Thesis Proposal Data Science & Society

Version S2023 – January 2023

0. Objectives of the Master thesis

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 (including 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 error patterns 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 strategy, 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 session organized by the evaluating staff members. If the thesis proposal and its presentation are successful (receive 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.

1. General

The thesis project proposal and proposal presentation jointly form a summative assessment in the Master Thesis course and determine the outcome of a Go/NoGo decision for the project in question. The goal of the thesis proposal is to provide a roadmap to the final thesis submission

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

that can be evaluated on its proposed data science methodology, its scientific and societal relevance, its novelty, its feasibility, and its planned experimental rigor.

The project proposal consists of a well written document (1250 ± 20% words, excluding the title page, references, tables, and figures). The proposals should be written in correct Academic English and adhere to the APA7 or IEEE Style. Proposals with spelling, grammar or style mistakes will not be evaluated; instead, you will be asked to resubmit a corrected version. An Overleaf template for the Thesis Proposal (recommended) is available. A separate template is available for the eventual full thesis. If you do not use the Overleaf template, make sure that the title page displays provide your full name, email address, thesis cohort, the name of your internal supervisor and the contact information of your external supervisor, if applicable. Submit your proposal as a pdf file in the Canvas Assignment. Your proposal will be evaluated by your supervisor of the project and, if necessary, a second reader. You will receive the evaluation of the proposal and, for the draft version, feedback on its contents. It is expected that you make use of the feedback on the draft proposal for the content of your Thesis Proposal Presentation. Thesis proposals that receive a NoGo (fail) can be submitted as a resit for this assignment, but this will not change the deadlines for the final thesis product. Note that your project plan may need to be adapted as you learn more about the data. This is fine as long as your overall goal (the task you are addressing, the data set you are using, the methods you are using) remains generally the same. Should your project change in a major way from what you initially proposed, you need to get renewed approval from your supervisor.

2. Avoiding Plagiarism

As with all assignments, you have to make sure that you do not commit plagiarism. Plagiarism is considered a serious case of fraud that, when suspected, will be reported to the Examination Board. Committing fraud can have serious consequences. At the minimum, when fraud is established by the Examination Board, the assignment is declared invalid and, in the case of a thesis, a new thesis will have to be written. Please see Article 16 of the Rules and Guidelines for TSHD (see below) for the procedure and sections in case of fraud. Note that TiU defines plagiarism as: "Using parts of a text written by someone else, or the reasoning or ideas of others, for a thesis or other assignment, without due acknowledgement." (Source: https://www.tilburguniversity.edu/students/studying/regulations/fraud/whatisplagiarism – this text contains a more elaborate explanation of what is plagiarism).

Note: Specific guidance on the use of chatGPT and other AI solutions will be presented shortly.

2.1. Overlap Detection – Feedback

To prevent accidental plagiarism, we want to ensure that students can gain experience with the overlap detection mechanism implemented by TurnItIn. TurnItIn is one of the tools that supervisors use to assist them in detecting potential cases of academic fraud. Please note that establishing fraud is a decision that is always made by the Examination Board of the School, not just by the TurnItIn algorithm. To provide you with this experience, the Canvas assignment for the *Draft* Thesis Proposals will give complete TurnItIn feedback to students as well as supervisors, so that both can learn from the feedback provided by this system. In addition, this Canvas assignment is open for multiple resubmissions, so that students get the opportunity to repair a draft proposal with regard to overlap before submitting their full proposal.

Useful Resources:
- TSHD Education and Examination Regulations (EER), including the Rules and Regulations: https://www.tilburguniversity.edu/students/studying/regulations/eer/humanities
- What is plagiarism?
https://www.tilburguniversity.edu/students/studying/regulations/fraud/whatisplagiarism

3. Outline and Contents

In the following subsections, a description of the general contents of the sections that will be assessed with the Assessment Rubric (see chapter 6) is laid out. This is a general description and should not be thought of as a complete "recipe" for the proposal. Your proposal can differ in order or grouping from this outline, as needed.

3.1. Project Definition, Motivation & Relevance

Provide a clear description of context of your thesis project, including a problem statement. Briefly explain why this problem is worth addressing, both from a societal and scientific point of view. Make sure that the problem you address has not been solved already.

3.2. Literature review (initial)

Provide a summary of what is known in the scientific literature about this problem. This should be based on at least 5 relevant recent sources and, if appropriate, some more classical sources. These recent sources need to satisfy the following requirements: (1) recency (published in the last 5 years), (2) quality (published in scientific peer-reviewed journals or conference proceedings), and (3) usefulness (they should help you frame the theoretical background of your project). At this stage of your project, a full literature review is not expected but it will be expected by the time of the final report.

3.3. Research Strategy and Research Questions

Outline the concrete Research Questions that the thesis project will answer. What will be contributed to the literature by answering these research questions. Avoid very general statements ("is it feasible to…") but try to formulate concrete research questions, split into subquestions where appropriate. The RQs should follow logically from the problem statement combined with the state of the art to inform your research strategy.

(Sub-)RQs should clearly specify which manipulations to the data, features and/or algorithms are contrasted and should be specific rather than general statements (e.g. name the algorithms you are considering, instead of posing general questions like "which Machine Learning model…")

Your strategy should address these elements that also appear in the evaluation rubric for the final thesis product (if already known at this point):

- Does the dataset/your target variable contain large class imbalances/non-normal distributions?
- Are there separable feature sets in the dataset(s)? Will additional features be generated? Which methods will be used for featureselection/ranking or model interpretability?

- On which aspect will model comparison be implemented? Com-parison of several tuned algorithms, comparison of different input datasets? What is the proposed method for validation and test setseparation? Will you use resampling or other statistical methods to assess model comparisons?
- How will error patterns be analyzed? Are there interesting sub-groups in the data for which bias/subgroup error analysis could be implemented?

(Sub-)RQs should specify which manipulations to the data, features, and/or algorithms are contrasted and should be specific rather than general statements (e.g., name the algorithms you are considering, instead of posing general questions like “which Machine Learning model...”).

You can write a short motivation leading up to a subRQ, for example: “previous research has shown that a larger proportion of man failed in X compared to women. Therefore, model performance and error analysis will also be split according to gender”.

3.4. Methodology & Evaluation

3.4.1. Dataset Description

Describe the dataset(s) that you will use in your project (size, format, accessibility). Provide rationale as to why you are choosing these data. **If, at the point of proposal submission you do not yet have your complete dataset (e.g., in a project with an external partner) there is a very real risk the project might fall through, and students are advised to look for a backup solution.**

3.4.2. Algorithms and Software

Describe what algorithms and software you plan to use in your project. Include a motivation for why you have chosen these implementations, with references to the literature.

3.4.3. Evaluation Method

Define how you will evaluate your results. For prediction problems (classification or regression), you will likely make use of standard techniques – they do not need to be explained in detail. How will you be able to judge on performance between competing models? Against what baseline method(s) will you compare your algorithm(s)? How do you plan to obtain ground-truth labeled data so that you can measure accuracy, precision, recall or some other metric? If you are planning to use unsupervised techniques, provide information of how the clustering algorithm will be tested and how model comparison will be implemented. Details about a plan for cross-validation or other out-of-sample evaluation should be included.

4. Milestones and Plan

Sketch out what you think will be the major intermediate milestones that you will need to achieve. Give a general idea of your planning.

5. Assessment Details:

Relationship to program learning outcomes

The Thesis Proposal assignment relates all the learning goals for the thesis, but specifically to the ILO KU1-2 and MJ2 from the DSS Program:
Knowledge and understanding (KU):

1. Students of the program: “Have broad knowledge and understanding of data science theories, methods, and techniques concerning data from socially relevant domains”.
2. Students of the program: “Are able to formulate novel ways of producing and processing information with the help of data analytics using existing knowledge in socially relevant domains”.

Making Judgment (MJ):

3. Students of the program: “judge the appropriateness of use for statistical and coding techniques employed in data analysis for a specific domain.

Learning goals (for the thesis proposal)

After finishing this assignment, the student can:

1) Motivation/Relevance
   a) Illustrate the societal relevance of the thesis research goal
2) Literature
   a) Summarise existing literature of methods and results applied to a particular data science problem or analysis (research goal).
3) RQ
   a) Formulate (a) clear and specific research question(s) based on identified gaps in literature that lead to solving the research goal.
   b) Organise research question(s) in a logical and feasible research strategy, with the help of sub-questions.
4) Method
   a) Argue why chosen data science method(s) is/are most appropriate, in contrast to other methods, to approach RQs
   b) Illustrate how model comparison and out-of-sample generalization will be implemented in the data science approach of the thesis
5) Form and Presentation
   a) Explain the research goal using their own words without relying on quotations
   b) Relate the literature, the research question(s), and methods that make up the thesis proposal in a coherent structure

6. Assessment Rubric for the thesis proposal (pass/fail)

All items below must receive a Pass mark for the presentation to receive a Pass mark. In case of a fail, a resit opportunity will be scheduled by the supervisor.

Item	Sufficient (Pass)	Insufficient (Fail)
Motivation/ Relevance	Proposal clearly presents societal and scientific relevance of the project	Proposal does not clearly present scientific or societal relevance
Literature	Proposal provides succinct summary of at least 5 relevant sources from scientific literature. Proposal highlights gaps in literature that lead to open research questions	Proposal omits or does not clearly illustrate relevance of cited literature. Cited literature does not inform research questions
Research Questions	Research Questions follow logically from project definition and literature review Research questions are appropriately specific, empirical, and feasible	Research Questions do not contribute to the project definition or disregard existing literature Research Questions are unspecific or unattainable or cannot be answered empirically with data science methodology
Method & Evaluation	Proposal has motivated choice of data science methods in light of the dataset and literature review Proposal has clearly outlined a scientifically rigorous evaluation strategy to answer the RQs, including model comparison, out-of-sample evaluation, and analysis of errors (if appropriate) and/or disparate impact/bias across classes/groups	Methods chosen are not appropriate for a data science thesis or redundant given the literature and dataset Evaluation strategy will not unambiguously answer research questions or does not include a model comparison approach Research Strategy does not include analysis of either errors or disparate impact/bias
Form, Structure & Presentation	Proposal conforms to guidelines regarding formatting, style, sections, and length, including citations and bibliography Proposal is written in a cohesive and structurally sound manner. Information content is placed in the appropriate places	Proposal does not conform to the necessary formatting of a TSHD Master thesis document, including approved citation styles. Structure of the proposal is incohesive or information content is placed in inappropriate places
Originality of writing	Proposal makes proper uses of citation, quotation and paraphrasing to avoid plagiarism	Thesis contains improperly paraphrased material, improper/incomplete citations, or improper attribution of direct quotations