Exercise #06

IT University of Copenhagen (ITU)
Data Mining KSD (DAMIN)
(Autumn 2025)

30 September 2025

Introduction This exercise focuses on deepening your understanding of data preparation, classification, and visualization using the `penguins` dataset. Through hands-on tasks, you will explore key techniques such as handling missing data, applying k-Nearest Neighbors, logistic regression, and decision trees. The dataset offers practical experience for analyzing classification problems in the context of real-world data. The objectives of this exercise are:

- Explore and prepare datasets by identifying differences and handling missing data.
- Apply k-Nearest Neighbors (k-NN) to classify missing values based on physical attributes.
- Develop logistic regression models to classify species, using data splitting methods for testing and training.
- Create a decision tree to classify species using a step-by-step, rule-based approach based on physical measurements.
- Utilize Python libraries such as Pandas, Scikit-learn, and Matplotlib to implement and visualize your solutions.

By completing these tasks, you will gain practical insights into the process of classification and data analysis, which are essential for data mining and machine learning.

Exercise 06.01. *Data Preparation* (10-20 minutes) – In this exercise, you will work with the penguins dataset provided in 3 different files: `original_penguins.csv`, `penguins_knn.csv`, and `penguins_no_chinstraps.csv`. The goal is to explore these datasets, identify differences, and use data visualization techniques to analyze their structure. This foundational step is critical in understanding how to handle and prepare data for further classification tasks.

- **Instructions:**
  - Explore the datasets.

    - Load each of the three datasets (`original_penguins.csv`, `penguins_knn.csv`, and `penguins_no_chinstraps.csv`) into separate Pandas `DataFrames`.
    - Inspect each dataset's structure by viewing the first few rows and checking the column names, data types, and presence of any missing values.
  - Compare the datasets.
    - Identify the key differences between the datasets.
    - Determine which dataset contains missing species entries and which dataset is the simplest in terms of features and structure.
  - Visualize the datasets.
    - Create basic plots (e.g., histograms or scatter plots) to gain insights into the distributions of key features like bill length, bill depth, flipper length, and body mass.
    - Use the visualizations to analyze and describe which dataset might be most suitable for different classification methods such as k-NN, regression, or decision trees.

- **Expected Output/Questions:**
  1. What are the main differences between the three datasets?
  2. Which dataset has missing species entries?
  3. Which dataset do you find to be the simplest in terms of features?
  4. Provide a brief description of the visual insights obtained from the dataset, particularly in relation to classification methods.

Exercise 06.02. *The Closest Neighbors Are the Trustful Ones!* (30-45 minutes) – In this exercise, you will use the k-Nearest Neighbors (k-NN) algorithm to classify missing species data from the `penguins_knn.csv` dataset. By leveraging different physical attributes such as bill length, bill depth, flipper length, and body mass, you will develop a routine to predict the missing species entries.

- **Instructions:**
  - Load the dataset.
    - Load the `penguins_knn.csv` dataset into a Pandas `DataFrame`.
  - Classify missing species using k-NN.
    - Develop separate routines to classify the missing species based on each of the following attributes individually:
      1. Bill length (mm).
      2. Bill depth (mm).
      3. Flipper length (mm).

        4. Body mass (g).
    - ∗ Implement a combined system that uses all of the above attributes simultaneously to classify the missing species.
  - Compare results.
    - ∗ Compare the classification results based on each attribute and the combined system.
    - ∗ Use the original `original_penguins.csv` dataset to validate your results by comparing them with the true species.
    - ∗ Calculate the accuracy of your method.

- **Expected Output/Questions:**

  1. Which attribute provided the best classification results when used individually?
  2. How does the combined classification system perform compared to using individual attributes?
  3. What is the overall accuracy of the system? Provide an error analysis.

Exercise 06.03. *Guess Which Species by Attribute!* (30-45 minutes) – In this exercise, you will apply logistic regression to classify penguins into two species based on selected physical attributes. You will experiment with different data splits for training and testing, and evaluate the accuracy of your models using the `penguins_no_chinstraps.csv` dataset.

- **Instructions:**

  - Load the dataset.
    - ∗ Load the `penguins_no_chinstraps.csv` dataset into a Pandas `DataFrame`.
  - Split the data.
    - ∗ Split the data into training and testing sets using two different ratios:
      1. 50% training data and 50% testing data.
      2. 80% training data and 20% testing data.
  - Perform logistic regression.
    - ∗ Select two attributes from the available ones (e.g., bill length, flipper length, body mass) to classify the penguins into two species using logistic regression.
    - ∗ Plot the decision boundaries for your logistic regression model to visually inspect the classification results.
  - Evaluate the model.
    - ∗ For each data split (50/50 and 80/20), calculate the accuracy and precision of the model.
    - ∗ Compare the results for each attribute pair and each data split, and determine which configuration provided the best results.

- **Expected Output/Questions:**
  1. Which attribute pair provided the best classification accuracy?
  2. How did the 50/50 and 80/20 data splits affect the model's performance?
  3. Provide accuracy and precision scores for both data splits, and interpret the model's effectiveness.

Exercise 06.04. *How Can a Tree Help!* (30-45 minutes) – In this exercise, you will develop a rule-based decision tree to classify penguins into species based on physical attributes such as island of finding, body weight, and flipper length. You will create a manual decision tree to classify the species using the `original_penguins.csv` dataset. – *Homework Exercise*

- **Instructions:**
  - Load the dataset.
    - Load the `original_penguins.csv` dataset into a Pandas `DataFrame`.
  - Analyze the data.
    - Study the data to identify patterns that can help exclude certain species based on the penguins' island of finding.
    - Calculate the average and standard deviation of body weight and flipper length for each species.
  - Create a decision tree.
    - Based on your analysis, construct a human-readable decision tree (using `if` and `elif` statements) to classify penguins into species.
    - Use attributes in the following order to classify the penguins:
      1. Island of finding.
      2. Body weight.
      3. Flipper length.
  - Test the decision tree.
    - Select a random subset of 10-20 penguins from the dataset and use your decision tree to classify them into species.
    - Calculate the accuracy of your decision tree by comparing the predicted species with the actual species.
- **Expected Output/Questions:**
  1. What are the average and standard deviation of the body weight and flipper length for each species?
  2. How did your decision tree perform? Provide the accuracy for your subset of penguins.
  3. Explain how the attributes used (island, body weight, flipper length) influenced the classification.