Instructions and hints for datasets

General

Each of tasks below should be processed by CRISP-DM methodology, which will be presented on lectures. We expect an analytic report that should contain at least:

- description of the solved problem – to be obvious that you understand the task
- input data description
- planned process steps, methods, techniques (a short description)
- data quality check and resolution about processing of data errors or missing data
- data exploration with proper charts/diagrams – to be obvious that you have understood data and relations
- (possibly simple) predictive or classification model (if not stated differently) for a proper target
- model performance evaluation
- summary, conclusion

The expected volume of a report is 10 to 15 pages (recalculated to the A4 format) of text, tables and charts. You can keep a source code in your report but it is not taken into account for the report volume, the same holds for big tables.

The report may be made in Czech, or English. We understand Slovak, too.

Changes of instructions are possible. Student may suggest any change if there is a reason for it; teacher will deal with this demand as soon as possible.

Assessment criteria

- Structure and arrangement (max. 9 points)
  - contains all required sections
  - does not contain useless things (source code included only reasonably)
- Data description and exploratory analysis (max. 11 points)
  - data sources are commented, structure and semantics of source data are described
  - source data is properly limited with respect to the solved task(s)
  - data quality check and sanity check are done on data
  - basic exploratory analysis is done, conclusions about data and about required transformation/cleaning are summarized
- Modeling (max. 11 points)
  - data is cleaned and properly transformed
  - target and features are well chosen
  - modeling methods and metrics are well chosen
  - a reasonable baseline model is fitted and evaluated; if needed, advanced models are fitted, evaluated and compared to the baseline model
  - outcomes of modeling are commented and interpreted
- Understandability (max. 11 points)
  - text can be understood without great effort, big grammar and stylistic mistakes are avoided
  - the reader can easily follow the autor, there no gaps and jumps in the text flow
  - each step is described or explained, tables and figures are properly commented

    - graph types are well chosen, figure design is consistent
    - the summary covers all report and gives answers to all asked questions
- Tidyness (max. 8 points)
    - the report is pleasant and engaging, reading is not tiring, text and figures are tidy and in good size
    - the report has proper layout and design, nothing is disturbing or distracting

Datasets

Datasets are saved in GitLab repository: https://gitlab.mff.cuni.cz/mlyni8am/ndbi048. Datasets have different size and complexity. At large or complex datasets the student may focus on a particular problem and work only with subset of columns and rows.

There are some datasets for demonstration purpose, **not for assignment**:

- hazard.zip
- hazard-sessions.zip
- homecredit_default.zip
- lyrics.zip
- stopwords.zip
- titanic2.zip

Almost all datasets have an own description text file included. If you still don't understand anything in the dataset, **do not hesitate** to ask (or to find an answer on the internet).

Each dataset below has a brief descriptions and examples of problems for an analysis. You may take them as an inspiration (do **not** take it literally as an assignment) or come with your own idea.

King James Bible (bible-kjv.zip)

The Bible in an ancient English translation (17th century). Plain text where each row is one verse with structure book name, chapter:verse number, verse text.

Examples of problems:

- If we take a random verse from the Bible, can we classify it (by its words) to the Old or New Testament (or to some specific set of books)?
- What book is the most similar to the book X? Can books of the Bible be clustered to "families" by the similarity of their vocabulary, text phrases etc.?

Birthdays (birthdays.zip)

Number of born children by US states and single days from January 1st, 1968 to December 31st, 1988. All 50 states of the USA and the District of Columbia have the standard two-letter abbreviations.

Examples of problems:

- How many children will be born next year in the state X?
- Are trends in number of newborns similar in all states or are there differences? How could we explain these differences (west X east, rich X poor, flat X hilly states etc.)?
- Does distribution of newborns over a year change in time?

Accidents on Brazil highways (brazil_nehody.zip)

Records of traffic accidents on federal highways in Brazil, 2007–2023: date, time, GPS, reason, weather conditions, number of dead etc.

Examples of problems:

- What is the probability of an accident in a particular highway section and on a particular date? What factors are important for this probability?
- How many dead will be next week (month, year) on the highway X (or on all highways together)?
- Did radar installation have a significant impact? Was there any reason for radar installation?

Chess ratings (chess_ratings_upravene.zip)

Official monthly lists of chess players in standard (Dec 2017 to Dec 2022), rapid and blitz games (Jan 2022 to Dec 2022).

Examples of problems:

- What rating will a particular player have next month or a year after (in absolute measure or binary "will be higher or not")?
- How many games will a particular player play next month (year)?
- What will be the average age of top 10 (100) players of particular country (federation) next year?
- Does age and sex have an impact on differences between standard and rapid (blitz) rating?
- If we know the age, sex and rating of the player, what title does he/she have?

Earthquakes (earthquakes_kaggle.zip)

The dataset is an extensive collection of data containing information about all the earthquakes recorded worldwide from 1990 to 2023. The dataset comprises approximately three million rows, with each row representing a specific earthquake event.

Examples of problems:

- What is the probability of the earthquake over certain magnitude in a particular area in next year?
- Are there any trends in magnitudes, frequency or moving epicentres over years?
- What factors are important for a tsunami following the earthquake?

Elections results in the Czech Republic (elections.zip)

Results of parliamentary elections in 2013 and 2017 years by election districts (the smallest unit for which votes are summed) together with some sociodemographic descriptors of the district population (e. g. share of men/women, share of unemployed people, share of children). If a student prefers, results of 2021 or 2025 elections may be studied instead but we cannot provide them.

Examples of problems:

- What vote share will have party X in district/municipality Y in 2017?
- What turnout will be in the district/municipality Y in year 2017 (in absolute numbers or binary "over/below total average")?

Jokes (jokes.zip)

150 jokes in English (text included) and their ratings.

Examples of problems:

- If we take a random joke and forget its evaluation, what average evaluation and its variability do we expect, taking all other jokes into account?

- How well can we predict whether reader X would like joke Y?
- How can we cluster jokes to reasonable "families"? What words or other factors are key for such clustering?

Population in Czech municipalities (populace_obce_cr.zip)

Population and its changes in all Czech municipalities yearly from 1971 to 2020. Additionally, a csv file with GPS coordinates for municipalities is added (note: the encoding of this csv differs from excel files).

Examples of problems:

- How many citizens will live in the municipality X next year (or N years ahead; in absolute numbers or binary "will/won't be higher")?
- How many newborns and deaths will be in the municipality next year?
- Are there any differences in trends in small and big municipalities or in different regions? What factors are important for such trends?

Population changes in all world's countries (populace_svet.zip)

Population and other demographic data on every country (so called "economic") in the world from 2000 to 2021. The zip file contains individual csv files for various demographic indicators (rows=countries, columns=years), an overview table and country groupings with respect to different categories. Some indicators are obvious, some need to be discovered or understood and compared to external sources.

Examples of problems:

- How many people will live next year in the country X (in absolute numbers or binary "will/won't be higher")?
- Will median age or life expectancy grow or sink in the country X next year?
- What is the similar feature of countries with growing population? Is the population trend correlated with trends in countries in neighborhood?

Tennis matches (tennis_matches.csv)

Statistics of sampled tennis matches (both men and women) in the period 2015–2023.

Examples of problems:

- Are all players universal, or is the surface important for some of them?
- How many aces or doublefaults (possibly on 10 games or so) will have player A in next 10 matches?
- How well can we predict the outcome of the match by ranking of oppponents? Do results of previous matches of them have a significant effect on the outcome?
- How long will a particular match last? How many games will be played?

Vaccination and corticosteroids (datacon.zip)

Records of clients of two health insurance companies including vaccination status and corticosteroid consumption.

Examples of problems (see also the commentary included):

- Did corticosteroids consumption vary in time, especially during covid period? If so, how it can be modeled (explained) and predicted?
- Does vaccinated and unvaccinated population differ in corticosteroids consumption? Are there differences between various age groups?

Vaccination and deaths (uzis_ockovani.zip)

Vaccination status and death date (if it happened in 2020–2022 years) of all citizens of the Czech Republic.

Examples of problems: see the articles included. You may try to replicate an analysis and then extend it some way or follow another goal (e.g. prediction of dead number next day).

Football world cup and matches (world_cup.zip)

This collection of file contains information of FIFA World Cup (WC) 2022 and of previous world cups together with sampled history of international matches.

Examples of problems:

- What will be the outcome of the match between teams A and B?
- How many goals will score the team A in the group stage of the world cup?
- How many goals will be scored at the whole world cup per one match?
- Do former matches of teams A and B have any special impact to the outcome of their next match?