TRINE UNIVERSITY

**Course Syllabus**

**Course Title:** Data Science and Big Data

**Term and Year:** Spring Term 1 2024

**Course Section and Number:** IS 5213 OL1

**Time and Place:** Online

**Number of Credit Hours:** 3

**Office Location:** Online

**Instructor:** Sasanka Panda

**Office Hours:** Sat/Sun6:00–7:30PMCST

**Email Address:** pandas@trine.edu

**Cell Phone #** 936-672-0093

**Course Description:** This course introduces the latest data analytics tools and platforms, explores the rapidly developing field of Data Science. You will learn how best to gain actionable insights from big data, as well as to develop data solutions and data transformation road maps for businesses of varying sizes and complexity levels. The goal of this course is to maximize the utilization of available data and optimize the efficiency of decision-making. Previous experience with Hadoop, Spark or distributed computing is not required.

**Learning Outcomes:** Upon completion of this course, the student should be able to:

1. Configure library packages formatted for their target environment.
2. Prepare data using modeling techniques to ensure quality results.
3. Develop predictive models using machine learning and statistical techniques.
4. Recommend business solutions to stakeholders based on big data insights.

**Prerequisite: None**

**Required Text:**

- **R for Absolute Beginners – Hands – on R Tutorial**
  **Free Online Version:**
  https://www.researchgate.net/publication/331209857_R_for_Absolute_Beginners_-_Hands-on_R_Tutorial
  **Author:** Duarte and Magno
  **Published Date:** 2018

- Additional material online or provided by instructor videos and notes

**References:** https://www.r-project.org/other-docs.html

**Course Requirements:**

**Attendance/Participation:** All students are expected to log in to their courses regularly throughout the week to receive instruction, materials, and updates from the instructor. It is your responsibility to check in and submit your assignments, complete your discussion board postings, and finish quizzes and exams by the due dates.

If you do not participate in the course, you will be counted absent. **Simply logging in is not enough; you must submit/complete an assignment, post to a discussion board, or other similar assignment tasks to avoid being counted absent. Instructors are required to submit attendance the Monday following each week of class.**

This attendance is reported to the Financial Aid Department and may result in the loss of any financial aid refund you are expecting if you have not been participating in your courses. **In addition, you will be administratively dropped from the course if you are reported absent a total of three weeks.**

**Content:**

Week 1: Install "R", Exploratory Data Analysis (EDA)
Week 2: Data Scrubbing
Week 3: Decision Trees
Week 4: Model Validation
Week 5: Random Forests and Gradient Boosting Models
Week 6: Linear and Logistic Regression
Week 7: Principal Component Analysis (PCA) and tSNE analysis
Week 8: Clustering and Segmentation

**Grading/Evaluation:**

| | | |
|---|---|---|
| Assignments | : | 500 Points |
| Quizzes | : | 350 Points |
| Discussions | : | 150 Points |
| | | ======== |
| Total | : | 1000 |

**Bonus Points May Also Available! ☺**

**Late Work Will Not Be Accepted**

**Trine Graduate Grading Scale:**

| Grade | Percentage | Quality Points | Meaning of Grade |
|---|---|---|---|
| **A** | 93-100 | 4.0 | Excellent |
| **B+** | 86-92 | 3.5 | Very Good |
| **B** | 81-85 | 3.0 | Good |
| **C+** | 75-80 | 2.5 | Above Average |
| **C** | 70-74 | 2.0 | Average (lowest passing grade) |
| **F** | 00-69 | 0.0 | Failure |
| **I** | Incomplete | Not figured into GPA | |
| **IP** | In Progress (grade deferred) | Not figured into GPA | |

| | | | |
|---|---|---|---|
| **W** | Withdrawal | Withdrawal before completion of 80% of semester | |
| **WP** | Withdrawal | Withdrawal after completion of 80% of semester issued only under special circumstances and with approval of the department chair/director | |

**Other Policies:**

**Academic Misconduct:**

The University prohibits all forms of academic misconduct. Academic misconduct refers to dishonesty in examinations (cheating), presenting the ideas or the writing of someone else as one's own (plagiarism) or knowingly furnishing false information to the University by forgery, alteration, or misuse of University documents, records, or identification. Academic dishonesty includes, but is not limited to, the following examples: permitting another student to plagiarize or cheat from one's own work, submitting an academic exercise (written work, printing, design, computer program) that has been prepared totally or in part by another, acquiring improper knowledge of the contents of an exam, using unauthorized material during an exam, submitting the same paper in two different courses without knowledge and consent of professors, or submitting a forged grade change slip or computer tampering. The faculty member has the authority to grant a failing grade in cases of academic misconduct as well as referring the case to Student Life.

**Plagiarism:**

You are expected to submit your own work and to identify any portion of work that has been borrowed from others in any form. An ignorant act of plagiarism on final versions and minor projects, such as attributing or citing inadequately, will be considered a failure to master an essential course skill and will result in an F for that assignment. A deliberate act of plagiarism, such as having someone else do your work, or submitting someone else's work as your own (e.g., from the Internet, fraternity file, etc., including homework and in-class exercises), will at least result in an F for that assignment and could result in an F for the course.

**Artificial Intelligence (AI) is prohibited:** All work submitted by students in this course must be generated by the student. Students may not have another person or entity contribute to an assignment for them, which includes using AI. Students may not incorporate any part of an AI-generated response in an assignment, use AI to formulate arguments, use AI to generate ideas for an assignment, or submit work to an AI platform for improvement. Using an AI tool to generate content may qualify as academic misconduct in this course.

**Electronic Devices:**

Use of electronic devices including smart watches and cell phones is prohibited during exams or quizzes unless directly allowed by the instructor.

**References**


Anderjef (2023). *Training, validation, and test data sets.* Wikipedia. Retrieved on June 1, 2023 from https://en.wikipedia.org/wiki/Training,_validation,_and_test_data_sets

DataDaft (2020). *Introduction to R: Data frames* [Video]. YouTube. https://www.youtube.com/watch?v=edifwfMEL_I

DataDaft (2020). *Introduction to R: Matrices* [Video]. YouTube. https://www.youtube.com/watch?v=p3EC-V9MiWU

DataDaft (2020). *Introduction to R: Vectors* [Video]. YouTube. https://www.youtube.com/watch?v=H4v4MRSc8k4

Data School (2015). *ROC curves and area under the curve explained* [Video]. YouTube. https://www.youtube.com/watch?v=OAl6eAyP-yo

dewangNautiyal (n.d.). *Underfitting and overfitting.* Geeks for Geeks. Retrieved on June 1, 2023 from https://www.geeksforgeeks.org/underfitting-and-overfitting-in-machine-learning/

Frost, J. (2023). *Principal component analysis guide & examples.* Statistics by Jim. Retrieved on June 1, 2023 https://statisticsbyjim.com/basics/principal-component-analysis/

Kamperis, S. (2021). Decision tress: Gini index vs. entropy. *Let's talk about Science.* Retrieved on June 1, 2023 from https://ekamperi.github.io/machine%20learning/2021/04/13/gini-index-vs-entropy-decision-trees.html

Kassambara (2018). *CART model: Decision tree essentials.* Statistical Tools for High-Throughput Data Analysis. Retrieved on June 1, 2023 from http://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/141-cart-model-decision-tree-essentials/

Knowledge Powerhouse (2021). *What is the different between bagging and boosting methods in ensemble learning?* [Video]. YouTube. https://www.youtube.com/watch?v=UeYG64Hm7Es

O'Loughlin, E. (2022). *How to… make a prediction using a multiple linear regression model in R* [Video]. YouTube. https://www.youtube.com/watch?v=AR6sLpcVcSU

O'Loughlin, E. (2016). *How to… perform simple linear regression by hand* [Video]. YouTube. https://www.youtube.com/watch?v=GhrxgbQnEEU

Obi, P. (2016). *Stepwise regression* [Video]. YouTube. https://www.youtube.com/watch?v=AdFT17sq53s

R Programming 101 (2022). *Visualize your data using ggplot. R programming is the best platform for creating plots and graphs* [Video]. YouTube. https://www.youtube.com/watch?v=rfR9Nrpfnyg

R Programming 101 (2022). *R programming for absolute beginners* [Video]. YouTube. https://www.youtube.com/watch?v=FY8BISK5DpM

R Programming 101 (2019). *R programming for beginners – Why you should use R* [Video]. YouTube. https://www.youtube.com/watch?v=9kYUGMg_14s

R Programming 101 (2019). *How to install R and install R studio* [Video]. YouTube. https://www.youtube.com/watch?v=orjLGFmx6l4

R Programming 101 (2021). *Ggplot for plots and graphs. An introduction to data visualization using R programming* [Video]. YouTube. https://www.youtube.com/watch?v=HPJn1CMvtmI

Simplilearn (2020). *Ensemble learning* [Video]. YouTube. https://www.youtube.com/watch?v=WtWxOhhZWX0

Soriano, P., & Kebabci, C. (n.d.). *Scree plot for PCA explained.* Statistics Globe. Retrieved on June 1, 2023 from https://statisticsglobe.com/scree-plot-pca

Soumya7 (n.d.) *Bagging vs boosting in machine learning.* Geeks for Geeks. Retrieved on June 1, 2023 from https://www.geeksforgeeks.org/bagging-vs-boosting-in-machine-learning/

The 365 Team (2023). *Overfitting vs underfitting: What is the difference.* 365 Data Science. Retrieved on June 1, 2023 from https://365datascience.com/tutorials/machine-learning-tutorials/overfitting-underfitting/

TheDataPost (2020). *K-Means clustering explanation and visualization* [Video]. YouTube. https://www.youtube.com/watch?v=R2e3Ls9H_fc

Thoughty2 (2021). *There was a secret science experiment at the 1906 Plymout fair* [Video]. YouTube. https://www.youtube.com/watch?v=lPnik1VamSM

TileStates (2021). *Logistics regression: The basics – simply explained* [Video]. YouTube. https://www.youtube.com/watch?v=yhogDBEa0uQ