**Assignment 3, data mining**

Put all deliverables into github repository in your profile. Defend by explaining deliverables and answering questions.

Deliverables: code, report (pdf)

Google form (before teams deadline): <https://docs.google.com/forms/d/e/1FAIpQLSe0GyNdOYlvM1tX_I_CtlPod5jBf-ACLGdHYZq1gVZbUeBzIg/viewform?usp=sf_link>

### **Basic Exercises**

#### **1. Basic Classification Methods**

**Exercise 1: Implementing Basic Classification Algorithms**

* **Objective:** Get familiar with basic classification techniques using Scikit-learn.
* **Tasks:**
  + Load a simple dataset (e.g., Iris or Wine dataset).
  + Implement a basic classification algorithm (e.g., Logistic Regression or K-Nearest Neighbors).
  + Train the model and evaluate its accuracy using a test set.

**Exercise 2: Confusion Matrix and Classification Report**

* **Objective:** Understand model evaluation metrics.
* **Tasks:**
  + Use the trained model to make predictions on the test set.
  + Generate a confusion matrix and classification report using Scikit-learn.
  + Discuss the results, focusing on precision, recall, and F1-score.

#### **2. Clustering Techniques**

**Exercise 3: Implementing K-Means Clustering**

* **Objective:** Apply K-Means clustering to a dataset.
* **Tasks:**
  + Load a dataset (e.g., customer data or a toy dataset).
  + Use K-Means to group the data into clusters.
  + Visualize the clusters using a scatter plot.

**Exercise 4: Evaluating K-Means Clustering**

* **Objective:** Evaluate the quality of clustering results.
* **Tasks:**
  + Calculate the inertia (sum of squared distances of samples to their closest cluster center).
  + Experiment with different numbers of clusters and plot the inertia to find the optimal number of clusters (Elbow Method).
  + Discuss the results.

#### **3. Introduction to Advanced Clustering Techniques**

**Exercise 5: Implementing Hierarchical Clustering**

* **Objective:** Explore hierarchical clustering.
* **Tasks:**
  + Apply Agglomerative Clustering to the same dataset used in K-Means.
  + Visualize the results using a dendrogram to illustrate the clustering hierarchy.
  + Compare the results with those from K-Means.

**Exercise 6: Introduction to DBSCAN**

* **Objective:** Understand density-based clustering.
* **Tasks:**
  + Implement DBSCAN on the dataset and visualize the clusters.
  + Discuss the parameters (eps and min\_samples) and their effect on clustering results.
  + Compare the performance of DBSCAN with K-Means.

### **Basic Report Structure**

#### **Title Page**

* Title of the Report
* Student Name
* Course Title
* Date

#### **Table of Contents**

* Introduction
* Basic Classification Methods
  + Overview
  + Exercises
  + Findings
* Clustering Techniques
  + Overview
  + Exercises
  + Findings
* Introduction to Advanced Clustering Techniques
  + Overview
  + Exercises
  + Findings
* Conclusion
* References

#### **Introduction**

* Introduce the relevance of classification and clustering in machine learning.
* Outline the objectives and structure of the report.

#### **Basic Classification Methods**

* **Overview:** Explain what classification methods are and their importance.
* **Exercises:**
  + Summarize the steps taken in each exercise and the algorithms used.
* **Findings:** Discuss the model performance and insights from the confusion matrix and classification report.

#### **Clustering Techniques**

* **Overview:** Describe the concept of clustering and its applications.
* **Exercises:**
  + Summarize the implementation of K-Means and its evaluation.
* **Findings:** Discuss the quality of the clusters obtained and the results of the Elbow Method.

#### **Introduction to Advanced Clustering Techniques**

* **Overview:** Briefly introduce hierarchical clustering and density-based clustering.
* **Exercises:**
  + Summarize the methods implemented in hierarchical clustering and DBSCAN.
* **Findings:** Compare the results of hierarchical clustering and DBSCAN with K-Means.

#### **Conclusion**

* Summarize key learnings from all exercises.
* Reflect on the applications of classification and clustering techniques in real-world scenarios.

#### **References**

* Cite any resources, articles, and documentation referenced throughout the report.

### **Submission Guidelines**

* Format: PDF
* Length: 10-15 pages, excluding title page and references.