Knight Foundation School of Computing and Information Sciences

Course Title: Advanced Data Science **Date:** 6/2/2024

Course Number: CAP 3764

Number of Credits: 3

Subject Area: Applications
Catalog Description: Advanced exploration topics such as machine learning, neural networks, reinforcement learning, time series, NLP, big data management, ethical AI, and emerging tech trends in data analysis.
Textbooks: Data Science from Scratch, 2nd Edition by Joel Grus. Released May 2019. Publisher(s): O'Reilly Media, Inc. ISBN: 9781492041139.
References (for further reading): Python for Data Analysis, 3rd Edition by Wes McKinney. Released August 2022. Publisher(s): O'Reilly Media, Inc. ISBN: 9781098104030.
Prerequisites Courses: CAP 2757 - Introduction to Data Science
Corequisite Courses: COP 3465 - Data Structures for IT

Type: Core Course for BS in Data Science; Elective for CS and IT Majors.

Prerequisites Topics:

1. Foundational data science concepts such as data science lifecycles, database management, data analysis, data visualization and concepts in ethics
2. Machine learning basics such as concepts, model evaluation, and validation
3. Strong programming skills with experience in data manipulation libraries such as pandas, and a basic understanding of machine learning libraries like scikit-learn

Course Outcomes:

1. **Analyze** the architecture and inner workings of deep neural networks and unsupervised learning techniques to cluster and reduce the dimensionality of datasets.
2. **Evaluate** reinforcement learning models in various scenarios.
3. **Differentiate** between various time series forecasting models and interpret seasonality patterns in time series data.
4. **Analyze** sentiment and topics from large textual datasets.
5. **Classify** different types of NoSQL databases and their use cases.
6. **Design** interactive visualizations using advanced libraries.
7. **Apply** geospatial visualization techniques to display location-based data.
8. **Evaluate** machine learning models for fairness and potential biases.
9. **Synthesize** the implications of ethical AI on societal structures.
10. **Analyze** complex optimization problems and select appropriate techniques.
11. **Present** findings and insights derived from large-scale projects in a coherent manner.
12. **Appraise** the potential of AR and VR in data visualization and analysis.

Association between Student Outcomes and Course Outcomes

BS in Computing: Student Outcomes Graduates of the program will have an ability to:	Course Outcomes
1) Analyze a complex computing problem and to apply principles of computing and other relevant disciplines to identify solutions.	1,2,3,4,5,10,12
2) Design, implement, and evaluate a computing-based solution to meet a given set of computing requirements in the context of the program's discipline.	6
3) Communicate effectively in a variety of professional contexts.	11
4) Recognize professional responsibilities and make informed judgments in computing practice based on legal and ethical principles.	8,9
5) Function effectively as a member or leader of a team engaged in activities appropriate to the program's discipline.	

Program Specific Student Outcomes

6) Apply theory, techniques, and tools throughout the data science lifecycle and employ the resulting knowledge to satisfy stakeholders' needs. [DS]	1,2,4,7

Assessment Plan for the Course and how Data in the Course are used to assess Student Outcomes

Student and Instructor Course Outcome Surveys are administered at the conclusion of each offering, and are evaluated as described in the School's Assessment Plan: https://abet.cis.fiu.edu/

Knight Foundation School of Computing and Information Sciences
CAP 3764 Advanced Data Science

Outline

Topic	**Number of Lecture Hours** (Total: 37.5 hours = 15 weeks * 2 lectures/week * 1.25 hrs/lecture)	Outcome
1. Advanced Machine Learning 1.1. Deep Learning and Neural Networks • Basics of Neural Networks • Convolutional Neural Networks (CNNs) • Recurrent Neural Networks (RNNs) • Transfer Learning and Pre-trained Models 1.2. Unsupervised Learning • Clustering (K-Means, DBSCAN, Hierarchical) • Dimensionality Reduction (PCA, t-SNE, UMAP) 1.3. Reinforcement Learning • Basics and Application Areas • Q-Learning and Deep Q Networks (DQN) 1.4. Advanced Model Evaluation • Learning curves • Cross-validation techniques • Hyperparameter tuning and optimization	10.5	1,2
2. Introductory concepts in Time Series Analysis 2.1. Time Series Components 2.2. ARIMA, Exponential Smoothing State Space Model (ETS), Prophet 2.3. Dealing with Seasonality 2.4. Time Series Forecasting	3	3
3. Introductory concepts in Natural Language Processing (NLP): 3.1. Text Representation: Bag of Words, TF-IDF, Word Embeddings 3.2. Sequence Models for NLP: LSTM, GRU, Transformers 3.3. Information Retrieval and Text Mining 3.4. Sentiment Analysis and Topic Modeling	4.5	4

4. Advanced Data Management and introductory concepts Big Data 4.1. Big Data Frameworks (e.g., Hadoop, Spark) 4.2. Distributed Databases and NoSQL (e.g., Cassandra, MongoDB) 4.3. Real-time Data Processing	3	5
5. Advanced Data Visualization 5.1. Interactive Data Visualization 5.2. Advanced Libraries (e.g., D3.js) 5.3. Geospatial Data Visualization	3	6,7
6. Model Interpretability and Explainability 6.1. Model Agnostic Methods (e.g., LIME, SHAP) 6.2. Model-specific Methods (e.g., feature importance)	1.5	8
7. Advanced Data Ethics and Governance 7.1. Ethical AI and Fairness Audits 7.2. Interpretability and Transparency in Machine Learning 7.3. Data Sovereignty and Decentralized Data Management	2.25	8,9
8. Advanced Optimization Techniques 8.1. Genetic Algorithms 8.2. Gradient-based optimization techniques 8.3. Bayesian Optimization	2.25	10
9. Advanced Project-based Learning 9.1. Students work on large-scale projects that simulate real-world challenges in data science. 9.2. Integration of multiple data sources and hybrid modeling techniques.	4.5	11
10. Trends and Future in Data Science 10.1. Quantum Computing in Data Science 10.2. Edge Computing and Data Science at the Edge 10.3. The Role of Augmented Reality and Virtual Reality in Data Analysis 10.4. The Intersection of Biotech and Data Science	3	12

Knight Foundation School of Computing and Information Sciences
CAP 3764 Advanced Data Science

Performance Measures for Evaluation

All assignments are assigned through the Canvas course site. Please note that the deadlines are strictly enforced. For example, if the deadline is 11:59 PM, any assignment submitted after this time is considered late. It is also each student's responsibility to submit correct files and ensure the submission is successful before the deadline (please double check your Canvas submissions). If you are unable to submit your assignment through Canvas, send a copy of your assignment to your instructor before the stated deadline. There will be two exams and each exam will be cumulative with an emphasis on the most recently covered material. Please note that every student is required to be physically present to take the exams with their own laptop. Exam details will be posted on the Canvas course site (https://canvas.fiu.edu).

Assignment	Points Each	Total Points	Percentage of Final Grade
Quizzes (11-Drop-1)	10	100	10%
Homework Assignments (2)	100	200	20%
Exam 1	200	200	20%
Exam 2	200	200	20%
Class Project	300	300	30%
		TOTAL	100%

Letter Grade Distribution Table

Letter	Range%	Letter	Range%	Letter	Range%
A	93 or above	B	82 - 85.9	C	70 - 73.9
A-	90 - 92.9	B-	78 - 81.9	D	60 - 69.9
B+	86 - 89.9	C+	74 - 77.9	F	less than 60

Description of Possible Homework Activities

Homework 1: Data Cleaning and Visualization
Description: Gain a practical understanding of data preprocessing, exploratory data analysis, and visualization techniques.

Task:

1. **Data Collection and Cleaning**
    - Obtain a dataset from UCI Machine Learning Repository or Kaggle. This dataset should have both numerical and categorical variables.
    - Perform initial data cleaning:
        - Handle missing values using suitable techniques.
        - Remove duplicate rows, if any.
        - Convert categorical variables to numerical representation.
2. **Exploratory Data Analysis (EDA)**

    - Compute summary statistics for the numerical variables (mean, median, standard deviation).
    - Create visual plots to understand data distribution (histograms, scatter plots, box plots).
3. **Data Visualization**
    - Use any advanced library of choice (e.g., Seaborn, D3.js) to create an interactive visualization.
    - Highlight any interesting patterns you find.

Submission: A Jupyter notebook detailing the process with appropriate comments and the visualizations. A brief report (1-2 pages) summarizing the findings.

Description of Possible Rubric:

Criteria	**Excellent (100)**	**Good (80)**	**Average (60)**	**Below Average (40)**	**Poor (20)**	**Weight**
Dataset Choice	Perfectly suited dataset from UCI/Kaggle.	Suitable dataset with minor issues.	Generic dataset with some relevance.	Poorly chosen dataset.	No dataset or irrelevant dataset.	5%
Handling Missing Values	Excellent handling with suitable techniques.	Good handling with minor issues.	Average handling, some missing values remain.	Poor handling, many missing values remain.	No handling of missing values.	10%
Duplicate Removal and Data Formatting	All duplicates removed, perfect formatting.	Minor duplicates remain, good formatting.	Some duplicates, average formatting.	Many duplicates, poor formatting.	No effort on duplicates or formatting.	10%
Categorical Variable Conversion	Perfect conversion to numerical representation.	Good conversion with minor issues.	Average conversion, some variables not converted.	Poor conversion, many variables remain.	No conversion effort.	5%
Summary Statistics	All statistics computed perfectly.	Minor errors in computation.	Some statistics missing or computed wrongly.	Many statistics missing or wrong.	No effort on statistics.	10%
Data Distribution Plots	Excellent plots covering all data aspects.	Good plots with minor omissions.	Average plots, some data aspects missing.	Few plots, many data aspects missing.	No plots or irrelevant plots.	15%
Choice of Library and Visualization Method	Advanced library used with perfect method.	Good library with minor issues in method.	Average library, some issues in visualization.	Poor choice of library or visualization method.	No library or irrelevant method used.	10%
Clarity and Presentation of Visualizations	Highly clear and well-presented visualizations.	Good clarity and presentation with minor issues.	Average clarity and presentation.	Poor clarity and presentation.	No visualizations or irrelevant presentation.	15%
Insights and Interpretation	Deep insights and perfect interpretation.	Good insights with minor interpretation issues.	Some insights, average interpretation.	Few insights, poor interpretation.	No insights or irrelevant interpretation.	5%
Report Clarity and Organization	Highly clear and well-organized report.	Good clarity and organization with minor issues.	Average clarity and organization.	Poorly organized and unclear report.	No report or irrelevant content.	10%

Homework 2: Basic Machine Learning Model Implementation

Description: Implement basic machine learning models to understand the process of training, validating, and evaluating models.

Task:

1. **Data Splitting**
    - Using the same dataset from Assignment 1 or another of your choice, split the data into training (70%) and testing (30%) sets.

2. **Model Implementation**
    - Implement a basic supervised learning model (either regression or classification based on the dataset).
    - Use cross-validation for hyperparameter tuning.
3. **Evaluation**
    - Evaluate the model's performance using appropriate metrics (e.g., accuracy, MSE, RMSE).
    - Compare the model's predictions with actual values using suitable visualization (e.g., confusion matrix, residual plots).

Submission: A Jupyter notebook detailing the model implementation, validation, and evaluation process. A brief report (1-2 pages) discussing the model's performance and potential improvements.

Description of Possible Rubric:

Criteria	Excellent (100)	Good (80)	Average (60)	Below Average (40)	Poor (20)	Weight
Appropriate Data Split	Perfect 70-30 split with appropriate data distribution.	Minor deviations from 70-30 split.	Approximate 70-30 split with some data issues.	Significant deviations from 70-30 split.	No split or completely inappropriate split.	10%
Choice of Model	Perfectly suited model for the dataset.	Suitable model with minor issues.	Generic model with some relevance.	Poorly chosen model.	No model or irrelevant model.	10%
Model Training and Validation	Excellent training and validation with no issues.	Good training with minor validation issues.	Average training, some validation issues.	Poor training and validation.	No training or validation effort.	20%
Hyperparameter Tuning	Excellent tuning using cross-validation.	Good tuning with minor issues.	Average tuning, some parameters not optimized.	Poor tuning, many parameters not optimized.	No tuning effort.	10%
Appropriate Evaluation Metrics	All metrics perfectly suited and computed.	Most metrics suitable with minor computation issues.	Some relevant metrics used, some computation issues.	Few relevant metrics, many computation issues.	No metrics or irrelevant metrics used.	15%
Model Performance Analysis	Deep analysis with perfect interpretation.	Good analysis with minor interpretation issues.	Some analysis, average interpretation.	Limited analysis, poor interpretation.	No analysis or irrelevant interpretation.	15%
Visualization of Results	Highly clear and relevant visualizations.	Good visualizations with minor issues.	Average visualizations, some aspects missing.	Few visualizations, many aspects missing.	No visualizations or irrelevant ones.	5%
Report Clarity and Organization	Highly clear and well-organized report.	Good clarity and organization with minor issues.	Average clarity and organization.	Poorly organized and unclear report.	No report or irrelevant content.	10%
Model Analysis and Recommendations	Deep analysis with actionable recommendations.	Good analysis with some recommendations.	Some analysis, few recommendations.	Limited analysis, vague recommendations.	No analysis or irrelevant recommendations.	5%

Class Project: Advanced Data Science Application

Description: Develop an end-to-end data science project implementing advanced techniques learned throughout the course.

Task:

1. **Problem Definition**
    - Choose a complex real-world problem that requires a combination of data preprocessing, machine learning, and advanced techniques (e.g., deep learning, NLP, time series analysis).
2. **Data Collection and Preprocessing**

- Collect data relevant to the problem. This can be from public datasets or simulated/generated datasets.
- Perform thorough preprocessing including data cleaning, normalization, and feature engineering.

3. **Model Development and Deployment**
    - Implement an advanced machine learning model or ensemble of models.
    - Optimize the model using advanced techniques (e.g., deep neural networks, ensemble learning).
    - Deploy the model using a simple web application or API.
4. **Analysis and Reporting**
    - Perform thorough analysis of the model's results.
    - Use advanced visualization techniques to represent the findings.
    - Discuss any ethical considerations, biases in the data or model, and implications of your findings.

Submission: A Jupyter notebook detailing the entire process. A web application or API (if applicable). A detailed report (5-7 pages) discussing the problem, solution approach, results, and implications. Optionally, a presentation summarizing the project.

Description of Possible Rubric:

Criteria	**Excellent (100)**	**Good (80)**	**Average (60)**	**Below Average (40)**	**Poor (20)**	**Weight**
Problem Definition	Clear, unique, highly relevant problem definition.	Minor ambiguity, relevant problem.	Generic, moderate relevance.	Vague, lacking relevance.	Undefined or off-topic.	10%
Data Collection	Comprehensive, highly relevant,, responsibly sourced data responsibly sourced data.	Mostly relevant data, responsibly sourced data	Relevant with notable gaps, responsibly sourced data	Limited relevance or gaps, responsibly sourced data	Little to no relevance, responsibly sourced data	20%
Data Preprocessing	Advanced techniques, deep understanding.	Standard methods, minor omissions.	Some preprocessing, some gaps.	Limited, inconsistencies.	Little to none.	10%
Feature Engineering	Innovative, enhancing model's power.	Good, minor improvements needed.	Basic, no advanced techniques.	Sparse, missing key features.	None or misguided attempts.	10%
Model Development	Advanced models, perfect for problem.	Relevant, minor room for improvement.	Basic, little customization.	Misaligned choice.	Inappropriate or none.	5%
Model Optimization	Cutting-edge techniques for peak performance.	Standard methods, minor omissions.	Basic, room for improvements.	Minimal techniques, underperforms.	No optimization.	10%
Deployment	Seamless, robust understanding of applications.	Good, minor bugs or limitations.	Basic, notable limitations.	Significant issues, unfriendly.	None or entirely non-functional.	5%

Analysis Depth	Deep, insightful analysis.	Good, minor gaps.	Basic, missed deeper insights.	Limited, missing major insights.	No or superficial.	10%
Advanced Visualization	Effective advanced visualizations for complex insights.	Good, minor improvements needed.	Basic, missed opportunities.	Limited or ineffective.	None or irrelevant.	10%
Ethical and Bias Considerations	Deep insights, solutions proposed.	Recognizes major biases, minor gaps.	Some recognition, lacks depth.	Limited recognition.	No mention.	10%