### **Midterm Project**

#### **Send to teams and google form before 20.10.2024 23:59.** Write report in pdf. Report must be according to the template. Plagiarism of code or report leads to 0.

#### **Link to google form** [**https://docs.google.com/forms/d/e/1FAIpQLSc7dGq5gnUqvBIQ3wW7E-TjnQmHQZI\_uj\_9g3OLofseSERUow/viewform?usp=sf\_link**](https://docs.google.com/forms/d/e/1FAIpQLSc7dGq5gnUqvBIQ3wW7E-TjnQmHQZI_uj_9g3OLofseSERUow/viewform?usp=sf_link)

#### **Project Overview**

This project focuses on applying data mining techniques to analyze data of your own choice (you can find in Kaggle, UCI). The project will involve data preprocessing, exploration, visualization, feature selection, and the implementation of both basic and advanced classification techniques.

#### **Key Technologies and Libraries Used**

* Python
* Pandas
* NumPy
* Matplotlib and Seaborn (for visualization)
* Scikit-learn
* Statsmodels

### **Project Requirements**

1. **Introduction to Data Mining**
   * Understand the fundamental concepts of data mining, its importance in business intelligence, and the types of data mining techniques.
2. **Data Preprocessing**
   * Collect and clean a dataset
   * Handle missing values, encode categorical variables, and scale numerical features to prepare the data for analysis.
3. **Data Exploration and Visualization**
   * Conduct exploratory data analysis (EDA) to understand the dataset.
   * Use visualizations (histograms, scatter plots, box plots) to uncover patterns and relationships in the data.
4. **Feature Selection and Dimensionality Reduction**
   * Identify relevant features that contribute to customer segmentation.
   * Apply dimensionality reduction techniques (e.g., PCA) to simplify the dataset while retaining important information.
5. **Classification Techniques**
   * Implement basic classification algorithms (e.g., Logistic Regression, Decision Trees) to predict customer segments.
   * Evaluate model performance using metrics such as accuracy, precision, and recall.
6. **Advanced Classification Methods**
   * Explore advanced classification methods (e.g., Random Forest, Support Vector Machines, Neural Networks).
   * Compare the performance of these methods against basic classifiers and select the best-performing model.

### **Project Report Template**

## **Project Report: Customer Segmentation and Prediction Using Data Mining in Python Times new roman 12 pt, single line spacing. Minimum number of pages is 10.**

### **Table of Contents**

1. **Executive Summary**
2. **Introduction**
3. **Project Objectives**
4. **Data Mining Overview**
5. **Data Preprocessing**
6. **Data Exploration and Visualization**
7. **Feature Selection and Dimensionality Reduction**
8. **Classification Techniques**
9. **Advanced Classification Methods**
10. **Testing and Model Evaluation**
11. **Conclusion**
12. **References**
13. **Appendices**

### **1. Executive Summary**

Provide a brief overview of the project's goals, the techniques used, and the outcomes achieved.

### **2. Introduction**

Introduce the concept of data mining and its significance in understanding customer behavior. Discuss the motivation behind choosing this project.

### **3. Project Objectives**

List the specific objectives of the project, such as identifying customer segments and predicting future purchasing behavior.

### **4. Data Mining Overview**

Describe the key concepts of data mining, including its purpose, methodologies, and applications in various fields.

### **5. Data Preprocessing**

* **Data Collection**: Outline the dataset used (source, size, features).
* **Cleaning and Preparation**: Detail the steps taken for data cleaning, handling missing values, and encoding categorical variables.

### **6. Data Exploration and Visualization**

* **Exploratory Data Analysis**: Summarize the findings from EDA, highlighting key insights.
* **Visualizations**: Include relevant visualizations that illustrate data patterns, distributions, and relationships.

### **7. Feature Selection and Dimensionality Reduction**

* **Feature Importance**: Discuss the methods used for feature selection.
* **Dimensionality Reduction**: Explain the application of PCA or other techniques, including the rationale for their use.

### **8. Classification Techniques**

* **Basic Classifiers**: Describe the implementation of basic classification algorithms, including code snippets and results.
* **Model Performance**: Evaluate and compare the performance of the models using various metrics.

### **9. Advanced Classification Methods**

* **Advanced Techniques**: Outline the advanced classification methods implemented, including the reasoning behind their selection.
* **Performance Comparison**: Compare the results of advanced methods against basic classifiers, using visual aids to support findings.

### **10. Testing and Model Evaluation**

Discuss the testing methodologies used to validate model performance, including cross-validation techniques and evaluation metrics.

### **11. Conclusion**

Reflect on the project outcomes, the effectiveness of the applied techniques, and suggestions for future work or improvements.

### **12. References**

List all resources, documentation, articles, and datasets referenced throughout the project.

### **13. Appendices**

Include additional diagrams, code snippets, or data that support the report but are too detailed for the main sections.