Week 4 – Regression (formative assignment)

In this assignment you will perform an analysis with different models and hyperparameters and compare the outcomes.

As part of the assignment you will create, as specified in the rubric (criteria 1 through 3):

1. *a written justification of the choice for one or more machine learning models suitable for the problem.*
2. *a detailed analysis using an existing implementation.*
3. *a written evaluation of the model's quality using an appropriate criterion.*

Work out your answers to the assignments/questions below in a Jupyter Notebook. Clearly separate natural text in Markdown cells (used for explanations, justifications, conclusions and reflections) from program code (used for analysis or visualisations).
At the same time ensure the notebook still provides a cohesive account of the analysis.

Assignment Description

The attached file ' Life Expectancy Data v2.csv' (from Kaggle) contains data from the World Health Organization on some health-related characteristics. The target for this dataset is life expectancy, and the purpose of this assignment is to build a regression model to predict life expectancy.

Start by studying the data: read attached the description, review the data, clean up if necessary, and determine which variables are appropriate to include in a regression model.

Build several regression models using the built-in functions in scikit-learn, always using the same features and train/test split. Compare the results of the different models and indicate which model you would prefer.

Compare the following models:

- Multivariate linear regression models
- Decision tree regression; tune the hyperparameters with grid search
- Bagging with regression trees
    - Tune the parameters with GS
    - Compare your models by plotting the amount of trees with your metric
    - Evaluate your best model visually and with a metric
    - Write a conclusion
- AdaBoosting with regression trees. Follow the same steps as 3.
- RandomForest regression. Also follow the same steps as 3.
- Visualise the feature importances of your random forest model in a sorted barplot. Which features are used often/rarely?

At the end of your notebook write a conclusion about which model performs best.

Assignment deliverables

Please submit the following three components:

1. **Documented analysis**
   Submit a compressed (zipped) folder containing your Python script or Jupyter notebook.
    - Do not include the dataset, output files generated in the script or your virtual environment.
    - If you have used a virtual environment, include a requirements.txt file listing all required libraries.

   Format: .zip
2. **Printable Version of Your Code**
   Provide a PDF version of your code (either notebooks or stand-alone scripts). This helps us offer detailed, line-by-line feedback.
    a. Ensure code readability: use landscape orientation if needed to accommodate longer lines.

   Format: .pdf
3. **Individual Contribution & AI Usage Report**
   Write a short document addressing the following points:
    a. **Individual Contribution**: Describe which parts of the assignment you completed and what responsibilities you took on.
    b. **AI Usage**: Specify whether you used AI tools and, if so, how. Be transparent about the extent and purpose of any AI support.

   Format:.pdf

Important Notes

1. Before submitting your notebook, restart the kernel and run all cells to ensure it executes cleanly from top to bottom. Remove any error messages or irrelevant outputs.
2. As with all your written assignments, you may build upon external ideas (including those from chatbots or AI tools), provided you cite your sources. Use APA or IEEE referencing style.
3. AI usage must be clearly documented. You may use AI tools as a source of feedback or inspiration, but the final work **must** be your own.