MCM ML 520 Assignment 2

For this assignment you will implement your first Machine Learning toolchains for both, a regression and a classification problem. Make sure to document every important step of the process in your report and answer the questions for both tasks.

Part 1 (50%): Iris Plant Species Classification

Dataset

Use the well known iris dataset (`iris.csv`) to perform your first ML classification. Original source: https://gist.github.com/curran/a08a1080b88344b0c8a7#file-iris-csv. Alternatively, scikit-learn has the iris dataset already included in their datasets collection, so you can also load it from there.

Task

1. Analyze the data using the same techniques as for the last assignment. Decide for yourself which and how to use the specific commands. Answer the following questions in the report and include figures supporting your answers:
   - Which classes exist? Are they (roughly) balanced?
   - Which noteworthy trends of features and relations between features as well as features and Classes do you see?
   - If you would need to distinguish the classes with those features, which features would you choose, any why?
2. In order to classify the three different Iris plant species, set up your first ML toolchain including the following steps:
   - Data and Feature Preprocessing (if necessary and applicable)
     - – Are there any outliers in the data which might need to be removed?
     - – Are there any missing values which need to be taken care of?
     - – Do you need to apply any feature preprocessing steps? (e.g Normalization, Feature Deletion/Reduction/Addition)
     - – Are there any categorical features that need to be transformed so that it can be used for classification task?
     - – Do you think it makes sense to derive any more features from the given ones? Why/why not?

- Split up the dataset into a training and a separate held back test set in a clever way
  - Why is such a train/test split important?
  - Which train/test split percentage do you choose and why?
  - Think about how can you make sure to include samples from all three classes in both datasets and why this is important.
- Train different classification models to distinguish between the three Iris Plant Species:
  - Use the following models: k Nearest Neighbour, Decision Tree, Support Vector Machine
- Use different hyperparameter settings for each model and explain why and how you chose them
- Use an appropriate cross-validation setup for the training
- Estimate the models' performances on the held back test set:
- Compare the models with their hyperparameter settings with two different error/performance measures * Why did you chose the specific error/performance measures? * What do they tell you?
  - Which model performs best with which hyperparameter settings and why do you think it does that way?
- Explain which model you would use in deployment and why

Part 2 (50%): Boston House Price Prediction

Dataset

Use the well known benchmark dataset boston housing (`housing.csv`) to perform your first ML regression. Original source: https://www.kaggle.com/vikrishnan/boston-house-prices Alternatively, scikit-learn has the boston housing dataset already included in their datasets collection, so you can also load it from there. Read up about the structure and content of the dataset and then use the 'MEDV' column as your regression target.

Task

1. Analyze the data using the same techniques as for the last assignment. Decide for yourself which and how to use the specific commands. Answer the following questions in the report and include figures supporting your answers:
   - Which noteworthy trends of features and relations between features as well as features and regression target do you see?
   - Which features would you choose to train the regression models, any why?
2. Build up your ML toolchain for this regression problem similar to the one you did for the classification and again take care of the following points:
   - Data and Feature Preprocessing (if necessary and applicable)
   - Train/Test split

- Use the following Regression models with different hyperparemter settings (where applicable) and an appropriate cross-validation setup for your training:
  - Linear Regression
  - Polynomial Regression
  - Logistic Regression
- Estimate the models' performances on the test set again with two different error/performance measurements
- Explain which model you would use in deployment and why