Knight Foundation School of Computing and Information Sciences

Course Title: Introduction to Data Mining **Date:** 6/2/2024

Course Number: CAP 4770

Number of Credits: 3

Subject Area: Artificial Intelligence
Catalog Description: Data mining applications, data preparation, data reduction and various data mining techniques such as association, clustering, classification, anomaly detection.
Textbooks: Data Mining: Practical Machine Learning Tools and Techniques (Fourth Edition, 2017) by: Ian H. Witten, Eibe Frank, and Christopher J. Pal. Publisher: Morgan Kaufmann Publishers. ISBN-10: 0128042915
References: Data Mining: Practical Machine Learning Tools and Techniques (Fourth Edition, 2017) by: Ian H. Witten, Eibe Frank, and Christopher J. Pal. Publisher: Morgan Kaufmann
Prerequisites Courses: (STA 3033 or STA 2023 or STA 2122 or STA 4322) and (COP 3530 or COP 3465)
Corequisite Courses: None

Type: Elective for CS (Applications), Elective for DS (AI-Robotics)

Prerequisites Topics:

1. Basic statistics and probability concepts.
2. Data structures.
3. Programming languages.

Course Outcomes:

1. Apply Data Preprocessing Techniques
2. Implement Data Mining Algorithms
3. Evaluate Predictive Models
4. Discover and Interpret Rules
5. Communicate Data Mining Insights Effectively
6. Practice selecting and applying data mining techniques to solve real-world problems.

Knight Foundation School of Computing and Information Sciences
CAP 4770 Introduction to Data Mining

Association between Student Outcomes and Course Outcomes

BS in Computing: Student Outcomes	Course Outcomes
1) Analyze a complex computing problem and to apply principles of computing and other relevant disciplines to identify solutions.	2, 4, 6
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.	2, 4, 6
3) Communicate effectively in a variety of professional contexts.	3, 5
4) Recognize professional responsibilities and make informed judgments in computing practice based on legal and ethical principles.	
5) Function effectively as a member or leader of a team engaged in activities appropriate to the program's discipline.	6

Program Specific Student Outcomes

6) Apply computer science theory and software development fundamentals to produce computing-based solutions. [CS]	2, 6

6) Apply theory, techniques, and tools throughout the data science lifecycle and employ the resulting knowledge to satisfy stakeholders' needs. [DS]	1, 2, 3, 4, 6

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 4770 Introduction to Data Mining

Outline

Topic	Number of Lecture Hours	Outcomes
• Introduction to Data Mining o Data mining applications o Machine Learning Methods o Careers in Data Mining o Data Mining Lifecycle o Ethics	3	1
• Probability and Statistics Review o Introduction to Random Variables o Normal Distribution o Student's t-Distribution	3	1
• Weka & Python for Data Mining o Weka for machine learning o Programming with Python and Jupyter o Data Mining Packages	3	1,2
• Data Preprocessing & Visualization o Data Formats o Handling Missing Values o Standardization and Normalization o Dimensionality Reduction o Training, Validation, and Test Sets o Outliers o Tools for Visualizing Data	6	1,2
• Supervised Learning o OneR o NaiveBayes o Decision Trees o Regression o Perceptrons o Classification Rules	12	2, 3, 4, 5
• Association Rules & Market Basket Analysis o Apriori Algorithm o Frequent Pattern Mining	3	4
• Model Evaluation	3	3
• Unsupervised Learning	3	2, 5
• Advanced Topics in Data Mining	6	6

Knight Foundation School of Computing and Information Sciences
CAP 4770 Introduction to Data Mining

Performance Measures for Evaluation

Assignment	Total Points	Percentage of Final Grade
Homework (5)	100 each	15%
Exams (2)	100 each	35%
Projects (3)	100 each	25%
Final Project	100	25%
	TOTAL	100%

Letter Grade Distribution Table

Letter	Range%		Letter	Range%		Letter	Range%
A	95 or above		B	83 - 86		C	70 - 76
A-	90 - 94		B-	80 - 82		D	60 - 69
B+	87 - 89		C+	77 - 79		F	59 or less

Knight Foundation School of Computing and Information Sciences
CAP 4770 Introduction to Data Mining

Description of Possible Projects

Project 1: Predictive Modeling with Supervised Learning

Description: Students will work on a project involving predictive modeling using supervised learning techniques. They will choose a real-world dataset, preprocess the data, select appropriate features, and apply various supervised learning algorithms such as Decision Trees, Naive Bayes, and Regression. The goal is to build a predictive model and evaluate its performance using relevant metrics.

Rubric:

Criteria	Excellent (100)	Good (80)	Average (60)	Below Average (40)	Poor (20)
Dataset Selection and Preprocessing (15%)	Chooses a complex, real-world dataset and preprocesses it effectively, addressing all relevant issues.	Selects a meaningful dataset and performs preprocessing with minor issues.	Chooses a dataset but encounters challenges in preprocessing, affecting the quality of the data.	Selects an inappropriate dataset or struggles significantly with preprocessing.	Fails to choose a suitable dataset or skips preprocessing entirely.
Algorithm Implementation (20%)	Implements the chosen algorithms correctly with well-documented code.	Implements algorithms with minor errors but maintains overall functionality.	Implements algorithms with significant errors impacting functionality.	Struggles to implement algorithms, resulting in poor functionality.	Does not implement the required algorithms.
Training and Convergence (20%)	Successfully trains the model with optimal hyperparameter tuning, leading to convergence.	Achieves successful training but with suboptimal hyperparameter choices.	Encounters difficulties in training or achieving convergence.	Struggles to train the model, resulting in poor or no convergence.	Fails to train the model.
Performance Metrics and Analysis (25%)	Achieves excellent performance metrics with insightful analysis of the model's strengths and weaknesses.	Achieves good performance metrics with adequate analysis.	Achieves acceptable metrics with limited analysis.	Attains poor metrics with minimal analysis.	Fails to achieve meaningful performance metrics.
Code Quality and Readability (20%)	Code is well-structured, well-documented, and follows best practices, making it easy to understand.	Code is mostly well-structured and documented but may lack some clarity.	Code is organized but may lack proper documentation and readability.	Code lacks structure, documentation, and readability, making it challenging to understand.	Code is disorganized and entirely lacking documentation

Knight Foundation School of Computing and Information Sciences
CAP 4770 Introduction to Data Mining

Project 2: **Market Basket Analysis with Association Rules**
Description: Students will explore market basket analysis using association rules on a transactional dataset. They will apply the Apriori algorithm and conduct frequent pattern mining to discover meaningful associations between items. The project will involve preprocessing the data, setting relevant parameters, and interpreting the discovered rules.

Rubric:

Criteria	Excellent (100)	Good (80)	Average (60)	Below Average (40)	Poor (20)
Dataset Selection and Preprocessing (15%)	Chooses a complex, real-world dataset and preprocesses it effectively, addressing all relevant issues.	Selects a meaningful dataset and performs preprocessing with minor issues.	Chooses a dataset but encounters challenges in preprocessing, affecting the quality of the data.	Selects an inappropriate dataset or struggles significantly with preprocessing.	Fails to choose a suitable dataset or skips preprocessing entirely.
Algorithm Implementation (20%)	Implements the Apriori algorithm and frequent pattern mining correctly with well-documented code.	Implements algorithms with minor errors but maintains overall functionality.	Implements algorithms with significant errors impacting functionality.	Struggles to implement algorithms, resulting in poor functionality.	Does not implement the required algorithms.
Association Rule Discovery (25%)	Successfully discovers meaningful association rules with insightful interpretation.	Discovers association rules with adequate interpretation.	Discovers rules with limited analysis or less meaningful interpretations.	Struggles to discover meaningful rules or provides minimal interpretation.	Fails to discover meaningful association rules.
Performance Metrics and Analysis (20%)	Analyzes the performance of association rules effectively, considering support, confidence, and lift.	Analyzes performance with minor oversights or less detailed examination.	Analyzes performance with limited consideration of relevant metrics.	Struggles to analyze performance effectively.	Fails to analyze the performance of association rules.
Code Quality and Readability (20%)	Code is well-structured, well-documented, and follows best practices, making it easy to understand.	Code is mostly well-structured and documented but may lack some clarity.	Code is organized but may lack proper documentation and readability.	Code lacks structure, documentation, and readability, making it challenging to understand.	Code is disorganized and entirely lacking documentation