Learning Goals

After successful completion of this course, the student will be able to:

1. Problem Statement & Research Goal
   a. Illustrate the societal relevance of the thesis research goal.
   b. Formulate (a) clear and specific empirical research question(s) based on identified gaps in literature that lead to solving the research goal.
   c. Organise research questions in a logical and feasible research strategy, with the help of sub-questions.
2. Literature Review
   a. Summarise existing literature regarding methods and results applied to a particular data science problem or analysis (research goal).
   b. Review literature to identify state-of-the-art and knowledge gap.
   c. Relate chosen methodology to knowledge gap, problem statement and research goal.
3. Methodology & Experimental Setup
   a. Summarize the methodology that will be applied towards the research goals.
   b. Contrast methodological alternatives.
   c. Illustrate the research methodology with a flowchart or similar visualization.
   d. Argue how chosen methodology will provide robust results and out-of-sample model evaluation.
4. Results
   a. Perform data science method(s) on chosen dataset in a robust, consistent, and transparent manner.
   b. Accurately illustrate the chosen exploration and evaluation method(s) numerically and/or graphically.
   c. Make a judgement on the performance ranking of competing models, based on robust evaluation metrics.
5. Discussion & Conclusion
   a. Interpret the significance, robustness, and novelty of the results in relation to the literature.
   b. Review the strengths and weaknesses of the method implementation and current results in relation to the research goal.
   c. Conclude on the contribution of the thesis to its research goal and societal relevance and plan follow up research.
6. Form & Presentation
   a. Perform formatting of a thesis conforming to TiU specifications, including rules for citations.
   b. Avoid plagiarism in the thesis.
   c. Put together a logically structured scientific manuscript.

Objectives

With the Master Thesis project “Data Science in Action”, students of the Data Science & Society (DSS) program demonstrate their mastery of the data science methodology (cf the Edison Data Science Framework[1]). The core of a DSS thesis is a machine learning approach (which may include deep neural networks) to data science, focused on exploring how existing or adapted features and algorithms contribute to regression or classification problems. Different algorithms are quantitatively contrasted in combination with multiple feature sets to arrive at the best predictive model, and model performance is validated on hold-out test set(s). Students interpret the models and explore the model errors to discuss the scientific and societal impact of their work.

For their projects, students use the R and/or Python programming language with accompanying libraries in an appropriate and correct manner. They are expected to approach the data scientific problems and questions pertaining to their project with curiosity, creativity, and as analytical thinkers. Students are required to translate complex and often extensive practical requirements (for instance, those of a commercial or governmental organization, or a research institution) into a work plan for developing, improving, or extending a data science solution. The proposed solution will support specific decision making and problem-solving processes and generalize to other similar contexts and new data.

Using an existing data set approved by their supervisor, students identify a substantive research question that can be addressed using the selected large data set(s). In order to formulate an appropriate research goal, students will produce a project definition that outlines the research goal and actively develop in-depth knowledge about existing solutions for the specific application area that will be discussed in the theoretical background for their thesis. Students are supported by experts in the domain provided by the data set owner (internal or external supervisor) and are advised to build on their prior expertise in a particular domain (e.g., their Bachelor studies) as much as possible.

The first stage in the project should be a well-crafted individual thesis proposal that provides the evaluating staff members with a clear view on the feasibility of the project. The thesis proposal is presented both in writing and orally during a presentation round organized by the evaluating staff members. If the thesis proposal and its presentation is successful (receives a “pass”), students continue with the thesis project. The end-product of the Master Thesis Data Science in Action (DSiA) project is the Master thesis. Next to that, students present the outcomes of their project to the general public during an open graduation session on a scientific poster accompanied by a short presentation.

[1] https://edison-project.eu/edison/edison-data-science-framework-edsf/

Thesis Content Requirements

Length

The length of the manuscript should be 8,000 words ±10%, excluding title page, preface, acknowledgements, references, and appendices. Students are required to list the number of words in the thesis data section on page 2 of the template. Theses that do not fall in this range will be automatically failed.

Elements of your thesis

The thesis consists of the following sections:

- Title page
- Preface (if needed)
- Abstract
- Data Source/Ethics/Code/Technology (DSECT) Statement
- Problem Statement & Research Goals
- Literature Review
- Methodology & Experimental Setup
- Results
- Discussion
- Conclusion
- References
- Acknowledgements
- Appendices and Supplementary Materials

Title page

Contains the title, author and other standard information. See the Latex template for details. The title summarises the substance of your thesis. Typically, it informs readers about what the research topic is and how it is being investigated; findings and other details are usually left out. Ideally, it should be less than 15 words.

Common problems and remedies:

- Be clear and avoid ambiguity
  - ✗ LCA of behavioural characteristics among TB patients
  - ✓ Latent Class Analysis of behavioural characteristics among tuberculosis patients
- Avoid being overly general or vague
  - ✗ Why do people evade taxes?

✓ Personality correlates of tax evasion behaviour among Dutch

- Be succinct; the finer details should be included in the Abstract (see section below)

  ✗ Support Vector Machines outperform other classification methods in forecasting stock market movement

  ✓ Forecasting Stock Market Movement with Support Vector Machines

Abstract

The summary is a very brief but self-contained account of your thesis. It should be around 150-250 words. The following points should be addressed:

- What problem definition of the thesis?
- What is your research question? The research question should follow from how other researchers addressed the problem (i.e., in terms of approach, focus, etc.) in the past.
- What distinguishes your approach from theirs? What are the essential features of your method?
- What dataset are you using?
- What are the main findings?

Data Source/Ethics/Code/Technology Statement

For purposes of transparency, oversight, reproducibility, and appropriate acknowledgement of the starting point of the thesis projects, students are required to write a DSECT statement. With the advent of AI-powered technologies that can assist in writing text, the assessment policy of the DSS Master Thesis / Data Science in Action has been updated to include the requirement for a **Statement of Technology** to be added to all deliverables. In the final thesis, this statement will be part of this broader **Data Source/Ethics/Code/Technology (DSECT) statement**. The templates will be updated to reflect the latest version of the Ethical Standards Compliance document for Students.

Obviously, the use of technologies like chatGPT runs the risk of committing (accidental) plagiarism. This concern is addressed in the compliance document. If you are unsure of a particular aspect, please discuss this during your supervision meetings.

If you can provide answers to all below questions for, this will provide the necessary information to include when writing your DSECT statement. There is no one-fits-all solution to these statements, and they heavily depend on the nature of the study. An example DSECT statement is included in the DSS master thesis templates (Overleaf and Word)

- Data Source
  - Who is the owner of the data?
  - Does the thesis project involve collecting data from human
  - participants or animals?
  - Does the owner of the data give consent to (re-)use of the data?

    - How and through what channel is the data acquired?
    - Figures
        - Did you create all the images and figures?
        - Do you have consent for using the images that you did not generate?
- Ethics
    - Did the study involve acquiring new data from human or animal participants? Under which ethical permission was the data obtained?
    - Is the data representative and inclusive or are there known biases in the dataset?
    - Under which de-identification regime has the data been released?
- Code
    - Did you use parts of the code from another study/someone else?
    - Did you list, including version number, all libraries and frameworks used?
- Technology
    - Did you use any tools or services to paraphrase the given text (for example a thesaurus or the Academic Phrasebank)? Please name them.
    - Did you use any tools or services to check spelling or grammar? Please name them.
    - Did you use any tools or service to typeset the given text? Please name them.
    - Did you use any reference management software other than the latex template? If so, please name it.
    - Did you use any tools or services to generate part of the text? If so, please name them.

Problem Statement & Research Goal

Explain briefly what the problem definition of your thesis is, what the societal and scientific relevance of your work is, what your research questions are, how you approached them, and what your findings were.

- Context
    - Start with the goal of your research and how this defines your problem statement.
    - Describe the context of your thesis in a very concise manner. Briefly explain the research domain, what the state-of-the-art is, and why the subject matter is interesting. Your readers should be left feeling that your thesis deals with an issue that is both important and interesting.
    - You do not need to go into great lengths to describe every relevant prior study yet; that should be reserved for the section on related work. For now, it is fine to state something along the lines of: *This issue has been addressed extensively (see Section 6).*
    - Devote one paragraph of this section explicitly to the scientific and societal relevance of your project. Note that scientific relevance could be derived from

the domain-specific research question addressed in your research, or in the proposal of a new algorithm or approach.

- Research strategy
    - Once the context is established, specify your research strategy, made up of one or more research questions (sometimes with subquestions). Research questions should be answered using techniques from data science and not via literature review or exploratory data analysis and should address issues of sampling, model comparison, and error analysis
- Findings
    - Give a brief (one paragraph) overview of your main findings.

Literature Review

Related work, sometimes labelled as theoretical framework or background, is a crucial element of your thesis. Explain the larger scientific context of the problem: what is the theory behind it, what previous research has been done related to it, and how your work builds on this related research. Below are some step-by-step guidelines for writing this section:

- Specify the area of research in which your work belongs and provide a context for the research focus. What research issue is your work focused on? Describe relevant work conducted in the same research area (with proper references). Has this issue been addressed in the literature? By whom? What have they done and found? What are the relevant theories? Are there any contradicting findings or competing models/theories? Provide the state of the art – give explicit metrics than can be compared across studies such as accuracy, R-squared, etc.
- Identify research gaps and/or shortcomings of existing methods; define research problems:
    - What is missing from prior research? What are the limitations of existing models?
    - Could there be alternative approaches to solving the same problem?
    - Specifically, what research problems are left unanswered? What insights or implications will tackling these research problems bring about?
- Use the literature to motivate the methodology you will use the research strategy. How will the chosen methods contribute to answering the research questions? What dataset will you use and why is it appropriate to use?

Common problems and remedies:

✗ Failure to maintain focus on the research question, by including references to studies that are only remotely related to yours.

✓ Start by searching for the most relevant recent papers on the problem, the dataset or the specific methodology. When the current problem is under-addressed, or the dataset is proprietary, the initial scope of the literature can be broader, including solutions to similar problems.

✗ Failure to support statements with adequate references.
✓ Always give credit where credit is due. If you are making a statement along the lines of: It has been established in prior research that…, make sure you follow the statement with references.

✗ Failure to express arguments or ideas in your own words.
✓ It is not acceptable to simply paraphrase the work of someone else by changing a few words here and there, without acknowledging the source. Instead, you should rephrase the idea in your own words. If you absolutely must include a direct quote, enclose it with quotation marks and specify the page number in your reference. Failure to do so is a case of plagiarism and can lead to severe consequences!

✗ Failure to include references to recent work.
✓ Whilst certain dated works remain important and are still widely cited (e.g., Gold, 1967, if the research concerns grammatical inference), try to stay on top of developments in the field and refer to the more recent literature.

✗ Failure to include references to relevant work
✓The literature review will be evaluated by the relevance of the cited literature. Did you include references for all major topics in your thesis (problem area, methodology, state of the art?)

✗ Failure to critically reflect on the literature.
✓ Demonstrate awareness of relations among existing models or studies by specifying any relevant commonality, contradiction, or inconsistency among them.

✗ Failure to give a convincing rationale for conducting the current study.
✓ Explain how the current work continues and improves upon previous lines of enquiry. Be explicit about the contribution of the current work.

Methodology & Experimental Setup

In the Methodology section you describe at a high level all elements of your data science pipeline. This should include the data mining and machine learning algorithms you used. Your methodology should include at least two models whose performance is contrasted or compared, and should be visualized with a flowchart.

You would usually explain the methods using a combination of mathematical formulas, diagrams, and descriptions. **Standard or common data science techniques do not need to be defined in depth** (e.g. do not include the formula for logistic regression, or for calculating accuracy). More importantly, make sure that you provide a justification for the methods used, compared to the alternatives. A motivation of the methods should be included or could also be placed in the concluding paragraph of your literature review.

Your Experimental Setup is where you describe in detail your dataset and the experimental procedure. Other researchers should have sufficient information to replicate your work based on this description (including flowchart visualisation) alone. The following information should be covered:

- Description of your raw dataset: the organization offering the dataset, sample size, how and when the data was collected, which features could be found in the data, and any other relevant information.
- Where appropriate, report (descriptive) statistics to offer a better impression of the dataset or selected results/visualisations of exploratory data analysis, especially how other variables relate to your target variable.
- Cleaning / preprocessing of your data; was there any oddity (e.g., error) in the dataset and what was done about it, which parts of the data were discarded and why, whether or not certain features were transformed and why, what was done about the missing values and why, and any other preprocessing done.
    - Always justify your decisions with theoretical and/or statistical arguments.
- Description of the experimental procedure: how was training and testing data selected, what was the order or flow throughout the data science pipeline, which algorithms were used at which stages, which parameter values were chosen and how.
- Description of evaluation criteria: for example, which error measure was used (e.g., classification accuracy, mean squared error, f1 score).
- Description of the robustness of the models, including (cross-)validation and out-of-sample generalization procedure to evaluate model performance on data not included in training.
- Description of the actual implementation, i.e., programming languages and versions, packages used, proprietary applications supporting the coding, etc.

It is vital that your data science procedures are robust and reproducible. Kapoor and Narayan (2022), when writing about data leakage as a threat to reproducibility, define data leakage as: “a spurious relationship between the independent variables and the target variable that arises as an artifact of the data collection, sampling, or pre-processing strategy. Since the

spurious relationship won't be present in the distribution about which scientific claims are made, leakage usually leads to inflated estimates of model performance". A model info sheet can help to prevent data leakage. It is available here: https://reproducible.cs.princeton.edu

Common problems and remedies:

✗ Symbols in formulas are not defined or explained.
✓ Make sure that it's clear what the notation stands for.

✗ Standard Data Science algorithms and/or metrics covered in detail
✓ You can omit detailed descriptions of approaches covered in the machine learning and data mining courses. You still need to motivate your choice of algorithms

✗ Failure to perform the correct data science methods.
✓ Make sure that the research steps you choose to perform are in accordance with the type of data and good data science practieces. For example, you cannot perform Principal Component Analysis (PCA) on categorical features.

✗ Failure to list all important details of research strategy.
✓ Always write with other researchers in mind and include all relevant details. When in doubt, ask yourself: If I were to leave this piece of information out, would other researchers be able to reproduce my work?

✗ Failure to justify choices made.
✓ Always be explicit about the rationale for making certain choices; they should be made on theoretical (e.g., prior research), methodological (e.g., algorithmic bias) or empirical grounds (i.e. tuning on validation data). For example, it is better to use only one or two algorithms properly than trying out every algorithm under the sun without proper justification and in a superficial way.

✗ Data leakage in the data science pipeline due to poor practices hampers reproducibility
✓ Make sure that your data science practices are robust and reproducible, e.g. by providing a model info sheet

Results

In this section, you report your results, often with the help of statistics, tables, and figures. Below are some guidelines:

- Present your results in a structured manner, often with the help of tables and figures. Pay attention to appropriate formatting of the graphs, and annotation.
- In your text, do not simply restate the information listed in the tables or figures. Try to make sense of the results, highlighting important or interesting findings that you might revisit in the discussion section. The figures are not the presentation of your results but their illustration.
- Do not leave information presented in tables or figures unexplained. You have included information there for a reason, so take the time to go through it (e.g., explain what each column is about).
- Provide high quality figures with clear axis labels, well-sized legends, and informative captions (providing a short description and explaining any symbols) and should be interpretable on their own. **Do not include screenshots of code or raw output.**
- Do not cluster tables and figures – there should always be some text in between.
- Larger tables (with more than ca. 10-15 rows) should be placed in the appendix. Figures that provide additional details but are not crucial to the main story can also be placed in the appendix
- An error pattern visualization is required (predicted vs. actual for regression, confusion matrices for classification). Where appropriate, explore the results further by means of statistical analysis, or visualizations.
- Present an error analysis, uncovering patterns that might not be obvious from the overall results (for example, does the overall pattern of results hold across ages and genders? Or, in the case of an AI model, does the predictive performance of the model differ greatly between classes).

Common problems and remedies:

✗ Failure to report the baseline performance.

✓ Always report the baseline performance, as it is difficult to interpret the results without knowing the basis for comparison (e.g., previous research, chance-level performance, majority class baseline etc.)

✗ Failure to interpret information listed in tables and figures.

✓ Instead of simply restating what is listed in the tables and figures, provide an interpretation of your findings so that your readers know what your results mean.

✗ Failure to use the correct type of figure.

✓ Consult scholarly articles and books to see which type of figure is appropriate for which visualization purpose.

✗ Failure to format numbers according to English-language conventions.
✓ Make sure you use decimal points, and commas as thousand separators (i.e., 1.2 and 10,000)

Example structure for reporting results:

- First, describe in one or two sentence(s) what results will be reported in this section. For example: In this section, classification performance for the feature types described in section 3 on X dataset will be presented.
- Describe the classification performance, which should be listed in a table too. Let your readers know which table you are referring to. Do not simply repeat the information found in the table; tell your readers: Which approach yields the best performance? Which is the worst? Is performance consistent across folds/resamples? Is there any other noteworthy result? Use statistical methods to compare distributions of metrics between algorithms to judge whether a particular method is significantly outperforming another. For example: The results of the classification tasks on the dataset are shown in Table 1. Both approach A and approach B yield the best classification performance.
- Explore your results further. For example: Does the classification performance differ across classes? What contributes to poor classification? It might be worthwhile to conduct additional classification tasks, for example on a subset of the dataset. Present the additional results in a table too and try to make sense of these additional findings. For example: "We analyzed the classifications of each Y [= what is being classified] individually to gain a better understanding of why approach C did not perform as expected. It appears that the poor classification is due to the following reasons: [list the plausible reasons here]". A discussion of model bias against a particular level of Y would be very appropriate here (and possible mitigation strategies)

Discussion

In this section, you should evaluate your results regarding the research questions listed in the introduction. The following elements should appear in your discussion:

- Summary and discussion of the results
    - Remind you readers what the goal of your study was and summarize your obtained results in one or two paragraphs.
    - Discuss the overall picture that emerges from the findings.
- Comparison to the literature:
    - Put your results in perspective by making links to the literature. Do not simply repeat all the findings that you have already reported in the Results section.

Show how they relate (quantitatively!) to your literature review and research questions.
    - For example: “Approach C performs considerably worse, a result that contradicts prior research [reference(s)] and our expectations.”
    - If a finding was surprising, you should offer reasonable speculations as to why this particular result was observed.
- Discussion of scientific and societal impact
    - Make very clear what the contribution of your study is within the existing framework. The discussion should show how your findings fit with existing knowledge, what new insights they contribute, and what consequences they have for theory or practice.
- Limitations and future directions:
    - It could be the case that your results only partially answered your research questions due to limitations of the model or the data. Acknowledge these limitations, offer possible solutions, and defend the validity of your results.
    - Include recommendations for future work; provide suggestions on the practical actions or scientific studies that should follow.

Common problems and remedies:

✗ Failure to provide context for the results.

✓ The discussion section should be understandable when standing alone. It is important that you spell out your research questions and goals again, so that researchers who only read this section can still have a good idea of what your findings are.

✗ Reporting new results in the Discussion section.

✓ All analysis and results should be reported in the Methods and Results sections. Do not introduce new results; you should only discuss the data that you have already reported in the results chapter.

Conclusion

Conclusion is a short section where you restate the problem statement and how your research contributed to problem at hand. You can include your research questions, though these might be too specific for a general conclusion. Finish by concluding on the implications of your work for the field.

Acknowledgements

In this brief section you can acknowledge sources of funding, data, code, or anyone who helped you with your research.

References

It is mandatory to use an approved citation style such as APA7 or IEEE.

Appendices and Supplementary Materials

Appendices are appropriate for extra, non-essential visualizations, examples and analyses. It is strongly encouraged, whenever possible, to store data and source code in an online repository and refer to it in the manuscript. Github.com is a common choice.

Copyright and collaboration

Do not reuse any figures or other images without a proper license or permission of the author; if you have the permission, the author still needs to be credited. See also the Data/Code/Ethics statement section
If you collaborated with someone else on some part of the project, indicate clearly which part of the work you are building upon was done by something else.

Plagiarism/Overlap

Make sure your thesis does not contain unacknowledged quotes or paraphrases as these could be detected as cases of plagiarism. Please consult the plagiarism FAQ in the Canvas course and the policy regarding overlap at the School level:
https://www.tilburguniversity.edu/students/studying/regulations/eer/humanities
Textual/conceptual overlap indicative of plagiarism is grounds for failing a thesis submission.

Checklist

Make sure to check the following items before you send your manuscript for review to your supervisor and/or second reader.

- Is the length within the specified limits?
- Did you run a spell checker?
- Did you proofread for grammar and clarity?
- Did you use English-language conventions for number formatting (decimal point, comma thousand separator)?
- Did you round excessively precise numbers?
- Are all quotes and paraphrases from other texts properly referenced?
- Are all figures either your own or used by permission?
- Are the symbols used in formulas defined or otherwise explained?

Useful sources

Below are some useful resources for writing your thesis.

APA style

- APA publication manual (7th ed.). Available from the library.
    - ISBN 13: 978-1433832161
    - ISBN 10: 143383216X
- Purdue online writing lab: https://owl.english.purdue.edu/owl/section/2/10/

General tips on writing and mechanics of style (e.g., punctuation, grammar, spelling, sentence structure, etc.).

- *How to write a paper* (7th edition), by Ashby, M. F. (2011). Available from: http://www.grantadesign.com/download/pdf/How_to_write_a_paper_6th_edi___tion_2005.pdf

References:

Kapoor, Sayash, and Arvind Narayanan. 2022. “Leakage and the Reproducibility Crisis in ML-Based Science,” July. http://arxiv.org/abs/2207.07048.