COURSE PROPOSAL

COURSE CODE / COURSE LEVEL: CS 212

COURSE NAME: Data Science (and Engineering) for Discovery and Diversity

TOTAL NO. OF CONTACT HOURS: 45

CREDITS: 3

PREREQUISITES: MA 100 or MA101 and CS 160

COURSE DESCRIPTION

This course introduces students to the main concepts of data science. It combines statistical and computational theories to create and implement Machine Learning and Deep Learning models for classification and prediction. Such models may have a significant impact on society, as they can be used to automate procedures and extract relevant information from large amounts of data. Also, the quality, objectivity, and preparation of training data are addressed to examine the cognitive bias that may affect the machine learning model, thus resulting in poor performance and inaccurate predictions. Students will learn how to detect and correct implicit/explicit bias often found in A.I. and Machine Learning algorithms, which is important to determining validity /veracity of information (such as found in social media) and in threat analysis (as in cybersecurity). The course includes a critique of the inherent biases of data science itself and their societal implications. Students will be guided through the process of formulating and carrying out data science analyses with real-world data, with a focus on open, pre-existing secondary data. Using popular languages such as Python, students will learn how to transform and manipulate structured and unstructured data and manage complex computational pipelines.

SUMMARY OF COURSE CONTENT

The course will follow a hands-on approach to the fundamental concepts and principles of machine learning, pattern recognition, and containerization. Topics include descriptive statistics (measures of central tendency and variability), elementary probability theory (an introduction to Bayes' theorem), linear algebra (matrix operations), nonparametric decision making (Euclidean distance, nearest neighbor, support vector machine, decision trees), and supervised and unsupervised learning techniques such as neural networks, kernel machines, convolutional neural networks. It will also cover containerization principles as an application for reproducibility of research work.

Students will be required to work in teams on a final project that focuses on open, pre-existing secondary data. They will be exposed to their own individually selected data-rich societal problem, sharing the selected problem with their peers as a final presentation at the end of the course. Consequently, expanding the knowledge of each student in this course.

Required course materials/study visits and expected expenditure for the students

- A laptop computer is required for this course.

LEARNING OUTCOMES

Upon successful completion of the course, students will be able to:

1. perform literature review of data science papers by giving a presentation;
2. finding decision boundaries that minimize the error for classification task;
3. list some data representation and transformation techniques to interpret results;
4. integrate perspectives from computational data by training an algorithm with different learning rules;
5. train a neural network model to solve a classification problem;
6. interpret performance metrics of classifier systems;
7. create and execute docker-files to build containers and images;
8. create a virtual environment for reproducibility and replicability.

TEXTBOOKS

1. Gose, Earl, Richard Johnsonbaugh, and Steve Jost. *Pattern Recognition and Image Analysis*. Upper Saddle River, NJ: Prentice Hall PTR, 1996.
2. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. The MIT Press, 2016. ISBN: 0262035618.
3. Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. ISBN: 1491962291.

GRADING POLICY

Assessment methods:
Your grade for the course will be based on the following percentages:

15% 4 homework assignments on concepts and reporting coding debugging and interpretation of results.
20% 5 quizzes on key concepts of the material covered during the week
20% 1 Midterm Exam: Basic concepts of data science as a field.
5% Attendance
40% **Final Project**
- 10% Selecting final project:
  - students will choose a problem from available datasets and real-world challenge.
  - provide a report addressing scientific questions related to the selected problem.
- 10% Presentation of the final project.
- 10% Functional code summited on GitHub.
- 10% Final report.

Assessment criteria:

Grade A characteristics:

Work of this quality directly addresses the question or problem raised and provides a coherent argument displaying an extensive knowledge of relevant information or content. This type of work demonstrates the ability to critically evaluate concepts and theory and has an element of novelty and originality. There is clear evidence of a significant amount of reading beyond that required for the course

Grade B characteristics:

This is highly competent level of performance and directly addresses the question or problem raised. There is a demonstration of some ability to critically evaluate theory and concepts and relate them to practice. Discussions reflect the student's own arguments and are not simply a repetition of standard lecture and reference material. The work does not suffer from any major errors or omissions and provides evidence of reading beyond the required assignments

Grade C characteristics:

This is an acceptable level of performance and provides answers that are clear but limited, reflecting the information offered in the lectures and reference readings.

Grade D characteristics:

This level of performances demonstrates that the student lacks a coherent grasp of the material. Important information is omitted and irrelevant points included. In effect, the student has barely done enough to persuade the instructor that s/he should not fail.

Grade F characteristics:

This work fails to show any knowledge or understanding of the issues raised in the question. Most of the material in the answer is irrelevant.

Grade scale

A	=	94- 100%	B	=	84-86%	C-	=	70-73%
A-	=	90-93%	B-	=	80-83%	D+	=	67-69%
B+	=	87-89%	C+	=	77-79%	D	=	60-66%
			C	=	74-76%	F	=	0-59%

ATTENDANCE REQUIREMENTS:

Class meetings will be held synchronously during the scheduled class. Attendance is essential for doing well in this class. Students who miss too many classes will do poorly or fail the course. Students are responsible for all information and materials given in class whether they were present or not. Being absent from a class does not excuse you from any assignments or exams that may occur the following class. You need to notify your instructor through email before class if you cannot come to class for any reason. Excused absences include documented illness, deaths in the immediate family and other documented emergencies, call to active military duty or jury duty, and official university activities. These absences will be accommodated in a way that does not arbitrarily penalize students who have a valid excuse.

Examination policy

Make-up tests will be allowed only in extraordinary circumstances with an excused absence. If you cannot take the test at the scheduled time, you must notify me through email before the test. You will be required to bring proper documentation such as a note from a doctor, coach, or counselor; an obituary or program from a funeral service, etc., to take the make-up test. All make-up tests must be completed within a week of the date of the test.

No test retakes are allowed in this course. Every student is allowed to take each test only once. After you submit the test, you cannot retake it.

The **Final Examination** will be given during the final week. The Final Exam is project-based, which is 30% out of the entire course's grade.

Doing **homework** is extremely important for your success in this course.

The **quizzes** will be due in your online Portal each week on Monday. Quizzes are made of questions from the homework of the previous week. For this type of assignment (Quiz), you submit the entire quiz before you can see which problems you missed and which ones you did correctly. FOUR attempts will be allowed for each quiz. You do not have to take all four attempts. Your best quiz attempt will be used for the final score of the quiz. You can solve problems on the quiz in any order you like. Each quiz is time restricted. Once the quiz is opened, you will have ONE hour to complete it. The remaining time will be displayed in the upper right corner. You must finish each quiz before the due date. You will not be able to work on the quiz after the due date unless you request an extension (See Late Work Policy below). All quizzes (except six quizzes that are due on the day of the test) must be completed by 11:59 pm on Monday. See Pacing Guide. The six quizzes that are due on the day of the test are exceptions. They must be completed by 8:00 am on the day of the test.

It is the **student's responsibility** to keep up with class assignments. Don't procrastinate! Don't wait until the last minute! Within a period of 7 days of the due date, students can request and receive ONE automatic extension of 3 days per assignment. Late assignments will be accepted with a 10% penalty on the points earned after the due date and will not be accepted more than 7 days late unless prior agreement has been made with the instructor for an alternate due date. If the due date of an assignment is changed, it will be announced via email.

ACADEMIC HONESTY

As stated in the university catalog, any student who commits an act of academic dishonesty will receive a failing grade on the work in which the dishonesty occurred. In addition, acts of academic dishonesty, irrespective of the weight of the assignment, may result in the student receiving a failing grade in the course. Instances of academic dishonesty will be reported to the Dean of Academic Affairs. A student who is reported twice for academic dishonesty is subject to summary dismissal from the University. In such a case, the Academic Council will then make a recommendation to the President, who will make the final decision

STUDENTS WITH LEARNING OR OTHER DISABILITIES
John Cabot University does not discriminate on the basis of disability or handicap. Students with approved accommodations must inform their professors at the beginning of the term. Please see the website for the complete policy.

SCHEDULE
Week 1
Module 1 Intro to Data Science and tools

- Introduction to Python and its main packages for data science.
- Git & GitHub Intro
- Quiz #1, Homework #1

Expected Outcome: This module will provide the main concept of programming and python packages used for data science.

Week 2
Module 2 Matrix Computation and Intro to Statistics and Probability

- Concepts of matrices and their fundamental operations.

- Hands-on matrix operations and implementation on python.
- Main concepts of random samples, frequency tables, histograms, central tendency, and variability.
- Some Probability Rules – Compound Events (Bayes' Theorem)
- Quiz #2, Homework #2

Expected Outcome: This module will teach how to handle and work with matrices and arrays using Python, implementing linear algebra concepts into image analysis applications.

Week 3

Module 3 Intro to Machine Learning (ML) with Python

- Unsupervised Learning Techniques and Applications
- Supervised Learning Techniques and Applications
- Quiz #3, Homework #3, Midterm Exam

Expected Outcome: This module will provide a set of techniques and skills that help develop a classification system for various data sets and evaluate the performance of the classifier system.

Week 4

Module 4 Intro to Deep Learning (DL) with Python and Docker for reproducibility and replicability

- Deep Learning Techniques and Applications
- Package up an application with docker
- Quiz #4, Homework #4

Expected Outcome: This module will provide a set of leading DL techniques to deal with a societal problem.

Week 5

Module 5 Wrapping Up and Final Research Projects

- The student will select and solve a societal problem using ML or DL techniques.
- Package up its application integrating GitHub and DockerHub for reproducibility.
- Quiz #5, Final, Oral presentation, project deadline.

Expected Outcome: This module will explain how data science is a field that can change and contribute to the development of new knowledge in AI.

TOOLS AND ADITTIONAL MATERIALS

Tools students will use are free of charge and available in the cloud:

1. Notion (Communication and reporting)
2. Overleaf (Latex for research templates)
3. Google Collab
4. Docker Hub
5. GitHub