COMPUTATIONAL ECONOMICS II
ASSIGNMENT 4
Author: Federico Durante

Machine Learning Classification for Economic Growth Prediction
Analyzing OECD NUTS3 Data

- **What does the code do?**

This script applies machine learning classification to predict whether a region's economy grew by more than 5% based on OECD data. The dataset includes economic and demographic indicators for NUTS3 regions in 2019. The script loads and preprocesses the data, splits it into training and test sets, and runs three classifiers: Random Forest, Logistic Regression, and Multilayer Perceptron. Each model's performance is evaluated using accuracy, precision, recall, F1-score, and a confusion matrix.

- **How to use the code?**

To run the code, ensure you have *Python* installed with required libraries (*pandas, numpy, sklearn*). Place the dataset files (*assignment_4_OECD_data_growth.csv* and *assignment_4_OECD_data_explaining_vars.csv*) in the same directory as the script. Execute the script (python script_name.py), and it will automatically load and process the data, train the classifiers, and display the results. The output includes classification metrics and a confusion matrix, helping to evaluate each model's performance in predicting economic growth.

- **Interpretation of the results**

```
Results for MLP Classifier
Precision: 0.7972972972972973
Recall: 0.8872180451127819
F1 score: 0.8398576512455516
Accuracy: 0.7680412371134021
Confusion matrix:
 [[ 31  30]
 [ 15 118]]
```

```
Results for Logistic Regression
Precision: 0.7245508982035929
Recall: 0.9097744360902256
F1 score: 0.8066666666666666
Accuracy: 0.7010309278350515
Confusion matrix:
 [[ 15  46]
 [ 12 121]]
```

```
Results for Random Forest
Precision: 0.75
Recall: 0.924812030075188
F1 score: 0.8282828282828283
Accuracy: 0.7371134020618557
Confusion matrix:
 [[ 20  41]
 [ 10 123]]
```

MLPClassifier had the highest accuracy (76.8%) and a strong recall (88.7%), meaning it correctly identified most growing regions. However, its false positive rate (30 cases) indicates that it often misclassified non-growing regions as growing.

Random Forest showed a well-balanced performance with the highest recall (92.5%), meaning it detected nearly all growing regions. Its accuracy (73.7%) was slightly lower than MLP, and while it had fewer false negatives (10), it misclassified 41 non-growing regions as growing.

Logistic Regression had the lowest accuracy (70.1%) and the highest false positive rate (46 cases), meaning it struggled the most in distinguishing non-growing regions. However, its recall (90.9%) was still strong, showing that it was effective at identifying growth cases.

In conclusion, **MLPClassifier** performed best, offering the highest accuracy and a strong balance of precision and recall. Random Forest is a close second, excelling in recall while maintaining solid accuracy. Logistic Regression performed the weakest, as its high false positive rate makes it less reliable for distinguishing non-growing regions.

It is important to note that results may change from one execution to another because of the random initialization of model parameters, stochastic training processes (especially in MLPClassifier), and floating-point precision differences. Additionally, if the random seed isn't fixed, the train-test split may vary, leading to different model evaluations.