GENERAL INFORMATION

Catalog Description: Foundation of data science, with focus on applications in software engineering. Different empirical methods such as surveys, case studies, and experiments. Threats to validity. Methods for data preparation. Statistics for data understanding and assessment. Commonly used supervised and unsupervised machine learning algorithms.

Course Description: This course addresses the emerging field of data science, with focus on applications in software engineering. It describes the scientific process and guidelines for empirical studies, particularly addressing different methods such as surveys, case studies, and experiments. The course also covers methods for data preparation, statistics that support data understanding and assessment, and commonly used machine learning algorithms. Students’ learning will be evaluated using multiple means, including critical analysis of published data analytics studies applied to automated software engineering, an exam, and application of methods discussed in class as a part of a multi-stage term project.

Course Learning Outcomes: Upon successful completion of SENG 670, students will be able to
- Demonstrate knowledge of the main empirical methods (i.e., surveys, case studies, and experiments) and threats to validity
- Critically analyze and evaluate data analytics studies
- Apply different data analytics methods
- Conduct data analytics studies and analyze, interpret, and report the results.

Class Schedule:
- Introduction to data-driven software engineering [1 week]
- Types of empirical studies: Surveys, case studies, and experiments [2 weeks]
- Types of threats to validity: Construct, internal, conclusion, and external validity [1 week]
- Data understanding & assessment: Measures of centrality, spread and association; hypotheses testing [1 week]
- Data preparation: Variable cleaning & feature creation [1 week]
- Supervised learning: Decision trees, Naïve Bayes, k-Nearest Neighbors [2 weeks]
- Supervised learning: Linear regression, Multivariate liner regression, Logistic regression [2 weeks]
- Unsupervised learning: k-means clustering, Expectation-Maximization clustering, Hierarchical clustering [2 weeks]
Homework assignments: As a part of homework assignments students will be required to critically analyze and provide recommendations for improvement of several published papers presenting data analytics studies. Homework assignments will typically consist of (1) written reports due at the latest by the beginning of the class, the day it is due, (2) in-class discussion, and (3) discussion forum postings. Late written reports will not be allowed. Students who cannot attend an in-class discussion due to an excused absence will be required to submit a recorded briefing using eCampus.

Exams: There will be a midterm exam which will cover the first 5-6 weeks of classes.

Term project: Each student will conduct a multi-stage data analytics term project which will be based on the methods discussed in class. The project will have the following four stages: (1) Choosing a topic from the provided list and compiling the related work section and bibliography, (2) Data preparation and understanding, including data analytics for assessment, (3) Using data analytics for prediction, and (4) Term project presentation. Each student will write a term paper that will summarize the results (stages 1-3) and give a presentation (stage 4).

Grading: Semester grades will be computed as follows

<table>
<thead>
<tr>
<th>Assignment/exam</th>
<th>Points %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homework assignments</td>
<td>30%</td>
</tr>
<tr>
<td>Homework 1 (10%)</td>
<td>10%</td>
</tr>
<tr>
<td>Homework 2 (10%)</td>
<td>10%</td>
</tr>
<tr>
<td>Homework 3 (10%)</td>
<td>10%</td>
</tr>
<tr>
<td>Midterm exam</td>
<td>25%</td>
</tr>
<tr>
<td>Project</td>
<td>45%</td>
</tr>
<tr>
<td>Stage 1: Select a topic, write introduction &amp; related work, include bibliography (10%)</td>
<td></td>
</tr>
<tr>
<td>Stage 2: Data preparation &amp; understanding, assessment (10%)</td>
<td></td>
</tr>
<tr>
<td>Stage 3: Data analytics for prediction (15%)</td>
<td></td>
</tr>
<tr>
<td>Stage 4: Presentation (10%)</td>
<td></td>
</tr>
<tr>
<td>Class total</td>
<td>100%</td>
</tr>
</tbody>
</table>

Grading scale: Grades will generally be A = 90 – 100%, B = 80 - 89%, C = 70 - 79%, D = 60 - 69%, and F = 0 – 59%. ‘+’ and ‘-‘ grade may be reported if the score is near boundary.

RULES OF OPERATION

Attendance: Students are expected to regularly follow the lectures. Students are responsible for all material covered in the course, keeping track of assignments’ due dates and examination dates. Students are required to attend the in-class discussions of homework assignments. Each unexcused absence from these classes will result in deducting 20% from the student’s homework score. However, an excused absence from an in-class homework discussion may be made up by submitting a recorded briefing. Students should make every effort to give prior notice of an excused absence and arrange the recording of the make-up briefing with the instructor.

Communication: All course material, important announcements, assignments, etc. will be provided using the eCampus features. Please use the eCampus Course Messages feature to communicate with the Instructor and other students in the class. The instructor will make her best effort to respond within 24 hours.

Grading policies:

- No late homework assignments will be allowed (i.e., a late assignment equals no assignment).
- Late project deliverables are strongly discouraged and will be penalized 10% for each day late.
- Make-up exam will be given only by prior arrangement and only under truly extraordinary circumstances. Consistent with WVU guidelines, students absent from regularly scheduled examinations because of authorized University activities will have the opportunity to take them at an alternate time.
- Grading appeals must be in writing on the day the assignment or exam is returned.
- Every attempt will be made to grade the assignments and provide feedback to the students within the same length of time the students have been given to complete the assignment (e.g., if the students are given one week to complete a homework assignment, the instructor will attempt to provide the grade within one week after the homework has been turned-in). The exam will typically be graded within one week.
**Academic Integrity:** The integrity of the classes offered by any academic institution solidifies the foundation of its mission and cannot be sacrificed to expediency, ignorance, or blatant fraud. Therefore, the instructor will enforce rigorous standards of academic integrity in all aspects and assignments of this course. For the detailed policy of West Virginia University regarding the definitions of acts considered to fall under academic dishonesty and possible ensuing sanctions, please visit the Office of Student Conduct Web page, which has links to the Student Conduct Code ([http://studentlife.wvu.edu/office_of_student_conduct/student_conduct_code](http://studentlife.wvu.edu/office_of_student_conduct/student_conduct_code)) and other relevant documents and forms. Should you have any questions about possibly improper research citations or references, or any other activity that may be interpreted as an attempt at academic dishonesty, please see the instructor before the assignment is due to discuss the matter.

**Inclusivity Statement:** The West Virginia University community is committed to creating and fostering a positive learning and working environment based on open communication, mutual respect, and inclusion.

If you are a person with a disability and anticipate needing any type of accommodation in order to participate in this class, please advise me and make appropriate arrangements with the Office of Accessibility Services ([URL http://accessibilityservices.wvu.edu/home](http://accessibilityservices.wvu.edu/home); Phone 304 293-6700). For more information on West Virginia University's Diversity, Equity, and Inclusion initiatives, please see [http://diversity.wvu.edu](http://diversity.wvu.edu).