STAT 2650: Statistical Methods for Data Science

2023 Summer Session	
Total Class Sessions: 25 **Class Sessions Per Week: 5** **Total Weeks: 5** **Class Session Length (Minutes): 145** **Credit Hours: 4**	**Instructor: Staff** **Classroom: TBA** **Office Hours: TBA** **Language: English**

Course Description:

The main purpose of this course is to help students obtain knowledge and use of Inferential statistics and machine learning methods while mastering existing packages from Python. Focus will be put on basic approaches for data science such as Bayesian methods, linear and nonlinear regression, correlation estimation and prediction, goodness-of-fit tests, and machine learning. At the end of the course, students will be armed with the necessary methods and tools to lead a data science project. Note that this is a second statistics course using python. It is assumed that you have knowledge of using python to perform basic statistics, and therefore you are familiar with editing and running python scripts.

Textbooks:

It is not mandatory to have these books, but I encourage you to use them as references to supplement the lecture contents.

- VanderPlas, Python Data Science Handbook.
- Grus, Data Science from Scratch.
- Bruce et al., Practical Statistics for Data Scientists.

Course Assignments:

3 lab-like assignments, 2 Midterm Exams, 1 Project and Final Exam.

- Each Friday, during class time you will work on a lab-like assignment where you will be guided through hints to achieve the expected outcomes. You are supposed to upload your answers in one PDF file at the end of the lecture session.
- A midterm test is scheduled during class time on Monday, June 13, 2022. ad on Wednesday June 22, 2022. The expected outcome is a python file containing the code and a PDF file with results and conclusions.
- A final Exam is scheduled during class time on Thursday June 30, 2022. The expected outcome is a python file containing the code and a PDF file with results and conclusions. Note that for the final assignment, you may not get instructions about the analysis methods to be used in some/all exercises.
- A "Project Report" is due by the end of the day, Wednesday June 29, 2022. Note that you have to prepare and submit a progress report during class time on Friday June 24, showing

your progress and obstacles you might face. Guidance might be provided to those who have difficulties running the analysis project. Details about the project are detailed below.

Project description:

The project will focus on solving a machine learning problem from a dataset prepared for this purpose. The choice of subject is free and you will find a list of suggestions at Awesome **Public Datasets https://github.com/awesomedata/awesome-public-datasets#healthcare**

The directory references many interesting data sources. However, for some resources, it will be necessary to follow several steps before obtaining the desired dataset in csv format. I recommend that you first choose a theme that interests you before looking for the data.

For example if you choose "Health" then "Coronavirus (Covid-19) Data in the United States" you will be redirected to the directory https://github.com/nytimes/covid-19-data. In the README.md, you will find the description of this data. If you choose "colleges" you will have access to the raw data which you can then upload and use https://github.com/nytimes/covid-19-data/blob/master/colleges/colleges.csv

Expected work:

You are being asked to pose a problem and answer it using data. The project is more machine learning oriented.

You are expected to deliver the outcomes in the form of a Python Project (code) and a Report (PDF) that covers the following points:

- An introduction: what is the problem to be solved, the questions to be answered, an overview of the data.
- Particular attention to the rigor of the approach (learning / test basis, overfitting, cross-validation, verification on a few examples, type of variable - continuous, discrete, categorical).
- Comparison of two models on the same dataset (either two different models, or the same model with different parameters). You will be interested in the observations for which the models are in disagreement. You can also compare learning speed and performance.
- You can also highlight the reasoning or intuition that leads you to try such a model, such a feature, such a method.
- The presence of at least one graph.
- A conclusion: what are the fundamental conclusions from the model and the results.

You will have to submit a draft report on June 24, where you are expected to present the progress of your work in PDF format.

I encourage you to work in groups (2-4 students per group) for the project and I strongly advise you to discuss your analyzes with your classmates. The spirit of the project is rather collaborative than competitive.

Approximate grading:

Report: 3 pts Graphics: 2 pts Scientific approach: 3 pts Code: 2pts

Course Assessment:

Your final grade will be computed as follows:

3 lab-like assignments: 15% (5% each)

Project : 30 %

2 Midterm Exams: 30% (15% each)

Final Exam: 25%

Grading Scale (percentage):

A+	A	A-	B+	B	B-	C+	C	C-	D+	D	D-	F
98-100	93-97	90-92	88-89	83-87	80-82	78-79	73-77	70-72	68-69	63-67	60-62	<60

Academic Integrity:

Students are encouraged to study together, and to discuss lecture topics with one another, but all other work should be completed independently.

Students are expected to adhere to the standards of academic honesty and integrity that are described in the Chengdu University of Technology's Academic Conduct Code. Any work suspected of violating the standards of the Academic Conduct Code will be reported to the Dean's Office. Penalties for violating the Academic Conduct Code may include dismissal from the program. All students have an individual responsibility to know and understand the provisions of the Academic Conduct Code.

Special Needs or Assistance:

Please contact the Administrative Office immediately if you have a learning disability, a medical issue, or any other type of problem that prevents professors from seeing you have learned the course material. Our goal is to help you learn, not to penalize you for issues which mask your learning.

Tentative Course Schedule:

Please use this as an approximate class schedule; section coverage may change depending on the flow of the course.

Monday	*Tuesday*	*Wednesday*	*Thursday*	*Friday*

May 30 Course Policy, Course Syllabus Course Overview	**May 31** Lecture 1: Statistical inference with a real example (Poisson and Exponential Distribution,Bootstrap Sampling, Confidence Interval)	**June 01** Lecture 2: Statistical hypothesis testing (Normal distribution, T-Test, Z-Test, ANOVA Test)	**June 02** Lecture 3: Regression (Linear, Polynomial, Logistic)	**June 03** **Lab-like assignment 1**
June 06 Lecture 4: Chi-squared Goodness of Fit Test	**June 07** Lecture 5: Predictions with scikit-learn	**June 08** Lecture 6: Gaussian naive Bayes classification	**June 09** Lecture 7: Dimensionality reduction algorithms	**June 10** **Lab-like assignment 2**
June 13 **Mid-term Exam 1**	**June 14** Lecture 8: Principal component analysis	**June 15** Lecture 9: Canonical correlation analysis	**June 16** Lecture 10: Linear discriminant analysis	**June 17** **Lab-like assignment 3**
June 20 Lecture 11: Support vector machines	**June 21** Lecture 12: Decision tree and random forests	**June 22** **Mid-term Exam 2**	**June 23** Lecture 13: K-means clustering	**June 24** **Project progress checkpoint**
June 27 Lecture 14: Gaussian mixture models	**June 28** Lecture 15: Evaluation metrics for machine learning	**June 29** Lecture 16: Introduction to deep learning	**June 30** **Final Exam**	**July 1** Project submission