Due Date: 23.11.2025 23:55

CENG 463 – Introduction to Machine Learning HW1

- Random Forest
- SVM
- KNN
- Gaussian Naive Bayes
- Logistic Regression

!!!Please read the homework instructions, rules and the sections declared as important carefully.

The dataset utilized in this assignment originates from the World Resources Institute (WRI) – Aqueduct Water Risk Atlas. It provides country-level indicators describing key hydrological and environmental factors, which are listed in the table below. The objective is to classify each country into a Water Resource Risk Category (0-4) using these indicators. Students are also expected to create two derived features — Composite Water Stress Index (CWSI) and Seasonal–Flood Interaction (SFI) — to enhance model performance.

Feature	Description
gid_0	Country Code
bws_score	Baseline Water Stress
gtd_score	Groundwater Depletion
drr_score	Drought Risk
rfr_score	River Flood Risk
sev_score	Seasonal Variability
w_awr_def_tot_cat	Target: Water Risk Category(0-4)

Please follow the steps below to fill out the provided .ipynb file.

1. Feature Engineering (35 pts)

Students are expected to create two new features based on the existing indicators. These derived variables aim to better represent the combined effects of multiple risk factors.

- **Composite Water Stress Index (CWSI):** CWSI combines three components — baseline water stress, groundwater depletion, and drought risk — into a single metric representing overall water pressure. Higher values indicate higher stress.

$$\mathbf{CWSI = 0.5 \times bws\_score + 0.3 \times gtd\_score + 0.2 \times drr\_score}$$

- **Seasonal–Flood Interaction (SFI):**

SFI represents the interaction between seasonal variability and river flood risk. Higher values indicate regions with both high variability and flood risk.

SFI = sev_score × rfr_score

2. Model Training & Evaluation(40 pts)

Train five classification models: Random Forest, SVM, KNN, Gaussian Naive Bayes, and Logistic Regression. You will use Sklearn algorithms for this task and apply them to obtain the accuracy scores for each technique and prepare an evaluation table.

Hint: Use scaled data for SVM, KNN, and Logistic Regression.

3. Hyperparameter Optimization(15 pts)

Tune each model using GridSearchCV with 5-fold cross-validation. Compare baseline and tuned results. Applying regularization and experimenting parameters of techniques to improve the initial accuracy scores.

Hint: Use accuracy as the scoring metric.

Hint: Add classification report.

4. Feature Importance Analysis(10 pts)

Choose one model and analyze feature importance. Present most influential features in a table and bar chart.

Hint: Visualize features sorted by importance.

!!!!!IMPORTANT

At the end of each task, add a short discussion section where you evaluate the results and comment on the model's performance and suitability for the dataset.
During hyperparameter tuning, compare baseline and tuned models and discuss how parameter optimization affects performance.
In the feature importance section, interpret the most and least influential features and explain how they relate to the model results.
Finally, prepare a comparison table including all five models (Random Forest, SVM, KNN, Gaussian Naive Bayes, Logistic Regression) and explain which model performs best and why.
All discussions and comparisons must also be clearly summarized in the HW1_Report.pdf file.

Assignment Rules:

1. In this homework, no cheating is allowed. If any cheating is detected, the homework will be graded as 0, and no further discussion will be entertained.

2. You are expected to submit your homework in groups. Therefore, it will be sufficient if only one member of the group submits the homework.

3. You must **fill in the provided Jupyter Notebook file** named:
   - CENG463_HW1_GROUPID.ipynb

   Follow the steps and instructions **provided in this file carefully.**

4. You must upload a **.zip file** to MS Teams. This file must include:
   - **CENG463_HW1_GROUPID.ipynb**

     Example: CENG463_HW1_G01.ipynb

   - **CENG463_HW1_GROUPID_Report.pdf**

     Example: CENG463_HW1_G01_Report.pdf

5. The .zip file must be named in the following format: **CENG463_HW1_GROUPID.zip**
   Example: CENG463_HW1_G01.zip
6. Please be aware that if you do not follow the assignment rules regarding export format and naming conventions, you will lose points.