RUTGERS
UNIVERSITY | CAMDEN
School of Business

53:716:535:01 BIG DATA ANALYTICS

Fall AY-2024-25 / Credits: 3
Pre/co-requisite(s): 53:716:535 Business Analytics

In-Class Learning

Instructor: Can Kucukgul

Office: BSB 331
Email: can.kucukgul@rutgers.edu
Office hours*: Use the Office Hour link in Announcements.
Thursdays 6:15 - 8:15pm

TABLE OF CONTENTS

COURSE DESCRIPTION

In search of a new frontier of competitive differentiation, companies are increasingly turning towards large-scale data gathering and analytics. As a business school student, do you feel ready for this era of competing through analytics in the data economy?

Adopting a learning-through-examples approach, this course focuses on leveraging the Python programming language and its data science ecosystem for exploratory and predictive analysis. This course provides students with an in-depth introduction to the (big data management platforms) Spark ecosystems, which is an environment used by companies to store and manipulate "Big Data" of a size and scale that cannot be handled by traditional databases.

Although no prior programming experience is required to enroll in this course, students without prior programming experience must be highly motivated to conquer a steep learning curve. To get the most out of this graduate course, students must plan to engage in intensive out-of-class practice based on the topics and techniques introduced in class sessions.

LEARNING OBJECTIVES

Students who complete this course are expected to develop a working knowledge of:
- The Python programming language and its ecosystem for data science
- Data pre-processing for analysis
- Exploratory analysis
- Statistical inference
- Predictive analytics
- Presentation and visualization for conveying data-driven insights
- Extraction, handling, and manipulation of big, unstructured data

COURSE GRADING

Component	Score
Homework Assignments (3)	54 points (18 each)
Exam	26 points
Final Project	20 points
Total = 100 points	

COURSE MATERIALS

- **Course website and communication**
  - o We will use Canvas as the primary place to host course-related information. All class materials including assignments will be posted on Canvas. To access this system, go to http://canvas.rutgers.edu log in, and click on the course on the dashboard.
  - o ***Lecture Notes and Recordings will be posted prior to inception of the respective week (for the exact dates, see detailed course schedule below).***
- **Resources**

  There is no need to buy a textbook for this course, *lecture notes will be comprehensive and self-sufficient*. The instructor has curated the materials for this course based on multiple books, academic papers, and online tutorials posted by other scholars.

  Suggested primary references:

  - Textbooks:
    - *Introduction to Data Mining, 2nd edition, by Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, Vipin Kumar; Pearson publishing.*
    - *Mining of Massive Datasets, 3rd edition, by Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman.*

  - Basic Python:
    - The *Python Tutorial* and the *Python for Everybody* webpage are excellent resources for learning the Python programming language.
    - *Think Python*, second edition, by Allen Downey. A PDF version of the book is available here.
    - Automating the Boring Stuff with Python by Al Sweigart.

  - Modeling concepts (for curious minds): The following books are not specific to Python, but they provide good coverage of the theoretical concepts we'll utilize in this course:
    - Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
    - The Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani, and Jerome Friedman.

- **Computing and Software**
  - Local installation option: all software packages we'll use in the course are freely available and can be installed in Windows, Mac OS, and Linux.
    - Python: We will use the Anaconda Package (Python 3.9 version) specifically customized for data science needs. If students are using their own computers, they should download and install Anaconda. This guide can be used to verify the installation.
    - Among the many convenient editors available for Python, we will primarily use Jupyter Notebook which is already packaged with Anaconda (no separate installation is required).

ASSESSMENTS - GRADED DELIVERABLES

HOMEWORKS

There will be three homeworks related to data analytics methods that are covered in the lectures. A total of 54 points (18 points per homework) can be earned by thorough completion of these tasks. Homeworks will help you build theoretical knowledge, and work on data analytics applications on your own. Exercises similar to homeworks will be solved/demonstrated within Python tutorials in-class.

EXAM

- There will be a comprehensive, open notes/open book, in-class exam.

FINAL PROJECT

- The course's final project requires students to create a report up to 10 pages with insights from examining a dataset using the techniques learned from the course.
- Submissions must include all supporting Python and data files.
- The final project will test both data analysis and presentation skills. The submissions for the project will be evaluated for: (1) rigor and thoroughness of data analysis, (2) creative representation of analysis results, (3) the level of challenge in the analysis, and (4) organization of the project deliverables. A detailed grading rubric will be available in Canvas later in the semester.
- In addition, the following rules apply:
    - A student's report must be fully supported by data-driven evidence demonstrated in the student's Jupyter Notebook file(s). All misalignments between the report and its supporting Jupyter Notebook file will receive a penalty.
    - The Report and the accompanying Jupyter Notebook file must be properly organized and will be graded for readability. They should be self-sustained documents explaining student's analysis.

COURSE OUTLINE AND SCHEDULE

Please note that this planned schedule may change. Refer to the course page in Canvas for the up-to-date schedule.

Weeks - Dates	Topic	Out-of-Class	Assignment
#1 – Mon, Sep 9	Introduction to data mining and big data	Tutorial 1	
#2 – Mon, Sep 16	Data types and data preprocessing	Tutorial 2	HW-1 Assigned
#3 – Mon, Sep 23	Data exploration + visualization	Tutorial 3	
#4 - Mon, Sep 30	Decision trees + Overfitting		HW-2 Assigned, HW-1 Due (Oct 2, 11:59pm)
#5 - Mon, Oct 7	Classification: Rule-based + knn + naïve bayes		
#6 - Mon, Oct 14	Classification: SVM + Ensemble	Tutorial 4	HW-3 Assigned, HW-2 Due (Oct 16, 11:59pm)
#7 - Mon, Oct 21	Clustering + Association rule mining	Tutorial 5	
#8 - Mon, Oct 28	Data-driven Optimization - Introduction to Gurobi		HW-3 Due (Nov 3, 11:59pm)
#8.5 - Mon, Nov 4	REVIEW for EXAM		
#9 - Mon, Nov 11	EXAM (In-class)		
#10 - Mon, Nov 18	Midterm wrap-up, Project introduction, Relevant readings		
#11 – Mon, Nov 25	Big Data: Hadoop + MapReduce		
#12 – Mon, Dec 2	Big Data: Spark	Tutorial 6	
#13 – Mon, Dec 9	Big Data: Applications with PySpark	Tutorial 7	
#14 – Mon, Dec 16	FINAL PROJECT SUBMISSIONS DUE		

Grade Ranges

Letter Grade Description

A   Highest grade (92% and above)
B+   Work of distinction (87.5% to 91.9%)
B   Work of distinction (81.5% to 87.4%)
C+   Average work (75.5% to 81.4%)
C   Average work (69.5% to 75.4%)
D   Passing, but unsatisfactory (60% to 69.4%)
F   Failure without credit (Below 60%)

UNIVERSITY POLICIES

Academic Integrity

Students The Academic Integrity policy can be found at http://studentconduct.rutgers.edu/student-conduct-processes/academic-integrity/. **Students are responsible for understanding the principles of academic integrity and abiding by them in all aspects of their work at the University.** Students are also encouraged to help educate fellow students about academic integrity and to bring all alleged violations of academic integrity they encounter to the attention of the appropriate authorities.

Academic Integrity means that you (the student) must:

- properly acknowledge and cite all use of the ideas, results, or words of others,
- properly acknowledge all contributors to a given piece of work,
- make sure that all work submitted as your own in a course activity is your own and not from someone else
- obtain all data or results by ethical means and report them accurately
- treat all other students fairly with no encouragement of academic dishonesty

Adherence to these principles is necessary in order to ensure that:

- everyone is given proper credit for his or her ideas, words, results, and other scholarly accomplishments
- all student work is fairly evaluated, and no student has an inappropriate advantage over others
- the academic and ethical development of all students is fostered
- the reputation of the University for integrity is maintained and enhanced.

Failure to uphold these principles of academic integrity threatens both the reputation of the University and the value of the degrees awarded to its students. Every member of the University community therefore bears a responsibility for ensuring that the highest standards of academic integrity are upheld. Violations are taken seriously and will be handled according to university policy.

Disability Services
The University is committed to supporting the learning of all students and faculty will provide accommodations as indicated in a Letter of Accommodation issued by the Office of Disability Services (ODS). If you have already registered with ODS and have your letter of accommodations, please share this with me early in the course. If you have or think you have a disability (learning, sensory, physical, chronic health, mental health or attentional), please contact https://success.camden.rutgers.edu/disability-services.

Accommodations will be provided only for students with a letter of accommodation from ODS. Their ser-vices are free and confidential. Letters only provide information about the accommodation, not about the disability or diagnosis.

Rutgers email
All communications to students will be done using the Rutgers email address provided to you.
Please forward your Rutgers email to your personal email if necessary. **Not checking your Rutgers email is not an excuse for missing any communications.**

Pronouns
This course affirms people of all gender expressions and gender identities. Feel free to correct me on your preferred gender pronoun. If you have any questions or concerns, please do not hesitate to contact me.

Chosen name (Preferred Name)
If you have a chosen name or preferred name other than what is listed on the roster, kindly let me know. If you would like to have your name changed within the rosters officially, go to: https://deanofstudents.camden.rutgers.edu/chosen-name-application.

Canvas
Posted will be the syllabus, resources (articles and examples), Power Point slides, announcements, guides, etc. To access this system, go to http://canvas.rutgers.edu log in, and click on the course on the dashboard.

Accessibility
If you have difficulties with accessing materials posted on Canvas, please alert the instructor.

Code of Student Conduct
Rutgers University-Camden seeks a community that is free from violence, threats, and intimidation; is respectful of the rights, opportunities, and welfare of students, faculty, staff, and guests of the University; and does not threaten the physical or mental health or safety of members of the University community, including in classroom space.

As a student at the University, you are expected adhere to the Code of Student Conduct. To review the code, go to the Office of Community Standards: https://deanofstudents.camden.rutgers.edu/student-conduct

Note that the conduct code specifically addresses disruptive classroom conduct, which means *"engaging in behavior that substantially or repeatedly interrupts either the instructor's ability to teach or student learning. The classroom extends to any setting where a student is engaged in work toward academic credit or satisfaction of program-based requirements or related activities."*

The following protocols on the codes of behavior reflect professional business norms on manners, courtesy, and respect. (***In general, you should treat others as you would like others to treat yourself. Be mindful that what is acceptable in a text or chatroom with friends may not be appropriate in a classroom or in an online conversation with an instructor***.)

Even though many of you are already aware of these protocols, they are explicitly stated here so that everyone is cognizant of the same protocols. These protocols should be followed by all students taking this course to help ensure the online experiences for everyone involved are pleasant. They are as follows:

- If you were to send an e-mail to the professor / tutor, please address the person appropriately such as "Dr. …." or "Mr./Ms. …." , not 'Hey'. Note that I will address you with your first name, unless you prefer that I address you differently.
- When sending an e-mail, you can get better attention by using the following guidelines:
  - use descriptive subject lines, (I am sure you have received a lot of Spam e-mails and I have occasionally and accidentally deleted student e-mails that I thought was Spam.)
  - please be as brief as possible by going straight to the point, and if possible, limit the use of attachments

Never send offensive and insulting messages* (this is a violation of the Student Code of Conduct). If you disagree, say so and state your reasons. Social media is a very powerful tool for communication. However, it can be badly misused if it is not used correctly or professionally. For example, you may have personal and legitimate concerns with this course. However, other students, who do not have similar feelings, can be negatively influenced by your concerns. This will unnecessarily and negatively affect their overall experience

of the course. Therefore, to minimize such an occurrence, please feel free to directly contact the instructor/tutor first to resolve any concerns that you may have to help ensure that everyone's online experience of this course is beneficial.

- Always guard against inciting others when it comes to content, opinions, etc. That is, avoid blaming or accusing others of wrongdoing.
- Do not start a volley of back-and-forth e-mails, with copies distributed to every student in the class.
- Copy the minimum number of people. That is, send e-mails to only the people you think should receive and will benefit from it.
- Treat all e-mails and postings as permanent forms of written record and do not expect that any your e-mail communications to be private, unless stated otherwise. Instead, assume that all e-mail communications are public.
- Do not publicize your own or others' personal information (such as email, phone numbers, last names, etc.).

CLASSROOM POLICIES

Exam Make-up Policy/Late Policy

If, for a university approved reason, you cannot take an exam at the scheduled time you must give the professor written notice at least two weeks in advance so that other arrangements can be made. If the situation does not allow for advance notification (for example, emergency hospitalization), contact the professor as soon as possible after a missed exam. Make-up exams for non-university approved reasons are not guaranteed. The professor reserves the right to request written documentation to support your absence (such as a doctor's note, an obituary, or military orders).

Assessments

More detail on the deliverables will be provided during the first lecture.

CLASS PARTICIPATION: It is expected that you will (1) attend class regularly and arrive on time, (2) listen attentively in class, and (3) contribute often to class discussions. Your in-class comments should be thoughtful and should reflect your careful reading of the assigned course material.

Participation and Late Work: Lack of participation will be reflected in the final grade. All assignments must be handed in on time; late work will receive reduced or no credit. No makeup exams will be scheduled without prior notification and a physician's excuse.

Incompletes and Problems: If you find that you are having trouble completing course work or need further explanation of class topics, please schedule an appointment with me immediately. If you need this class for graduation, you should be sure that your performance is up to standard throughout the course. It is too late to wait until the last week of classes to ask for help. I'm available to meet throughout the entire semester if you need help. "Incompletes" will only be given through prior consultation, under extreme circumstances