  |  |  |
| Session 13  February 24, 2025  3:30PM - 5:00PM | Midterm Exam |  | Christian Alis, Ph.D. |  |  |  |  |  |
| Session 14  February 27, 2025  3:30PM - 5:00PM | Information retrieval and searching by similarity: vector space representation, similarity measures, quality measures |  | Christian Alis, Ph.D. | Lecture notebook | Visually Explained, “Principal Component Analysis (PCA),” YouTube  Rachel Thomas, “Topic Modeling with SVD NMF (NLP video 2),” YouTube  Victor Powell, "Principal Component Analysis Explained Visually", https://setosa.io/ev/principal-component-analysis/  Aggarwal, Data Mining: The  Textbook, Chapter 3  Han et al., Data Mining:  Concepts and Techniques, 4e,  Chapter 2 | Class Participation |  |  |
| Session 15  March 3, 2025  3:30PM - 5:00PM | Information retrieval and searching by similarity: vector space representation, similarity measures, quality measures |  | Christian Alis, Ph.D. | Lecture notebook | Visually Explained, “Principal Component Analysis (PCA),” YouTube  Rachel Thomas, “Topic Modeling with SVD NMF (NLP video 2),” YouTube  Victor Powell, "Principal Component Analysis Explained Visually", https://setosa.io/ev/principal-component-analysis/  Aggarwal, Data Mining: The  Textbook, Chapter 3  Han et al., Data Mining:  Concepts and Techniques, 4e,  Chapter 2 | Class Participation | Exercise 6 (Due: 10 Mar 2025) |  |
| Session 16  March 6, 2025  3:30PM - 5:00PM | Linear dimensionality reduction and factor analysis: PCA, SVD, LSA, NMF |  | Christian Alis, Ph.D. | Lecture notebook | Visually Explained, “Principal Component Analysis (PCA),” YouTube  Rachel Thomas, “Topic Modeling with SVD NMF (NLP video 2),” YouTube  Victor Powell, "Principal Component Analysis Explained Visually", https://setosa.io/ev/principal-component-analysis/  Aggarwal, Data Mining: The  Textbook, Chapter 2  Han et al., Data Mining:  Concepts and Techniques, 4e,  Chapter 2 | Class Participation |  |  |
| Session 17  March 10, 2025  3:30PM - 5:00PM | Linear dimensionality reduction and factor analysis: PCA, SVD, LSA, NMF |  | Christian Alis, Ph.D. | Lecture notebook | Visually Explained, “Principal Component Analysis (PCA),” YouTube  Rachel Thomas, “Topic Modeling with SVD NMF (NLP video 2),” YouTube  Victor Powell, "Principal Component Analysis Explained Visually", https://setosa.io/ev/principal-component-analysis/  Aggarwal, Data Mining: The  Textbook, Chapter 2  Han et al., Data Mining:  Concepts and Techniques, 4e,  Chapter 2 | Class Participation |  |  |
| Session 18  March 13, 2025  3:30PM - 5:00PM | Linear dimensionality reduction and factor analysis: PCA, SVD, LSA, NMF |  | Christian Alis, Ph.D. | Lecture notebook | Visually Explained, “Principal Component Analysis (PCA),” YouTube  Rachel Thomas, “Topic Modeling with SVD NMF (NLP video 2),” YouTube  Victor Powell, "Principal Component Analysis Explained Visually", https://setosa.io/ev/principal-component-analysis/  Aggarwal, Data Mining: The  Textbook, Chapter 2  Han et al., Data Mining:  Concepts and Techniques, 4e,  Chapter 2 | Class Participation |  |  |
| Session 19  March 17, 2025  3:30PM - 5:00PM | Linear dimensionality reduction and factor analysis: PCA, SVD, LSA, NMF |  | Christian Alis, Ph.D. | Lecture notebook | Visually Explained, “Principal Component Analysis (PCA),” YouTube  Rachel Thomas, “Topic Modeling with SVD NMF (NLP video 2),” YouTube  Victor Powell, "Principal Component Analysis Explained Visually", https://setosa.io/ev/principal-component-analysis/  Aggarwal, Data Mining: The  Textbook, Chapter 2  Han et al., Data Mining:  Concepts and Techniques, 4e,  Chapter 2 | Class Participation | Exercise 7 (Due: 24 Mar 2025)  Assignment 3 (Due: 31 Mar 2025) |  |
| Session 20  March 20, 2025  3:30PM - 5:00PM | Lab 2 |  | Christian Alis, Ph.D. |  |  |  |  |  |
| Session 21  March 24, 2025  3:30PM - 5:00PM | Representative-based clustering; Performance measures |  | Christian Alis, Ph.D. | Lecture notebook | Aggarwal, Data Mining: The  Textbook, Chapter 6 Han et al., Data Mining:  Concepts and Techniques, 4e,  Chapter 8 | Class Participation |  |  |
| Session 22 March 27, 2025  3:30PM - 5:00PM | Representative-based clustering; Performance measures |  | Christian Alis, Ph.D. | Lecture notebook | Aggarwal, Data Mining: The  Textbook, Chapter 6 Han et al., Data Mining:  Concepts and Techniques, 4e,  Chapter 8 | Class Participation |  |  |
| Session 23  April 3, 2025  3:30PM - 5:00PM | Representative-based clustering; Performance measures |  | Christian Alis, Ph.D. | Lecture notebook | Aggarwal, Data Mining: The  Textbook, Chapter 6 Han et al., Data Mining:  Concepts and Techniques, 4e,  Chapter 8 | Class Participation |  |  |
| Session 24  April 7, 2025  3:30PM - 5:00PM | Representative-based clustering; Performance measures |  | Christian Alis, Ph.D. | Lecture notebook | Aggarwal, Data Mining: The  Textbook, Chapter 6 Han et al., Data Mining:  Concepts and Techniques, 4e,  Chapter 8 | Class Participation | Exercise 8 (Due: 14 Apr 2025) |  |
| Session 25  April 10, 2025  3:30PM - 5:00PM | Hierarchical clustering |  | Christian Alis, Ph.D. | Lecture notebook | Aggarwal, Data Mining: The  Textbook, Chapter 6 Han et al., Data Mining:  Concepts and Techniques, 4e,  Chapter 8 | Class Participation |  |  |
| Session 26  April 14, 2025  3:30PM - 5:00PM | Hierarchical clustering |  | Christian Alis, Ph.D. | Lecture notebook | Aggarwal, Data Mining: The  Textbook, Chapter 6 Han et al., Data Mining:  Concepts and Techniques, 4e,  Chapter 8 | Class Participation | Exercise 9 (Due: 21 Apr 2025) |  |
| Session 27  April 21, 2025  3:30PM - 5:00PM | Density-based clustering |  | Christian Alis, Ph.D. | Lecture notebook | Aggarwal, Data Mining: The  Textbook, Chapter 6 Han et al., Data Mining:  Concepts and Techniques, 4e,  Chapter 8 | Class Participation |  |  |
| Session 28  April 24, 2025  3:30PM - 5:00PM | Density-based clustering |  | Christian Alis, Ph.D. | Lecture notebook | Aggarwal, Data Mining: The  Textbook, Chapter 6 Han et al., Data Mining:  Concepts and Techniques, 4e,  Chapter 8 | Class Participation | Exercise 10 (Due: 1 May 2025)  Assignment 4 (Due: 8 May 2025) |  |
| Session 29  April 28, 2025  3:30PM - 5:00PM | Final Exam |  | Christian Alis, Ph.D. |  |  |  |  |  |
| Session 30 May 5, 2025  3:30PM - 5:00PM | Final Exam |  | Christian Alis, Ph.D. |  |  |  |  |  |

**Required References**

Charu Aggarwal, “Data Mining: The Textbook”, Springer, 2015  
J. Han, J. Pei and M. Kamber, “Data Mining: Concepts and Techniques”, 4th ed., Morgan Kaufmann, 2023

**Additional References**

C. Shalizi, Lecture notes for Statistics 36-350 [http://www.stat.cmu.edu/~cshalizi/350/]