Assignment 2 for "Business Analytics II" course

The assignment should be done individually, the deadline is the 13th of November, 23:59. The exercises worth 40 points in total. You can find the required data files in the CSC platform, in the shared folder. There is a starting notebook, Assignment_2.ipynb, that has all the libraries that you may need to use listed for install and import, and it contains some code that you will need to use in Task 3. You have to submit the Jupyter notebook with the code that you used to solve the problems (download the notebook from the CSC platform and upload it in Moodle); in the script, discuss the results and interpret the output in comments.

1. (8 points, Association rules analysis) In this assignment, you will perform an association rules analysis on a dataset from the file social_media_user_engagement.csv. The platform is used for content sharing and interaction, with users engaging through likes, shares, comments, and follows across various content categories such as technology, fitness, travel, food, and education. Each row in the dataset represents a user's interaction with a piece of content.

   The dataset includes:

   - User ID
   - Content Category (e.g., Technology, Fitness, Food, Travel, Education)
   - Type of Interaction (Like, Comment, Share, Follow)
   - Time of Interaction (Morning, Afternoon, Evening)
   - Device Used (Mobile, Desktop, Tablet)

   Data Exploration:

   - Engagement Trends by Time of Day: which types of interactions are most common in different time periods (morning, afternoon, evening)?
   - Content Categories and Interaction Types: investigate the most common types of interactions across different content categories. Are there patterns showing that certain types of content are more likely to be liked, shared, or commented on?
   - Device Preferences: examine if users interact differently with content depending on the device they are using (mobile, desktop, tablet). For example, are mobile users more likely to "like" content, while desktop users are more prone to "comment" or "share"?

   Market Basket Analysis:

   - First, filter your data and focus on interactions made during the morning. Transform the data into transactional format (where each transaction corresponds to a user's morning interactions), extract frequent item-sets, and create association rules. Use different support and confidence values to identify the appropriate number of rules.
   - Repeat the same steps for interactions made in the evening to see if there are differences in user engagement compared to the morning.

   Interpretation and Recommendation:

   - Analyze the association rules. Are the rules reflecting common engagement patterns, or do they reveal unexpected insights about how users interact with content?

2. (6 points, Network centrality measures) In this assignment, you will analyze a dataset representing a referral network among hospitals and healthcare providers. The referrals.csv file includes patient referral connections between healthcare providers, while the hospitals.csv file contains information about the

hospitals, including their name, location, and specialization (e.g., cardiology, oncology, etc.). The ID column of hospitals.csv can be used to connect it to the links in referrals.csv. Your task is to identify the most central hospitals within this network, which signify key players in healthcare services.

The first task is to calculate the introduced centrality measures for the network:

- Degree Centrality: Identify the hospitals with the highest number of referral connections.
- Betweenness Centrality: Uncover the hospitals that act as bridges in the network, facilitating patient transfers between different healthcare providers.
- Closeness Centrality: Determine the hospitals that can reach others quickly, indicating their accessibility and importance in the network.
- PageRank: Identify the hospitals with high influence based on the structure of the referral network.

Perform the following tasks:

- Create a table showing the top 10 hospitals based on each centrality measure
- Calculate the number of common hospitals in the top 10 list for each pair of centrality measures
- Are certain specializations (e.g., cardiology or oncology) that dominate the top positions across multiple centrality measures?
- Compute the correlations between the centrality measures to evaluate the similarity in rankings

3. (16 points, Descriptive and Predictive modeling, Feature selection, Neural networks) In this exercise, you will have to analyze a dataset (cancer.csv) that includes information about breast cancer patients, with the outcome being whether it is malignant (1) or benign (0), as specified in the 'diagnosis' column.

   You have to analyse the data set from different perspectives, with the main focus on understanding what are the most important variables that determine the diagnosis.

   - As the first task, you need to perform some descriptive analysis to get a picture of the data, keeping in mind that you are mainly interested in understanding the impact of the variables on 'diagnosis'. In order to do so, you are free to choose on what other variables you focus on, but choose **at least 4** and investigate their relationship with 'diagnosis'. You may do this buy creating summary statistics and visualizations.
   - Build a Random Forest classification model for predicting 'diagnosis' using all the variables (except for the id column). What is accuracy you can achieve? What are the three most important predicting features?
   - Perform feature selection using Recursive Feature Elimination and Random Forests as it was done in the course. Calculate the accuracy when setting n_features_to_select to 3, 5, 7, and 9. Which value gives the highest accuracy?
   - The code written by a Large Language Model when prompted to build a neural network classification model with PyTorch, can be found in the Assignment_2.ipynb notebook. It is working already, but it has no comments and explanations included, so you have to understand what it does based on the lecture on the use of PyTorch. Does the model provide better performance than the Random Forest model? Try to modify different parameters that were discussed in the course (at least try structure of layers, learning rate, activation function). Add comments in the code explaining where the parameters were modified in the code, what alternatives you tried, and record for which version you got the best performance.

4. (10 points, Text analytics) In this exercise you will have to work with the data provided in the file 'trustpilot.csv'. The data contains information about 917 reviews of different e-commerce platforms from the review aggregator site Trustpilot. In the data, you have the following information:

   - Card_id: review id
   - Header: the header of the review
   - Review: the text of the review
   - Location: the location of the user submitting the review
   - shop: the e-commerce store reviewed (one of Zalando, Wish, Sheinside, Boozt or Nelly)

Start with performing some descriptive analysis on the data: check at least distribution of locations, frequency of shops, average rating per shop type.

In the analysis, you have to work with the column 'Review'. First, perform the steps of text analysis we went through in order to preprocess the data and obtain the most frequent words across all the articles. Try to iterate it at least two times, and in each iteration you should extend the set of stopwords with new ones based on the words you obtained as frequently occurring but are not particularly informative when you try to understand what the reviews are about.

In the second step, perform topic modeling on the data with specifying four topics to be extracted. Based on looking at the top 15 words from each topic, can you differentiate them and explain how they are different?