AFLYST Social Data Science (Advanced methodology course – collection, treatment and analysis of data)

Title	AFLYST Social Data Science (Advanced methodology course – collection, treatment and analysis of data)
Semester	E2022
Master programme in	Socialvidenskab / Virksomhedsstudier / Politik / Politik og forvaltning / Virksomhedsledelse / Business Administration and Leadership / Politik og Forvaltning / Socialvidenskab / Virksomhedsledelse
Type of activity	
Teaching language	English
Study regulation	

REGISTRATION AND STUDY ADMINISTRATIVE

Registration	Sign up for study activities at STADS Online Student Service within the announced registration period, as you can see on the Study administration homepage. When signing up for study activities, please be aware of potential conflicts between study activities or exam dates. The planning of activities at Roskilde University is based on the recommended study programs which do not overlap. However, if you choose optional courses and/or study plans that goes beyond the recommended study programs, an overlap of lectures or exam dates may occur depending on which courses you choose.
Number of participants	
ECTS	5
Responsible for the activity	Kim Sass Mikkelsen (ksass@ruc.dk)
Head of study	Margit Neisig (neisig@ruc.dk)
Teachers	
Study administration	ISE Studyadministration (ise-studyadministration@ruc.dk)
Exam code(s)	U60397

ACADEMIC CONTENT

Overall objective	

Detailed description of content

Data are everywhere, and big data has become a buzzword within public policy. Organizations and decision makers both in and outside the public sector have more data available at their disposal than ever before, and are using them in ever more ways. Policies are being adopted and adapted in response to data flows about user groups; big data are changing and augmenting public service delivery; and everywhere decisions and outcomes for target groups from unemployed to tax payers and back are being predicted using machine learning techniques. In the academy as well, this type of social data science is of increasing relevance as researchers use machine learning techniques for a host of purposes, such as analysis of text and other unstructured data and studying how different groups react differently to the same policy interventions.

The vision of this course is that students introduced to the practice of predictive data analysis are better equipped for operating in a data driven public sector. Understanding social data science can help students solve practical problems, know what can be done – and what should not be done – using big data and predictive modeling, and support colleagues with data science backgrounds in developing solutions that can make policy and service production better for everyone.

The course starts from what you know from introductory statistics courses, and introduces you to social data science from a practical perspective. We expand your ability to handle and visualize data. We deepen your understand of the statistical models you already know. And we introduce you to a series of relatively simple machine learning models (e.g., clustering algorithms, decision trees, and k-nearest neighbor techniques). Along the way, we introduce issues with the implementation of machine learning for policy and public sector purposes.

The course also introduces the statistical software R. R is a powerful, freely available, open-source program, and is widely used for data science and machine learning among both practitioners and researchers. As an added benefit of the course, therefore, students are freed to implement statistical models and machine learning models also in work settings where employers are unwilling or unable to pay for commercial software.

Course material and Reading list

The course syllabus draws heavily on James et al.'s (2013 & 2021) Introduction to Statistical Learning (cited below). The book is available through the university library. The remaining readings for the course are book chapters and academic articles that either introduce techniques and how to implement them in software or show what these techniques can be used for in a policy or public administration setting. This is not a reading intensive course, as we focus heavily on doing analyses, but some readings take a little while to get through.

A complete list of readings will be available on the course moodle site.

Examples of readings include:

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2021. An introduction to statistical learning. Springer.

Knaflic, C.N., 2015. Storytelling with data: A data visualization guide for business professionals. John Wiley & Sons.

Molnar, C., 2020. Interpretable machine learning. LeanPub.

Rodolfa, K.T., Lamba, H. and Ghani, R., 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. Nature Machine Intelligence, 3(10), pp.896-904.

Asquith, B., Hellerstein, J.K., Kutzbach, M.J. and Neumark, D., 2021. Social capital determinants and labor market networks. Journal of Regional Science, 61(1), pp.212-260.

Overall plan and expected work effort

The course gives 5 ECTS points, corresponding to an expectation that students spend 135 hours in relation to it.

The course comprises 10 double sessions, totalling 20 hours.

It is expected that students have read and worked on syllabus materials prior to each session. It is expected that students will spend around 90 hours on this work. Students can expect a substantial number of these hours to be spent on small at-home exercises using techniques on real data.

The exam is a portfolio comprising small sets of tasks and exercises implementing, interpreting, and discussing results from data analyses. Most of the work on this portfolio is done throughout the semester, with possibilities for editing portfolio elements based on feedback. The expectation is that work on the exam will total 25 hours.

Teaching sessions depart from readings and students work on small at-home assingments, progressively building toward reflected application of machine learning models on real data. Teaching in the sessions may include the following elements:

Questions for the teaching team regarding syllabus content.

Presentations and demonstrations by the teaching team based on syllabus content.

Exercises alone or in ad hoc groups solving concrete problems in data analysis.

Class discussions about the feasibility and utility of perspectives and techniques from the syllabus in practice.

At-home exercises are small data problems posed for students to work on outside class to train syllabus content. Examples include running a particular model or operation on a dataset handed out by the teaching teams.

In-class exercises are similar to at-home exercises but may be longer and more challenging since the in-class setting affords students opportunities to get support.

The portfolio for the exam consists of a select number of in-class and at-home exercises students work on during the course. In addition, the portfolio may contain an element requesting a discussion of the use of machine learning or similar techniques for a particular public sector purpose, e.g. discussing fairness or interpretability.

Format

Campus

Evaluation and feedback

Aktiviteten evalueres regelmæssigt ud fra studienævnets evalueringsprocedure. Den aktivitetsansvarlige vil blive orienteret om en eventuel evaluering af aktiviteten ved semesterstart se link til studienævnets evalueringspraksis her https://intra.ruc.dk/nc/for-ansatte/organisering/raad-naevn-og-udvalg/oversigt-over-studienaevn/studienaevn-for-samfundsstudier/arbejdet-med-kvalitet-i-uddannelserne.

Programme

See moodle.

ASSESSMENT

Overall learning outcomes

Efter endt kursus vil de studerende:

- med faglig relevant terminologi kunne redegøre for og vurdere fordele og ulemper ved at indsamle og analysere data ved hjælp af en givne metode
- være i stand til sikkert og selvstændigt at anvende den given metode i forhold til en specifik faglig problemstilling
- kunne reflektere over forskningsetiske spørgsmål relateret til metoden
- kunne formidle resultater opnået gennem anvendelse af metoden på en faglig præcis måde.

Form of examination

Individuel portfolio

Portfolioen skal have et omfang på maksimalt 24.000 tegn inkl. mellemrum. Produkterne kan f.eks. være øvelsesbesvarelser, talepapirer til præsentation, skriftligt feedback, skriftlige refleksioner og skriftlige opgaver. Udfærdigelsen af produkterne kan være underlagt tidsbegrænsninger.

Omfangskravene er inklusive eventuel forside, indholdsfortegnelse, litteraturliste, figurer og andre illustrationer, men eksklusive eventuelle bilag.

Produkterne til portfolioen udarbejdes helt eller delvist under kursusforløbet.

Portfolioen afleveres samlet (uploades på eksamen.ruc.dk). Evt. løbende delaflevering til den kursusansvarlige med henblik på feedback erstatter ikke den samlede aflevering.

Der foretages en samlet bedømmelse af portfolioen.

Bedømmelse: 7-trinsskala

Form of Re-examination

Samme som ordinær eksamen / same form as ordinary exam

Type of examination in special cases

Examination and assessment criteria

After the course, students are expected to have the ability to:

Assess and discuss advantages and disadvantages of simple supervised and unsupervised machine learning techniques.

	Independently work with data pre-processing, exploratory data analysis, and simple unsupervised and supervised machine learning models in the R statistical environment. Interpret, evaluate, validate, compare, and communicate results from simple machine learning models. Assess and discuss limitations, ethical considerations, and interpretability concerns related to applications of machine learning in a social science and public policy context.
Exam code(s)	Exam code(s) : U60397