CENG463 Assignment 1

Spring 2024-2025

Deadline
23.05.2025, 23.55

Assignment Instructions

- You are expected to complete the assignment using the **Google Colab** environment. However, to prevent potential issues, both the **Colab notebook link** and the corresponding **.ipynb file** must be submitted.
- The submitted Python file must be named according to the following convention, listing student numbers in **ascending numerical order**:
  CENG463_Studentnum1_Studentnum2_Studentnum3.ipynb

Rules and Restrictions

- Collaboration across groups, the use of publicly available code, and the generation of code via artificial intelligence tools are **forbidden.**
- If you have any further inquiries, please feel free to contact me via email at: **cerensozeri@iyte.edu.tr**

Tasks

- The provided dataset has already been divided into **training, validation, and test** sets. You are required to use the dataset as-is, without modifying the splits. A detailed description of the dataset is available on the final page of this document.
- You are expected to perform **feature selection**. The choice of algorithm is up to you; however, the number of **selected features must be 10.**
- You are required to implement and evaluate the following three classification models: **Decision Tree**, **Nearest Neighbour**, and **Naive Bayes**. You are expected to **implement both the Naive Bayes** and **Nearest Neighbour classifiers from scratch**, without relying on Scikit-Learn Classes.
- You are also required to implement the hard voting **yourself**, using the predictions from all three models.

Hyperparameters

- For the **Decision Tree**, use *random_state=42* and accept all other default parameters.
- For **Nearest Neighbour**, set the **number of neighbors to 5** and use the **Hamming** distance metric.

Evaluation

- To assess the performance of each model, use the **confusion matrix** along with **precision, recall, accuracy,** and **F1-score** metrics.
- Based on the results, you are expected to provide a **detailed explanation** of why one model performs better or worse than the others.

Hamming Distance

Hamming Distance is a measure of dissimilarity that counts the number of positions at which the corresponding symbols in two equal-length strings are different. It is used to identify and quantify errors or differences between two data sequences.

Example: p_1=10**10**1; p_2=10**01**1;

$d(p_1,p_2)$= 2 because the bit-vectors differ in the 3rd and 4th positions.

Target: edible = 0, poisonous = 1

dangerous_shape: 1 = Cap (mushroom head) shape is considered risky. 0 = Cap shape is typical and safer.

irregular_surface: 1 = Cap surface is rough or uneven. 0 = Cap surface is smooth.

dark_cap_color: 1 = Cap has a dark or unusual color. 0 = Cap has a light or common color.

has_bruises: 1 = Visible bruises are present on the mushroom. 0 = No bruising observed.

strong_odor: 1 = Odor is strong or unpleasant. 0 = Odor is mild, pleasant, or absent.

non_free_gills: 1 = Gills are attached to the stalk. 0 = Gills are not attached to the stalk.

dense_gills: 1 = Gills are tightly packed. 0 = Gills are widely spaced.

narrow_gills: 1 = Gills are thin and narrow. 0 = Gills are broad.

dark_gill_color: 1 = Gills have dark pigmentation. 0 = Gills are lightly colored.

tapering_stalk: 1 = Stalk narrows toward the base. 0 = Stalk thickens toward the base.

root_missing_or_unusual: 1 = Root is absent or has an irregular form. 0 = Root is present and has a standard form.

irregular_surface_above: 1 = Surface above the ring is textured. 0 = Surface above the ring is smooth.

irregular_surface_below: 1 = Surface below the ring is textured. 0 = Surface below the ring is smooth.

dark_stalk_above: 1 = Stalk above the ring is dark-colored. 0 = Stalk above the ring is light-colored.

dark_stalk_below: 1 = Stalk below the ring is dark-colored. 0 = Stalk below the ring is light-colored.

non_white_veil: 1 = Veil is any color other than white. 0 = Veil is white.

multiple_rings: 1 = There is more than one ring. 0 = There is one or no ring.

complex_ring_type: 1 = Ring structure is large or visually complex. 0 = Ring is simple or missing.

dark_spore: 1 = Spore print is dark in color. 0 = Spore print is light in color.

dense_population: 1 = Mushrooms grow in dense clusters or groups. 0 = Mushrooms are sparse or scattered.

disturbed_habitat: 1 = Mushroom grows in artificial or disturbed environments. 0 = Mushroom grows in natural environments.