INFOB2DA | Practical Assignment 1

Data understanding and preprocessing (Total 100 points)

Utrecht University | Visualization and Graphics Group
Dr. Michael Behrisch
Alister Machado dos Reis, Elio Verhoef, Vincent Haverhoek, Kalee Said, Sacha Vermeer

Submission deadline:
Sunday, 17.9.2023, 23:59.

General information

- You must form groups of 3 students. Individual submissions are only accepted in special cases. Each group member must understand the entire assignment, including the code which you will create!
- You may use the following form to register your group and get assigned to a TA: https://edu.nl/4bm4j; Please register only once with three group members; We will write you an email confirming your group, TA, and werkcollege room assignment.
- Submission deadline: Sunday, 17.09.2023, 23:59pm. You will have to present your submissions on Monday, 18.09.2023 in the regular exercise/werkcollege slot. If you are unable to present your work there, send us an email.
- You will get a (more) detailed description and a general introduction of the project in the exercise/werkcollege slot Monday, 11.09.2023 17:15-19:00. *It will be beneficial for you to be there*.
- You are only allowed to use the Python programming language in your code.
- For this practical assignment, you can achieve 100 points in total.

Handing in your assignment
Hand in the Jupyter Notebook, dataset, notebook checkpoints (the folder which is automatically created in your project by Jupyter) and presentation slides as a ZIP-file on MS Teams. Make sure you press the submit/inleveren button after uploading your files in Teams, otherwise the submission is not completed.

Introduction
The aim of this practical assignment is to simulate a hands-on (visual) data analysis session in the early "data exploration and understanding" phase.

Imagine the following scenario: Your boss approaches you on Thursday evening with the following instructions: "*By the way, our client gave us this cool new dataset. Can you give me an overview of it and tell her something she did not know, yet? Ah and before I forget, …until tomorrow evening!*" Well, time is ticking… (⊙ o ⊙)

You fire up your laptop to start analyzing the dataset, which you do by creating a data analytics pipeline, focused on data understanding and visualization.

After this assignment you will be able to …
1) … reason on basic statistical insights from a (mostly numerical) dataset.
2) … use fundamental visualization functionality to get an overview about the dataset.
3) … reflect on the impact of preprocessing steps within the data analytics pipeline.
4) … retrieve the importance of dimensions in your dataset.
5) … have a structured approach to (early-stage) visualization-driven data analysis.
6) … understand which dimensions are worth exploring/harvesting in ML algorithms in the future.

Dataset overview
The mammographic_masses_data.csv file contains mammogram metrics for 961 different mammograms.

Feature	Description
BI-RADS assessment (BA)	1 to 6 (ordinal, non-predictive)
Age	Patient's age in years (integer)
Shape	Mass shape: round = 1, oval = 2, lobular = 3, irregular = 4, (nominal)
Margin	Mass margin: circumscribed = 1, microlobulated = 2, obscured = 3, ill-defined = 4, spiculated = 5, (nominal)
Density	Mass density: high = 1, iso = 2, low = 3, fat-containing = 4, (ordinal)
Severity	Benign = 0 or malignant = 1 (binominal, target/class label)

Task 0: Setup Environment (0 Points)

Goal: After you have achieved this task you are able to successfully set up a Jupyter Notebook environment.
Result: A ready to go Jupyter Notebook.
Recommend software and libraries: Python, Visual Studio Code, Jupyterlab

As always in coding, there are several paths to accomplish your goals. In INFOB2DA we support the following installation setup:

- Install Python 3.8 (to avoid compatibility issues). Make sure to check the 'add Python to PATH' option when installing Python. If you already have Anaconda installed, you can skip this step. Just make sure it is the right Python version.
- Install Visual Studio Code (VSC).
- Create a folder in VSC.
- Install JupyterLab via the terminal in VSC. Use the command '`pip install jupyterlab`' (you do not have to do this if you have Anaconda already installed).
- Use the command '`jupyter-notebook`' in the VSC terminal to open the Jupyter Notebook environment or open Jupyter Notebook through Anaconda.
- Create a new Jupyter Notebook.

Task 1: Download and import dataset (5 Points)

Goal: After you have achieved this task you are able to import datasets for further analysis.
Graded result: The result of question 1.1 must be visible in the notebook.
Recommend libraries: Pandas

- Install the Pandas library. You can do this by either using the command '`!pip install pandas`' in a notebook cell and then running it or by using the command '`pip install pandas`' in the VSC terminal.
- Import the Pandas library (try to import it as 'pd').

1.1 Use Pandas to import the dataset from the provided .CSV file. Look at the Pandas documentation to see if you can find a suitable function for this. If done correctly, you end up with a Pandas dataframe. **(5 points)**

Task 2: Get dataset on screen (15 Points)

Goal: After you have achieved this task you are able to explore the dataset with summary statistics and visualizations.
Graded result: The results of questions 2.1-2.3 must be visible in both the notebook and the presentation.
Recommend libraries: Pandas, Plotly

2.1 Use the basic summary statistics functionality of Pandas to take an initial look at the dataset. Which functions do you think everyone should be aware of to render out summary statistics? Make sure you can explain the meaning of the statistical values for this dataset. **(4 points)**

2.2 The Pandas library has advanced functions to gain more specific insights. An example of this is dataframe filtering is the '`loc`' function. Use this function to show the margin attribute of every instance of the data where the severity is 1. **(5 points)**

- Install and import the Plotly library. It is allowed to use other visualization libraries, such as Matplotlib or Seaborn if you are looking for different functionality. However, note that Plotly will be used as the main visualization library in the upcoming assignments as well.

2.3 Render at least three basic visualizations that capture the essence of the dataset. Examples of visualizations include (but are not limited to) scatterplots, heatmap/correlation matrices, and distribution plots. **(6 points)**

Task 3: Preprocessing (15 Points)

Goal: After you have achieved this task you have experienced the importance and impact of data transformations.
Graded result: The results of questions 3.1 and 3.2 must be visible in both the notebook and the presentation. The manually coded implementation of question 3.2 will be checked.
Recommend libraries: Pandas, Plotly

- Hint: use the '`copy`' function to copy the dataframe before performing the next steps.

3.1 Clean the dataset from any missing values (hint: look at the Pandas documentation). Render plots showing the difference between a cleaned and uncleaned dataset. **(5 points)**

3.2 Manually code a normalization algorithm to be performed on any variables you deem worthy of being normalized (hint: the lecture slides). Think about what kind of normalization suits the dataset. You are not allowed to use libraries or predefined functions. Render plots that show the difference between a normalized and not-normalized dataset. **(10 points)**

Task 4: Feature engineering (30 Points)

Goal: After you have achieved this task you are able to use feature engineering techniques to perform dimensionality reduction on the data and understand its impact.
Graded result: The results of questions 4.1-4.3 must be visible in both the notebook and the presentation.
Recommended Libraries: Plotly, Sklearn

- Install the Sklearn library. You do not have to import the whole library. Try to only import the modules or functions you need for the upcoming questions.

4.1 Apply one of Sklearn's automatic feature selection algorithms on the preprocessed dataset. Render plots or matrices showing the impact of this technique. **(10 points)**

4.2 Apply Sklearn's PCA dimensionality reduction algorithm on the preprocessed dataset. Match the number of principal components with the number of features which are selected in task 4.1. Render plots or matrices showing the impact of this technique. **(10 points)**

4.3 Apply Sklearn's Truncated SVD algorithm on the preprocessed dataset. Match the desired dimensionality with the number of features selected in task 4.1. Render plots or matrices showing the impact of this technique. **(10 points)**

Task 5: Reflection (35 Points)

Goal: Practice communicating your findings to domain-experts and domain-novices.
Graded result: Presentation (***max 15 minutes per group including Q&A from TAs***) and presentation quality/structure, PDF with screenshots along with descriptive text (e.g., annotations, captions, bullet points to highlight specific findings), data and performance tables, code quality and cleanness, knowledge of each team member of the code (make sure to have the code running during the presentation/Q&A).

Back to our guiding scenario: Your boss had a nice weekend, but a _lot_ of communication with the client and invited her ON MONDAY for a 12-15min status update on "how the data set understanding" is going. Do prepare this presentation and do not "just" answer the questions below question but let your fictive company shine :). Keep in mind that clients know their domain.

Guiding questions (meant solely as a loose guideline on the kind of questions you should answer in your presentation; list is non-exclusive; does not impose any order; answering all questions is not obligatory for full points):

- Give an overview of the dataset and highlight some interesting findings. (2.1-2.3)
- Why does visualization help you to get an overview of your dataset? Mention the three goals of visualization? (Whole assignment)

- Explain the reasons behind your data cleaning procedure. (3.1)
- Name three techniques which can be applied in the data preprocessing step (3)
- What are the general reasons for normalization? (3.2)
- Explain which normalization procedure you have used on which variables and why. (3.2)

- Explain the results of Sklearn's automatic feature selection. Why has it selected these features? (4.1)
- Explain the results of PCA and TSVD and its impact. (4.2, 4.3)
- Compare the applied feature reduction methods. Which performs best? (4.1-4.3)

- Is PCA always reducing the dimensions? What is the output of a PCA? (4.2)
- Is PCA robust against rotations of the whole data space (is the result still the same after rotating the data space)? (4.2)
- For which case is the usage of a PCA not reasonable? (4.2)