UNIVERSITY OF
SASKATCHEWAN

COURSE SYLLABUS

COURSE TITLE:	**Data Science for Epidemiology**		
COURSE CODE:	CHEP 898	**TERM:**	2025 Winter
COURSE CREDITS:	3	**DELIVERY:**	In person
CLASS SECTION:	CRN 26905	**START DATE:**	January 8th, 2025
CLASS LOCATION:	HLTH 3100	**LAB LOCATION:**	
CLASS TIME:	Wednesdays 9:00- 11:50 am	**LAB TIME:**	
WEBSITE:			

Course Description

This course introduces students to the principles of data science as applied to epidemiological research. Emphasis is on data wrangling, version control with Git and GitHub, high-performance computing, and machine learning techniques. It also compares traditional epidemiologic analysis approaches with contemporary machine learning methods.

Prerequisites

1. Graduate level course in biostatics or statistics
2. Introductory or graduate level course in epidemiology

Enrollment Limit

10

Land Acknowledgement

I acknowledge our shared connection to the land and recognize that Indigenous and Métis peoples on Treaty 6 Territory and all Indigenous peoples have been and continue to be stewards for social justice, equity, and land-based education. In the spirit of reconciliation may we all strive to learn and support the work of Indigenous communities as allies.

Artificial Intelligence

This course will follow the general USask Guidelines about AI for Educators and Students (https://leader ship.usask.ca/initiatives/ai/index.php). The University has developed high level guidance based on the European Network for Academic Integrity (ENAI) recommendations. The principles are descriptions of USask intentions for, and beliefs about, the use of AI. They include 4 categories:

- Ethical and Responsible Use
- Literacy
- Tool Use
- Change and Innovation

AI Rules for this course

In general, my opinion is that you should be exploring these tools, what they can do, and how you can integrate them into your work. These tools are great for editing, formatting, generating ideas, and writing very basic code. USask faculty and students have access to Microsoft Co-Pilot (https://teaching.usask.ca/learning-technology/tools/microsoft-copilot.php). It's critical that when you use these tools you are very aware of bias and that you intervene to correct the text. Here are my general rules for AI in this course.

1. You can use AI tools for any or all parts of the work.
2. If you do you must cite your work (as above). 2.1. Acknowledge AI tools: “All persons, sources, and tools that influence the ideas or generate the content should be properly acknowledged” (p. 3). Acknowledgement may be done in different ways, according to context and discipline, and should include the input to the tool. 2.2. Do not list AI tools as authors: Authors must take responsibility and be accountable for content and an AI tool cannot do so. 2.3. Recognize limits and biases of AI tools: Inaccuracies, errors, and bias are reproduced in AI tools in part because of the human produced materials used for training.
3. If you do you must include a 500 word reflective essay about the experience as part of your self-evaluation.
4. Be very careful with reference. Many of these tools just make up random references.
5. I will not use tools like GPTZero to detect whether you have used AI tools or not. We are making an agreement to be honest with each other here. This is small class. We have that luxury.

Contact Information

Dr. Daniel Fuller daniel.fuller@usask.ca

Learning Outcomes

1. Understand the basics of data wrangling and data management in epidemiology.
2. Gain proficiency in using Git and GitHub for version control.
3. Learn to leverage high-performance computing resources for epidemiologic data analysis.
4. Explore various machine learning techniques and their applications in epidemiology.
5. Compare and contrast traditional epidemiological analysis methods with machine learning approaches.

Readings/Textbooks

There is not one textbook for this course. We will use various components of different open access resources.

- R for Data Science (2e). 2024. Hadley Wickham, Mine Çetinkaya-Rundel, and Garrett Grolemund. https://r4ds.hadley.nz/
- An Introduction to Statistical Learning with Applications in R (2e). 2024. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. https://www.statlearning.com/
- Learn Tidymodels. https://www.tidymodels.org/learn/

Other Required Materials

Use of a statistical software program (R) is required for this course. You will also be asked to install other software including PostGRES (SQL) and Git.

Dataset

In this course we will use the CanPath Student Dataset that provides students the unique opportunity to gain hands-on experience working with CanPath data. The CanPath Student Dataset is a synthetic dataset that was manipulated to mimic CanPath's nationally harmonized data but does not include or reveal actual data of any CanPath participants.

The CanPath Student Dataset is available to instructors at a Canadian university or college for use in an academic course, at no cost. CanPath will provide the Student Dataset and a supporting data dictionary.

- Large sample size (Over 40,000 participants)
- Real-world population-level Canadian data
- Variety of areas of information allowing for a wide range of research topics
- No cost to faculty
- Potential for students to apply for real CanPath data to publish their findings

General Class Schedule

Week	Date	Topic	Data Work	Assignment Due
1	January 8	Intro to Data Science	Intro R	
2	January 15	R Wrangling and Visualization	Data Wrangling	
3	January 22	Version Control with Git/Github	Data Visualization	Data Wrangling
4	January 29	Missing Data	Missing Data	Version Control
5	February 5	Linear Regression	Linear Regression	Missing Data
6	February 12	Logistic Regression	Logistic Regression	
7	February 19	Reading Week		
8	February 26	Scientific Computing	Scientific Computing/Big Data	Independent Analysis 1
9	March 5	Causal Inference	Causal Quartet	Scientific Computing
10	March 12	Support Vector Machines	Random Forest	
11	March 19	Random Forest	Matching	Random Forest
12	March 26	Matching Methods	SVM	Matching
13	April 2	Artificial Neural Networks	ANN	Independent Analysis 2

- Subject to change depending on speed

Attendance and Participation

Attendance and participation and reading ahead are critical to this course. There will a lot of time for discussion and working on assignments allocated in this course but reading ahead is a critical aspect of the learning process.

Assignment Grading Scheme

Assignment	Grade %
Data Wrangling and Visualization	10%
Github	10%
Missing Data	15%
Independent Analysis - Part 1	10%
Random Forest	15%
Scientific Computing/Big Data	10%
Matching	15%
Independent Analysis – Part 2	15%
Total	100%

Assignment Descriptions

Data Wrangling and Visualization

Value: 10% of final grade
Due Date: See Course Schedule
Type: This assignment will have students work with fundamental skills of data science and submit via an RMarkdown file.
Description: In this assignment you will complete a data wrangling assignment that will involve data cleaning, descriptive statistics, understanding missing data, and joining datasets together.

Github

Value: 10% of final grade
Due Date: See Course Schedule
Type: This assignment will have students work version control systems and submit their assignment to their own Github repository.
Description: In this assignment you will create a Github account, install Git on your local computer, create a Github repository and commit and push your work to that Github repository.

Missing Data

Value: 15% of final grade
Due Date: See Course Schedule
Type: This assignment will have students understand the basic and advanced concepts of missing data handling.
Description: In this assignment you will apply and compare different methods for imputing missing data on large health administrative dataset.

Independent Analysis 1

Value: 10% of final grade
Due Date: See Course Schedule
Type: This assignment will have students conduct the first part of an independent data science workflow
Description: This is part 1 of the independent analysis. You will need to find a dataset, develop an analysis plan to includes the major components of the course (ie., Github, Scientific Computing), and conduct descriptive statistics and data wrangling on your chosen dataset.

Random Forest

Value: 15% of final grade
Due Date: See Course Schedule
Type: This is a code-based assignment where you conduct a Random Forest analysis.

Description: In this analysis you will complete a Random Forest analysis using the Can Path student dataset. You will need to run the analysis, conduct detailed hyperparameter tuning, and conduct model comparisons.

Scientific Computing/Big Data

Value: 10% of final grade
Due Date: See Course Schedule
Type: This is a code-based assignment where you will learn to use an HPC.
Description: In this assignment you will use the USask Plato High Performance Computing to run a large scale machine learning on a large (~1GB) dataset.

Matching 15%

Value: 15% of final grade
Due Date: See Course Schedule
Type: This is a code-based assignment where you will learn to use an HPC.
Description: In this analysis you will complete a machine learning based matching analysis using the Can Path student dataset.

Independent Analysis 15%

Value: 15% of final grade

Due Date: See Course Schedule
Type: This assignment will have students conduct the second and final part of an independent data science workflow.
Description: This is part 2 (final part) of the independent analysis. You will need to conduct a complete analysis including data wrangling, missing data handling, and apply at least 2 different machine learning methods to your data.

Self-Evaluation

Value: 0% of final grade (Formative Evaluation)
Due Date: See Course Schedule
Type: Written report (200 words)
Description: Complete the student self-evaluation form. This is **required** for **each** assignment where you use AI.

Submitting Assignments

All assignments should be submitted to the appropriate place in Canvas or Github. All assignments are due at 5pm (CST) on the due date. Please don't stay up until midnight to get the work done. Remember there are no late penalties so just take an extra day if you need and get some sleep.

Late and Missing Assignments

There is no penalty for late assignments. However, because many assignments have two parts, it is critical to the first assignment of the sections in around the due date. Missing assignments that are not submitted by the end of the course will receive a grade of zero.

Readings

1. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is Machine Learning? A Primer for the Epidemiologist. Am J Epidemiol. 2019 Dec 31;188(12):2222-2239. doi: 10.1093/aje/kwz189. PMID: 31509183.

2. Serghiou S, Rough K. Deep Learning for Epidemiologists: An Introduction to Neural Networks. Am J Epidemiol. 2023 Nov 3;192(11):1904-1916. doi: 10.1093/aje/kwad107. PMID: 37139570.
3. Wiemken TL, Kelley RR. Machine Learning in Epidemiology and Health Outcomes Research. Annu Rev Public Health. 2020 Apr 2;41:21-36. doi: 10.1146/annurev-publhealth-040119-094437. Epub 2019 Oct 2. PMID: 31577910.
4. Fuller D, Buote R, Stanley K. A glossary for big data in population and public health: discussion and commentary on terminology and research methods. J Epidemiol Community Health. 2017 Nov;71(11):1113-1117. doi: 10.1136/jech-2017-209608. Epub 2017 Sep 16. PMID: 28918390.

Additional readings will be provided using material from the online course material.

University of Saskatchewan Grading System (for graduate courses)

The following describes the relationship between literal descriptors and percentage scores for courses in the College of Graduate Studies and Research:

90-100 Exceptional: A superior performance with consistent strong evidence of

- a comprehensive, incisive grasp of subject matter;
- an ability to make insightful, critical evaluation of information;
- an exceptional capacity for original, creative and/or logical thinking;
- an exceptional ability to organize, to analyze, to synthesize, to integrate ideas, and to express thoughts fluently;
- an exceptional ability to analyze and solve difficult problems related to subject matter.

80-89 Very Good to Excellent: A very good to excellent performance with strong evidence of

- a comprehensive grasp of subject matter;
- an ability to make sound critical evaluation of information;
- a very good to excellent capacity for original, creative and/or logical thinking;
- a very good to excellent ability to organize, to analyze, to synthesize, to integrate ideas, and to express thoughts fluently;
- a very good to excellent ability to analyze and solve difficult problems related to subject matter.

70-79 Satisfactory to Good: A satisfactory to good performance with evidence of

- a substantial knowledge of subject matter;
- a satisfactory to good understanding of the relevant issues and satisfactory to good familiarity with the relevant literature and technology;
- a satisfactory to good capacity for logical thinking;
- some capacity for original and creative thinking;
- a satisfactory to good ability to organize, to analyze, and to examine the subject matter in a critical and constructive manner;
- a satisfactory to good ability to analyze and solve moderately difficult problems.

60-69 Poor: A generally weak performance, but with some evidence of

- a basic grasp of the subject matter;
- some understanding of the basic issues;

- some familiarity with the relevant literature and techniques;
- some ability to develop solutions to moderately difficult problems related to the subject matter;
- some ability to examine the material in a critical and analytical manner.

<60 Failure: An unacceptable performance.

Program Requirements

- Percentage scores of at least 70% are required for a minimal pass performance in undergraduate courses taken by graduate students;
- Percentage scores of at least 70% are required for a minimal pass performance for each course which is included in a Ph.D. program;
- Percentage scores of at least 70% are required for a minimal pass performance in all courses used toward JSGS Public Policy and Public Administration programs and all core courses for Master of Public Health students, whether included in a Ph.D. program or a Master's program;
- For all other graduate courses, percentage scores of at least 60-69% are required for a minimal pass performance for each course which is included in a Master's program, provided that the student's Cumulative Weighted Average is at least 70%;
- Graduate courses for which students receive grades of 60-69% are minimally acceptable in a Postgraduate Diploma program, provided that the Cumulative Weighted Average is at least 65%;
- Students should seek information on other program requirements in the Course & Program Catalogue and in academic unit publications.

Access and Equity Services (AES)

Access and Equity Services (AES) is available to provide support to students who require accommodations due to disability, family status, and religious observances. Students who have disabilities (learning, medical, physical, or mental health) are strongly encouraged to register with Access and Equity Services (AES) if they have not already done so. Students who suspect they may have disabilities should contact AES for advice and referrals at any time. Those students who are registered with AES with mental health disabilities and who anticipate that they may have responses to certain course materials or topics, should discuss course content with their instructors prior to course add / drop dates. Students who require accommodations for pregnancy or substantial parental/family duties should contact AES to discuss their situations and potentially register with that office. Students who require accommodations due to religious practices that prohibit the writing of exams on religious holidays should contact AES to self-declare and determine which accommodations are appropriate. In general, students who are unable to write an exam due to a religious conflict do not register with AES but instead submit an exam conflict form through their PAWS account to arrange accommodations. Any student registered with AES, as well as those who require accommodations on religious grounds, may request alternative arrangements for mid-term and final examinations by submitting a request to AES by the stated deadliness. Instructors shall provide the examinations for students who are being accommodated by the deadlines established by AES. For more information or advice, visit https://students.usask.ca/health/centres/access-equity-services.php, or contact AES at 306-966-7273 (Voice/TTY 1-306-966-7276) or email aes@usask.ca.

Academic Integrity

The University of Saskatchewan is committed to the highest standards of academic integrity and honesty. Students are expected to be familiar with these standards regarding academic honesty and to uphold the policies of the University in this respect. Students are particularly urged to familiarize themselves with the provisions of the Student Conduct & Appeals section of the University Secretary Website and avoid any behavior that could potentially result in suspicions of cheating, plagiarism, misrepresentation of facts and/or participation in an offence. Academic dishonesty is a serious offence and can result in suspension or expulsion from the University. All students should read and be familiar with the Regulations on Academic Student Misconduct (https://governance.usask.ca/student-conduct-appeals/academic-misconduct.php) as well as the Standard of Student Conduct in Non-Academic Matters and Procedures for Resolution of Complaints and Appeals (https://governance.usask.ca/student-conduct-appeals/non-academic-misconduct.php) For more information on what academic integrity means for students see the Academic Integrity section of the University Library Website at: https://library.usask.ca/academic-integrity.php You are encouraged to complete the Academic Integrity Tutorial to understand the fundamental values of academic integrity and how to be a responsible scholar and member of the USask community - https://libguides.usask.ca/AcademicIntegrityTutorial There are also valuable resources on the Integrity Matters website: https://academic-integrity.usask.ca/

Copyright

Course materials are provided to you based on your registration in a class, and anything created by your professors and instructors is their intellectual property and cannot be shared without written permission. If materials are designated as open education resources (with a creative commons license) you can share and/or use in alignment with the CC license. This includes exams, PowerPoint/PDF slides and other course notes. Additionally, other copyright-protected materials created by textbook publishers and authors may be provided to you based on license terms and educational exceptions in the Canadian Copyright Act (see http://laws-lois.justice.gc.ca/eng/acts/C-42/index.html). Before you copy or distribute others' copyright-protected materials, please ensure that your use of the materials is covered under the University's "Use of Materials Protected By Copyright" Policy available at https://policies.usask.ca/policies/operations-and-general- administration/copyright.php. For example, posting others' copyright-protected materials on the open internet is not permitted by this policy or by the university Copyright Guidelines (available at https://library. usask.ca/copyright/general-information/copyright-guidelines.php) and requires permission from the copyright holder For more information about copyright, please visit https://library.usask.ca/copyright/ where there is information for students available at https://library.usask.ca/copyright/students/your-course-materials.php, or contact the University's Copyright Coordinator at copyright.coordinator@usask.ca or 306-966-8817.

Student Supports

Academic Support for Students

Visit the Learning Hub to learn how the University Library supports undergraduate and graduate students. Attend online or in-person workshops, review online resources or book 1-1 appointments for help with * First year experience * Research * Study strategies and skills * Writing * Math and Statistics

Teaching, Learning and Student Experience

Teaching, Learning and Student Experience (TLSE) provides developmental and support services and programs to students and the university community. For more information, see the students' website http://students.usask.ca.

Financial Support

Any student who faces unexpected challenges securing their food or housing and believes this may affect their performance in the course is urged to contact Student Central https://students.usask.ca/student-central.php.

Aboriginal Students' Centre

The Aboriginal Students' Centre (ASC) is dedicated to supporting Indigenous student academic and personal success. The ASC offers personal, social, cultural and some academic supports to Métis, First Nations, and Inuit students. The ASC is in the Gordon Oakes Red Bear Students Centre, which is an intercultural gathering space that brings Indigenous and non-Indigenous students together to learn from, with and about one another in a respectful, inclusive, and safe environment. Visit https://students.usask.ca/indigenous/index.php or students are encouraged to visit the ASC's Facebook page https://www.facebook.com/aboriginalstudentscentre/

International Student and Study Abroad Centre

The International Student and Study Abroad Centre (ISSAC) supports student success and facilitates international education experiences at USask and abroad. ISSAC is here to assist all international undergraduate, graduate, exchange, and English as a Second Language students in their transition to the University of Saskatchewan and to life in Canada. ISSAC offers advising and support on matters that affect international students and their families and on matters related to studying abroad as University of Saskatchewan students. Visit https://students.usask.ca/international/issac.php for more information.