Business Analytics II first course assignment

The three
exercises worth 50 points in total; in order to pass this assignment you need at least 25 points. You have to submit a Jupyter notebook with the code that you used to solve the problems; in the script, discuss the results and interpret the output in comments.

1. (Association rules analysis, 15 points) You have to perform market basket analysis using the data from 'assignment_basket.csv'. The file includes shopping data, with additional info on the period of day and whether the transaction happened on a weekday or during the weekend. The data is in the format we have encountered before: we have several rows for each transaction, one row for each item. Before you create the association rules, check what are the most frequently sold items. Are the most frequent items the same on a weekday and in the weekend? What can you say about afternoons and mornings?

   In this case, you have to perform the analysis two times in order to discover some patterns/differences in customers' purchase behaviour in the mornings and in the afternoons. First focus on transactions that took place in the mornings, transform the data into transactional format, extract frequent item-sets and create association rules. (Note: experiment with different support and confidence values until you get a reasonable amount of item-sets and rules). Then perform the same steps with the data of transaction in the afternoons.

   - Do the extracted rules seem to contain useful information or only trivial observations (do you mostly see the items that are the most frequent in general)?
   - What would you recommend to a person (i.e. what is the consequent in association rules terminology) when you know that in the basket there is already (i) Eggs, and the purchase is in the morning, (ii) Coke and Juice, and the purchase is in the afternoon, (iii) Toast, either in the morning or in the afternoon?

2. (Classification, 25 points) In this exercise you have to work with the data in the file 'patients.csv', that contains some measurements about patients, who experienced angina, which can typically be a symptom of coronary artery disease (maybe familiar from BAI). You can find the following variables in the data:

   - age: age of the patient
   - gender: gender of the patient (0 - female, 1 - male)
   - pain: intensity of the chest pain (integer value, 0-3)
   - blood_pressure: blood pressure of the patient
   - cholesterol: cholesterol in blood, mg/dl
   - blood_sugar: indicating whether the blood sugar level is normal or not (1 when it is above 120 mg/dl, and 0 otherwise)
   - heart_rate: maximum heart rate
   - exercise: whether the chest pain was induced by some physical exercise or not (1 or 0)
   - outcome: 1 for patients with heart attack, and 0 for patients who did not have heart attack

First, try to understand the different variables using some basic statistical measures and some visualization tools. As the minimum, calculate mean values/value counts, correlation of the columns, and a couple of scatter-plots and box-plots to understand the relationship between the outcome column and the other variables.

- To run the classification models, create training and test sets (25% test set)
- Create a baseline model as a logistic regression without any parameter optimization. What is the accuracy you can achieve?
- Create models by optimizing the parameters of (i) decision trees, (ii) bagging, and (iii) random forest classifiers. What is the best accuracy you can achieve across all the models?
- What are the four most important predictors according to the best decision tree model? Create a new decision tree model that uses only those four variables as predictors. What is the best accuracy you can achieve using only the four variables?

3. (Regression, 10 points) In this exercise, you will perform tasks faced by a data scientist working in the real estate industry. Your job is to build a predictive model to estimate selling price for houses. You can find more details about the dataset (House_assignment.csv), at the following link `https://www.kaggle.com/mssmartypants/paris-housing-price-prediction`

   - Area: area of the property
   - Rooms: number of rooms
   - Yard: indicator on whether the property has a yard
   - Owners: number of previous owners
   - Year: year of construction
   - Basement: area of basement
   - Attic: area of attic
   - Garage: area of garage
   - Storage: indicator on whether there is a storage room
   - Guest: number of guest rooms
   - Price: market price of the building

   You need to perform the following tasks:

   - Exploratory data analysis: try to understand the different variables in the data. Identify the variables, based on exploratory data analysis methods (similar as in the previous task), that you think have an effect on the price of the house
   - Develop a decision tree regression model that the company can use to predict the selling price for new houses on the market. Try to optimize the parameters. What is the best MSE that you can achieve?