Analytics and Application[AA] WS 2024/25

Master of Science
BM Information Systems II / BM Artificial Intelligence and Visual Analytics III
Faculty of Management, Economics, and Social Sciences
Department of Information Systems for Sustainable Society
University of Cologne
Version 1 - October 7, 2024

Instructor Prof. Dr. Wolfgang Ketter
TA Janik Muires
Website `www.is3.uni-koeln.de` and ILIAS
Term WS 2024
Class Time Wed, 14.00-17.30h
Class Location Hörsaal XXIV (WiSo "Schlauch")

Welcome to **Analytics and Application [AA]**. This course and the accompanying reading materials aim to provide the knowledge and skills required for data analytics using information systems that drive business success.

Business Analytics is the use of data-driven decision-making. Companies, governments, and other organizations now collect and have access to large amounts of data about suppliers, clients, employees, citizens, transactions, etc. Data Mining and predictive analytics provide a powerful toolkit for detecting actionable patterns in data and generating predictions. These methods are used in many industries: Mobile companies use their customer database to predict customer churn or to personalize SMS messages to improve customer service; Financial institutions use past loan data to predict defaulting chances for loan applicants; Charities use data from a campaign in one location to target the right people in another location; Politicians use databases of supporters to segment and best target each audience; Movie rental and e-Commerce websites provide recommendations based on users' online behavior; Renewable energy providers use weather forecasts to predict their electricity generation to better trade in different types of markets.

Most business analytics applications are geared toward benefiting the company, often at the expense of the individual, community, and society. This course focuses on human and socially-responsible business analytics, especially with a focus on sustainability.

In this course, we will work with real business problems and real data. We will examine the types of questions that data mining can answer and will develop a variety of data-driven tools to answer these questions. The emphasis is on understanding the concepts and logic behind a wide set of data mining techniques and their relation to specific business analytics situations. The course is not about mastering the theoretical underpinnings of the techniques. The gained knowledge will be applied to a business analytics team project that encapsulates the learnings, while is not expected to master the theoretical underpinnings of the techniques.

You will learn about the process of data analytics. You will learn to identify problems, define the structure of an information system, evaluate competing solutions for a business problem, and you will start to "speak business and analytics". You will gain an insight into how important these activities are in creating information systems that are truly aligned with business needs. Throughout the course, you will also learn about selected topics in sustainability, especially in terms of renewable electricity.

Some of the most important thoughts that we will present throughout the lecture are summarized in this reader and the references therein. This material is designed to help you find additional references, and to help you recapture what was taught in the lectures. It is not designed, however, to serve as an exclusive source of the exam preparation material. In our lectures and workshops, we will present facts, and examples, and teach you skills that are not on the following pages, and we reserve the right to ask you about them during the final exam of your assignments. Please make sure you take advantage of all modes of learning that we offer in this course!

Cologne, October 7, 2024

Prof. Dr. Wolfgang Ketter, Janik Muires
Department of Information Systems for Sustainable Society

Class Schedule

#	Title	Topics	Reading[1]
#1 Oct 09	**(L) Kick-Off** **(L) Intro Data Science**	Intro to course contents & objectives Core ideas and concepts in data mining tasks	Ch. 1 Ch. 2
#2 Oct 16	**(L) Regression 1/2** **(L) Regression 2/2**	Linear Regression & Polynomial Features Cross-Validation & Regularization	Ch. 6
Oct 23	-	No Class	-
#3 Oct 30	**(L) Classification 1/2** **(W) Intro Data Science**	Separating Boundary Classification (SVM) Set-Up Workshops & Coding Intro	Ch. 6
Oct 30	**Start Team Assignment**	See Section 2	-
#4 Nov 06	**(W) Regression 1/2** **(W) Regression 2/2**	Recap & Coding everything Regression "	-
#5 Nov 13	**(L) Naive Bayes** **(W) Classification 1/2**	Probability Recap, Naive Bayes Classifier Recap & Coding SVM	Ch. 8 & 10
#6 Nov 20	**(L) Classification 2/2** **(W) Naive Bayes**	Probabilistic Classification Recap & Coding Naive Bayes	-
#7 Nov 27	**(L) Ensemble Methods** **(W) Classification 2/2**	Trees, Bagging, Boosting Recap & Coding Probabilistic Classification	Ch. 13
Nov 27 23:59	**Deadline Milestone 1**	See Section 2.2	-
#8 Dec 04	**(L) Neural Networks** **(W) Ensemble Methods**	Intro to Artificial Neural Networks (NN) Recap & Coding Ensemble Methods	Ch. 11
#9 Dec 11	**(L) Unsupervised Learning 1/2** **(L) Unsupervised Learning 2/2**	K-Mean, K-means++, Hierarchical clustering Dimensionality Reduction, Fuzzy Clustering	Ch. 15
#10 Dec 18	**(W) Neural Networks** **(W) Unsupervised Learning 1/2**	Recap & Coding NN Recap & Coding Hard Clustering	-
-	Christmas / New Years	Happy Holidays!	-
#11 Jan 08	**(L) Recommender Systems** **(W) Unsupervised Learning 2/2**	Association Rules & Collaborative Filtering Recap & Coding Dim. Reduction & Fuzzy Clustering	Ch. 14
Jan 08 23:59	**Deadline Milestone 2**	See Section 2.2	-
#12 Jan 15	**(L) Time Series** **(W) Recommender Systems**	Time Series Analysis and Forecasting Recap & Coding Recommender Systems	Ch. 16-19
#13 Jan 22	**(L) Text Mining** **(W) Time Series**	Natural Language Processing (NLP) Recap & Coding Time Series	Ch. 20
#14 Jan 29	**(W) Text Mining** **(W) Q&A Session**	Recap & Coding NLP General Q&A session at the end	-
Jan 29 **23:59**	**Deadline Team Assignment**	see Section 2.2	-
Feb 07	**Exam 1**	10:00 - 11:30, Hörsaal XIII	-
Mar 14	**Exam 2**	10:00 - 11:30, Hörsaal XIII	-

Table 1. Tentative Schedule, (L): Lecture Slot, (W): Workshop Slot
[1]: Recommended Reading: Shmueli et al. (2019)

Course Requirements and Resources

1 Programming Software

An important feature of this course is hands-on learning using data mining and data visualization software. For data mining, you will use the free, open-source Python programming software. We recommend using the Anaconda distribution of Python. There are two options:

- **Option 1** (probably the easiest): Install the full Anaconda version, which comes with a lot of pre-installed packages (incl. Jupyter Notebook) and a package manager graphical user interface (GUI). It is available for download **here**.
- **Option 2** (for slightly more advanced users and if you are short on disk space): Install the much lighter MiniConda distribution, which essentially is just the Conda package manager and Python. From there, you can simply install Python packages as required via the Terminal (Mac) or the Command Prompt (Windows). Miniconda is available for download **here**.

We will do a quick run of how to get Python and Jupyter Notebook up and running in Workshop 1. Every student is responsible for bringing a laptop to each workshop session.

2 Online Git Repository

We operate an official git repository for AA. We will use this repo as the central point for sharing workshop materials, codes, data, and instructions with you over the course of this semester. This repo contains the code and (some) data for workshops.

The git repository is hosted on GitHub. The link is the following:

- **Git Address**: `https://github.com/IS3UniCologne/AA_2024.git`

You should check this repository regularly to view and/or download newly released or updated files. If you have experience using git, you may also wish to clone the repository on your local machine for convenience.

3 ILIAS Platform

ILIAS will be used for sharing all lecture-related content, including copies of lecture slides. Lecture slides will be uploaded after each lecture. You can access ILIAS using the following link and password:

- **Course Folder Name: Analytics and Application** (can be accessed **here**)
- **Password:** AA_2024!

4 Textbooks

Core Reading

- **Data Mining for Business Analytics: Concepts, Techniques and Applications in Python** by Galit Shmueli, Peter C. Bruce, Peter Gedeck & Nitin R. Patel, Wiley, 2019, ISBN: 978-1-119-54986-4. [don't buy it, find smarter ways]

Further Reading

- **Python Data Science Handbook** by Jake VanderPlas, O'Reilly, 2016. [available as free .pdf from **here**]
- **An Introduction to Statistical Learning** by Gareth James, Daniela Witten, Trevor Hastie & Robert Tibshirani, Springer, 2013. [availabe as free .pdf **here**]
- **Pattern Recognition and Machine Learning** by Christopher Bishop, Springer, 2006. [availabe as free .pdf from **here**].

Assessment Guidelines

1 General Guidelines

Assessment of individual performance in **AA** will be based on two main components: (1) a **team assignment** (see Section **??** for details) conducted over the second and third part of this course and (2) a **written exam** (see Section 3 for details) taking place at the end of the semester. For the team assignment, teams will be required to deliver a project report. All material covered as part of this course (lectures and workshops) will be exam-relevant, and your understanding of these topics will be tested in the final exam. Please refer to **Table 1** for weights and deadlines for the individual assessment components.

Assignment	Percentage of Grade	Date/Deadline
Team Assignment	33%	Jan 29 23:59
Exam	67%	see Class Schedule
Total	100%	

Table 1. Grading Scheme & Dates/Deadlines

You must pass both the Team Assignment and the Exam to pass this course. This means that you will need to achieve a grade of 4.0 or better in both examination components.

Note: Those students who passed the bachelor's course "Data Science and Machine Learning (DSML)" **cannot** take over their grades from that course, to pass the course Analytics and Applications.

2 Team Assignment

The team assignment runs in parallel to the lectures. Deadline is **February 07th, 23:59**, submission will be through ILIAS. In the following, we will provide details on team composition, working mode, and deliverables for this assignment.

2.1 Team Composition & Working Mode

Task Selection: All teams will work on the same task, which will be announced during lecture #3 (Oct 30th).
Team Composition: The team assignment will be conducted in teams of **5 students**. We will make no exceptions to this rule, save for one *remainder team* if the total number of course participants is not divisible by 5. You will be responsible for forming these teams, so try to make contact with other students in the first weeks.
Team Registration: We will allocate time for you to get to know each other and organize yourself in teams during the first few sessions. We expect you to register these teams until the end of the third week (Oct 27th) by sending an email with a creative **team name** and the **names** and **student IDs** of each group member to **is3-teaching**. The student sending out the mail will be considered point of contact for that team.

2.2 Deliverables

The main deliverable of the team assignment is a **5-page project report** (excl. figures, cover page, executive summary, references, and appendices) which is to be submitted as a .pdf file via ILIAS no later than the deadline specified in Table 1. In addition to the report, we also expect you to upload your Python code in the form of annotated Jupyter notebooks (.ipynb).

Milestones: To guide your progress a little and ensure active participation, there will be two milestone deliverables after four and eight weeks of receiving the task, respectively. **Submission of milestones is mandatory for passing the team assignment**, but will not be graded. This is a learning from previous semesters, where it has shown that students who sign up for the course but do not plan to actively participate corrode teams' performances.

Milestone	Description	Date/Deadline
Data Preparation	Show us that you have started working with the data	Nov 27, 23:59
Analysis & Model	Show us some progress to produce results	Jan 08, 23:59
Team Assignment	Final submission	Jan 29, 23:59

Table 2. Team Assignment Deadlines

Report: The project report details the team project, from the business problem through the data mining problem and solution, to recommendations and practical relevance. As presentation is a key component of a successful data science project, we will consider it in our evaluation. We, therefore, advise you to write your report in LaTeX. To facilitate the writing process, we provide a **LaTeX-template** on Overleaf. Using Overleaf's **documentation**, students can easily learn how to use Latex for scientific projects. Alternatively, you may use Microsoft Word using a similar typeset and line spacing. Please name the file `AA_Team_Report_<yourTeamNumber>.pdf` (e.g., `AA_Team_Report_03.pdf`).

The report should be written clearly and professionally and include the following sections:

1. **Cover page** with informative title, team number, and member names
2. **Detailed report:**
   (a) Problem description (business goal and data mining goal)
   (b) Data description
   (c) Brief data preparation details (how your data was created from the raw data) and key charts. Details can be provided in an Appendix.
   (d) Data analytics: Analytical methods applied (with sufficient detail and screenshots; use Appendix if needed) and appropriate performance evaluation (proper choice of measures, benchmarking).
   (e) Conclusions (advantages and limitations) and business recommendations

Code: Each team submits the code used to generate the findings in their report (solutions as well as figures). You may use Python (the language the course is taught in) for your project; we will not regard other programming languages. There are two options to submit your code:

1. Repository (Preferred): Submit a link to a (public) GitHub repository in which all scripts a collected. Make sure to tag a version that should be regarded for grading before the submission deadline.
2. Archive: Submit a .zip-archive containing all necessary files.

In both cases, please give the notebook files indicative names such as "01_Data_Preparation", "02_Descriptive_Analysis", etc and/or provide a README file explaining the order of execution.

Note that this course does not focus on software engineering, i.e., we do not expect you to write production-ready code. However, your submission should be sufficiently commented for us to understand it.

3 Exam

The final exam contains questions covering material in the workshops and lectures. The exam will be of a 90-minute closed-book format and comprise multiple-choice and written sections. You will receive detailed advice on preparing for it towards the end of the course. We will offer two exams this semester. Please refer to the Class Schedule for exact dates and relevant KLIPS sign-up deadlines.

Appendix A - Course Guidelines

This section contains general rules that apply during the course. These are designed to help us provide you with a high-quality learning experience, and to make transparent what we expect from you in return. Please review these guidelines carefully. If you have any questions, please contact the instructors after looking at the paragraph on communication **(G.2)** below.

(G.1) Course Setup Classes are taught in a seminar-style format where conceptual material is discussed and related to real-world cases. Practical skills are developed in workshops. It might be difficult for you to follow the lectures if you do not study the **recommended readings** for the sessions. We encourage you to also read around the topics to further deepen your understanding of the material. We will upload presentation slides by 12:00 noon on the day of lecture such that you can use it to make notes during the lecture. If updates are made during lecture, we will upload an updated version later that day.

(G.2) Communication To foster a lively conversation within the class, make sure you direct your questions in the following order of priority:

(a) There will be **weekly workshops** that serve as office hours at the same time. We will reserve time for you to ask questions of general interest during these workshops. This is the best place to ask organizational questions and clarification questions on the content of the lectures.

(b) If your question is of general interest but cannot wait until the next workshop, or if your question is better asked in writing, post it to the **AA forum on ILIAS**. We will check the forum regularly and provide answers for everyone's benefit. In particular, all questions about drafts of your assignments must be directed to this forum, quoting from your writeup as necessary. Make sure that your questions are specific and include all relevant details. An example of a specific question is: "In the following performance requirement: [requirement text], is the number of transactions per second a good performance metric or do you know of a better alternative?" An example of a vague question is: "Attached is a file with our second group assignment. Is it good?"

(c) If your question is personal in nature (team conflict, grades, etc.), please send it to **is3-teaching**. This email address is accessible only to the instructors and the teaching assistants of the course. In order for us to be able to optimally follow up with your request, use a subject line of the following format: *[AA] 'subject'*.

(d) If we cannot address your problems in any of these ways, use the course email address (**is3-teaching**) to arrange a personal meeting with us. Please respect the fact that we cannot accommodate walk-in questions, no matter how brief your question may be.

(G.3) Integrity (and LLM) We strictly enforce the university's policies on scholarship and plagiarism. Implicit in handing in homework, assignments, papers, and exams is that they represent your own work (or the result of sanctioned collaboration). Any exceptions must be explicitly noted. Representing someone else's work as your own is grounds for failing the course. We explicitly stress this point in light of recent advances in large language models (LLM) such as GPT. Output largely generated using LLMs does not represent your own work. This does not mean, however, that you are not allowed to use such tools - as complementary assistance, not as ghost-writers. We expect you to be explicit about the usage of tools such as ChatGPT. Specifically, you are expected to clearly indicate generated code (just like you would reference any other external source) including the specific prompt used to generate the output. For further information and guidelines, please refer to `https://verwaltung.uni-koeln.de/stabsstelle02.1/content/faq/data/chatgpt/index_ger.html` (in German).

(G.4) Late submissions Team and individual assignments are due at the dates and times we announce during the course. Late work is not accepted and will be treated as a fail or no-show.

(G.5) Grading Dispute Resolution Please follow the steps below in case of doubts about a grade you received:

- Sign up for a spot on the exam inspection date. As per university guidelines we will publish the dates for this six weeks in advance.
- Look at the grading scheme and your assignment again and send an email to **is3-teaching** explaining specifically where you think the grade is unjustified. Please give references to pages and specific grading comments; we have a copy of your graded assignment in our archives.

– Please respect that we cannot discuss other people's assignments, the relative merits of their writings, or their grades with you. The point of reference for us is the grading scheme and whether or not the grade you received is correct under that scheme.