Introduction to Data Science Project Information 2023

Formal requirements

- Teams of 2-3 students, at least 30 hours of work per student
  - Team can have 1 student if working at least 60 hours and if this is agreed beforehand with the course instructors
- Each team introduces the chosen topic in a practice session on Nov 13, 14, 15
- By **deadline on Nov 13 at noon (12:00)** - Each team must insert a slide with the project's title, description and team members into the List of Projects document corresponding to in which group's practice session you want to do the introduction:
  - Group 1: https://docs.google.com/presentation/d/1qfB5unNfexMeYYVr3UK4eKTXomG1CbkzQIezgqy7T3g/edit?usp=drive_link
  - Group 2: https://docs.google.com/presentation/d/1UrLrnm6E-Jt4xpj7315pk0M6jYmXv5shiQN8grZX8es/edit?usp=drive_link
  - Group 3: https://docs.google.com/presentation/d/1DqxKzghGC24eywNAszOYDjqnD-NkEzKpyRJsm1Ay2E8/edit?usp=drive_link
  - Group 4: https://docs.google.com/presentation/d/1msjMBJY4RMK4Cb8tJfoMvDRidByV_UacP2gYB_YPQJg/edit?usp=drive_link
  - Group 5: https://docs.google.com/presentation/d/13g-8dqAj4Mw-tDK6grchqbrs7pmtQNT8Eu0ox7iJQ_o/edit?usp=drive_link
  - Group 6: https://docs.google.com/presentation/d/19hWzZBjddRivqEEHtFB2fhLVayczFOz0Uzw5rIFBZgM/edit?usp=drive_link
  - Group 7: https://docs.google.com/presentation/d/1Jpkud9Yy01r98b4MbbG5C0erW8xf8gTJDIfnNqzu67s/edit?usp=drive_link
  - Group 8: https://docs.google.com/presentation/d/17M836SMDsTDQOs9prXY3GtOHE4j5b3gCcK2wQm0Rww0/edit?usp=drive_link
- If a team has members from multiple groups, then you can choose which of these groups you present at. If some members cannot participate at the presentation, then that is ok. Keep in mind that practice session attendance is still checked during that week, but it is ok to participate in another group's practice session instead of your own.

Further requirements

- Every team must present the project in the **poster session** (Dec 14 at 2pm)
- Every team must provide access for the grading instructor to the project code repository hosted either at Github or BitBucket
- The project will be graded by the instructors and will get maximally 30 points
- If the project gets X points then each team member gets X points
- Getting at least 15 points for the project is a prerequisite for passing the course

Evaluation of projects

- Projects will be evaluated after the project session
- The grade consists in the following two factors
- Technical quality (15 points)
  - Your project can get 15 points for technical quality, if you have:
    - stated clear objectives
    - applied relevant data mining methods on relevant data, and
    - the achievements are sufficient, considering the required working hours
- Presentational quality (15 points)
  - Your project can get 15 points for presentational quality, if the poster is:
    - explaining well the motivation and objectives of the project
    - describing used data science methods so that others could in principle replicate your work
    - presenting the main results of your work in a visually appealing and understandable way
  - and if your project code repository (GitHub or BitBucket) contains:
    - all the code as well as readme-files describing what the code does and how to run it

Instructions

- There are following options for choosing the topic of your project:
  - Choose a topic on the data offered by us
  - Choose a topic from Kaggle
  - Choose your own topic (on some open data or on any data you might have)
- For each of those options the formal requirements differ slightly, so please look at the dedicated slide below
- There can be multiple projects on the same topic

Choosing a topic from Kaggle

- Kaggle (https://www.kaggle.com/) hosts many interesting machine learning competitions and datasets
- You can choose to compete in one of the challenges (https://www.kaggle.com/competitions) or work on one of the datasets (https://www.kaggle.com/datasets)
- In each of the Kaggle challenges and competitions there are available ‘kernels’ where people have explained a way to analyze the data and have provided their code
- You must declare all kernels that you use in your project

Choosing your own topic

- You can freely choose your own topic, as long as it requires you to demonstrate mastering some topics from the Introduction to Data Science course
- In your readme-files of your code repository page you must specify the origin of the data and provide a short description of the data
- The data do not need to be public but the grading instructor must be able to see the data, at least from your computer
- Here are some more ideas to help you:
  - http://opendata.riik.ee
  - https://www.wikidata.org/
  - http://openclimatedata.net/
  - https://data.worldbank.org/
  - https://www.data.gov/
  - https://data.unicef.org/resources/resource-type/datasets/
- Finally, have a look at the project presentations from 2021 (different format than we will have):
  - https://courses.cs.ut.ee/2023/ids/Main/Projects2021

Topics proposed to the IDS course:

Lung Cancer Survival Prediction with Synthetic Data [Markus Haug, University of Tartu]

Objective:
Utilize synthetic medical data to predict 5-year survival rates post-lung cancer diagnosis by effectively selecting key features from patient treatment trajectories.

Possible approach:
1. Feature Selection: Use algorithms like PCA and Stepwise Selection to identify crucial features.
2. Subsequence Analysis: Find frequent subsequences within treatment trajectories for additional predictive value. Effectiveness (performance) of the algorithm is important.
3. Use network science approach. Using graphs and extracting key-player nodes as features and using them for prediction.
4. Any custom approach.

End result:
Predictive Model: Develop a workflow that, given patient treatment trajectory data, outputs a 5-year survival prediction. Workflow will be tested on a test set not seen by students and compared with other groups work.

Outcome:
A concise list of significant features, an understanding of vital treatment subsequences, and a prediction function for stakeholder use. Possibility to get published!

Impact of covid 19 on energy consumption in Baltic countries [Neha Sharma, University of Tartu]

- We would like to find the impact of covid 19 on energy consumption for Baltic countries. The idea is to find how the covid 19 measurements such as lock down, shut down of schools, universities, commercial places etc has impacted the energy consumption. There is data available for all 3 countries from 2018 to 2022. Methods such as data analysis, visualisation, statistical models and time series analysis could be used to get results.

Automated classification of open datasets to improve data findability on open government data portals [Anastasija Nikiforova, University of Tartu]

- Problem:
  Many open government data (OGD) portals offer vast amounts of free datasets, but a significant portion remains unused. A primary reason is that these datasets are challenging to locate due to poor categorization or inaccurate tags, with some OGD portals having up to a third of their datasets uncategorized. This makes datasets hard to discover unless users search using precise terms from titles or descriptions.

- Goal:
  Design an automated data classification mechanism to suggest relevant categories and tags for datasets based on their title, description, and possibly other parameters. This would improve the findability of datasets on OGD portals.

- Methods:
  Explore existing OGD portals to understand dataset presentation and user search behaviors.
  Define a list of indicators that would serve as input for data classification.
  Develop a solution, likely involving simplified text analytics, to automatically assign appropriate categories and tags to datasets.

- ORIGINAL DESCRIPTION:
  While many open government data (OGD) portals provide a large number of open datasets that are free to use and transform into value, not all of these data are actually used. In some cases, this is because these data are difficult to find due to the low level of detail presented in them, including, but not limited to the absence or inaccuracy of the category(-ies) and tags assigned to a particular dataset, which is a part of the data publisher task. In the case of some OGD portals, 1/3 of the datasets are not categorized, although the portal provides a rich list of data categories that are in line with best practices and allow to classify these datasets. This leads to cases where the dataset cannot be found if the user searches for data using catalog or tags (only using the search bar will return the dataset, if the search query matches the title or description of the provided dataset). This thesis is intended to propose an automated data classification mechanism, which, based on a dataset and the data provided on it (title, description of the dataset (! please, take into account that you will be asked to carry out at least a simplified text analytics), parameters of the dataset (if sufficiently expressive)), will suggest a categories and tags to be assigned to it.
  First, you will be expected to explore OGD portals and how datasets look like, and what can be scenarios for OGD user to search for a particular dataset. Then, a list of indicators will be defined, which should constitute the input for data classification (mostly in line with the above but can be enriched, if possible), and an appropriate solution will be developed. This would contribute to the FAIRness of the open data, although mainly referring to F – findability, but indirectly affecting other features that the OGD should meet in order to provide social, economic and technological benefits from individual users, SMEs and governments.

Towards automating data quality specification [Anastasija Nikiforova, University of Tartu]

- Problem:
  Data quality management is crucial due to the increasing volume and variety of data, especially third-party open data. Given the complexities involved in understanding data quality dimensions, rules, and metrics, there's a need for automated data quality specification.

- Goal:
  Develop an ad-hoc approach to automate data quality (DQ) specification. This involves extracting DQ requirements from data features, facilitating the creation of rules based on patterns present within the data, such as recognizing email formats, date consistencies, or postcode structures.

- Methods:
  Leverage Machine Learning (ML) techniques for this extraction, possibly in combination with predefined rules or metadata-driven methods. Existing ML-supported DQ tools can serve as a reference.

- ORIGINAL DESCRIPTION:
  Data quality management, although is not new, but still very relevant topic that becomes even more relevant with the increase of the amount and variety of the data. However, data quality is a very multi-faceted topic, where a proper management of the above is a complicated task, including but not limited due to the need for domain knowledge with the reference to both the data and topic of data quality. For the latter, this is due to the concepts of data quality dimensions, rules and metrics, which makes it complicated to the end-user without respective DQ knowledge conduct a DQ analysis. Hence, attempts towards automating data quality specification become popular. This is all the more needed in the light of wide popularity of third-party data such as open data - data that were generated / collected and processed by entity other than data user. Current approaches can be divided into rule-based, metadata-driven and ML-based, where the latter is the most promising but the least represented in both academia and practice. Hence, the objective of this project would be to propose such a ad-hoc approach towards automating DQ specification by means of extracting data quality requirements from data features, which would be expected to be done employing ML (composite/combined approach with the use of predefined rules or using metadata is welcome and can appear to be the most promising). This would mean that DQ rules would be extracted from the analysing the data (attributes / columns for tabular data) and determining the patterns such as email (if @domain is determined), date, post codes or any others, as well as consistency rule extraction (e.g., attribute names or metadata states that the dataset contains start and end date, so the comparison of the respective values, so that end date is after the start date would make sense). The requirements for the above can be partly retrieved from several existing DQ analysis tools supporting ML-based rule/check definition.

Tool for analyzing financial business reports [Kristiina Lillo, SEB Bank]

Very challenging!

- Our project is to have an independent tool that will read in business annual report information (tables but as well unstructured description fields) and will provide output of pdf comparing historical information about business financial metrics but as well description fields about business type, changes in management boards, description, most important business events and so one.
- Problem is that currently there isn't any good automated tool that will give us possibility to easily compare how business has changed over the years, all of that need to be done manually.
- User input should be business name and number of years to compare (default value should be 5 years).
- Input for script should be annual business reports, which are available in Estonian business registry (e-Äriregister – (rik.ee)). Script should be able to detect agreed data fields (for example revenue, sales, debts) and their changes over the years. As well if in description fields (text fields) information is changed about important business related inputs (for example management board members, type of business, partners and so one).
- Output is pdf file – where overview of different parameters is added, how they are changed over the times, what important information were able to detect form unstructured data fields.
- Additionally (not mandatory) – if possible then also to have possibility to compare different companies to each other. For example based on their location, type of business or industry.

Topics using large language models
[Meelis Kull, University of Tartu]

- We can provide access to the API of OpenAI large language models such as GPT3.5, GPT4, as well as the Embeddings API (costs covered by the university, quota-based)

SOME IDEAS:

- LLM-based search engine for courses at the University of Tartu
  “I am studying computer science but want to take some basic physics course”
  “As a master's student in the quantitative economics curriculum, I'm looking for advanced seminars on market strategy. What are my options?"

- LLM-based researcher search engine at the University of Tartu
  “Are there researchers in environmental studies who have published on climate change impacts?”
  “Find me all researchers in the University of Tartu who have applied machine learning in their papers”

Questions or comments?

Please ask in Campuswire under the tag ‘project’.