**Assignment 4, data mining**

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>

#### **Exercise 1: Anomaly Detection**

**Objective:** Implement an anomaly detection system using Python libraries.

**Tasks:**

1. **Data Collection:**
   * Use a publicly available dataset (e.g., KDD Cup, Credit Card Fraud Detection) or generate synthetic data for anomaly detection.
   * Load the dataset using pandas and perform initial exploratory data analysis (EDA).
2. **Data Preprocessing:**
   * Handle missing values and outliers in the dataset.
   * Normalize or standardize features as necessary.
3. **Anomaly Detection Techniques:**
   * Implement at least two different anomaly detection algorithms, such as:
     + **Statistical Method:** Use Z-score or IQR for univariate anomaly detection.
     + **Machine Learning Method:** Implement Isolation Forest or One-Class SVM from scikit-learn.
     + **Deep Learning Method:** Build an autoencoder for anomaly detection.
4. **Model Evaluation:**
   * Split the dataset into training and testing sets.
   * Evaluate the performance of the models using metrics such as precision, recall, and F1-score.
   * Visualize the results using confusion matrices and ROC curves.
5. **Visualization:**
   * Use matplotlib or seaborn to visualize the detected anomalies against the normal data points.
   * Create plots that showcase the distribution of features and the anomalies.
6. **Reporting Findings:**
   * Summarize the detection results, including the number of detected anomalies and the effectiveness of each method.

#### **Exercise 2: Time Series Analysis**

**Objective:** Analyze and forecast a time series dataset using Python libraries.

**Tasks:**

1. **Data Collection:**
   * Select a time series dataset (e.g., stock prices, weather data, or sales data).
   * Load the dataset using pandas and perform initial EDA.
2. **Data Preprocessing:**
   * Handle any missing values in the time series data.
   * Resample the data to a uniform time interval if necessary (e.g., daily, weekly).
3. **Exploratory Data Analysis:**
   * Visualize the time series data using line plots to identify trends, seasonality, and patterns.
   * Decompose the time series into trend, seasonality, and residuals using seasonal decomposition.
4. **Modeling:**
   * Implement forecasting techniques, such as:
     + **ARIMA Model:** Fit an ARIMA model to the data and determine the appropriate parameters (p, d, q).
     + **Exponential Smoothing:** Use Holt-Winters method for forecasting seasonal data.
     + **Machine Learning Approach:** Implement a model using scikit-learn (e.g., Random Forest, LSTM for deep learning).
5. **Model Evaluation:**
   * Split the dataset into training and testing sets.
   * Evaluate the forecast accuracy using metrics such as MAE, RMSE, and MAPE.
   * Visualize the predicted vs. actual values.
6. **Reporting Findings:**
   * Summarize the results of the analysis, including the accuracy of different forecasting methods and insights gained from the time series analysis.

### **Report Structure**

1. **Title Page**
   * Title: “Anomaly Detection and Time Series Analysis Using Python”
   * Author’s Name
   * Date of Submission
   * Organization Name (if applicable)
2. **Executive Summary**
   * Brief overview of the key findings and methodologies used for anomaly detection and time series analysis.
3. **Table of Contents**
   * List all sections and subsections with corresponding page numbers.
4. **Introduction**
   * Overview of the importance of anomaly detection and time series analysis in data mining.
   * Purpose and scope of the report.
5. **Anomaly Detection**
   * **Data Collection and EDA**
     + Description of the dataset used and initial exploratory analysis.
   * **Data Preprocessing**
     + Steps taken to clean and prepare the data.
   * **Anomaly Detection Techniques**
     + Overview of the methods implemented (e.g., statistical, machine learning, deep learning).
     + Comparison of their performance and effectiveness.
   * **Visualization and Results**
     + Visual representations of the detected anomalies.
     + Summary of findings, including metrics used for evaluation.
6. **Time Series Analysis**
   * **Data Collection and EDA**
     + Description of the time series dataset used and initial exploratory analysis.
   * **Data Preprocessing**
     + Steps taken to clean and prepare the time series data.
   * **Exploratory Data Analysis**
     + Visualizations showing trends, seasonality, and patterns.
   * **Modeling**
     + Description of the forecasting models used (e.g., ARIMA, Exponential Smoothing).
     + Evaluation of their performance using accuracy metrics.
   * **Visualization and Results**
     + Comparison of predicted vs. actual values.
     + Summary of insights gained from the analysis.
7. **Conclusion**
   * Summary of key findings from both the anomaly detection and time series analysis.
   * Discussion on the significance of these analyses in real-world applications.
8. **Recommendations**
   * Practical suggestions for future work, improvements, or applications of the methods discussed.
9. **References**
   * List of all sources, libraries, and datasets referenced in the report.
   * Use a consistent citation style (e.g., APA, MLA).
10. **Appendices**
    * Additional material, such as code snippets, detailed visualizations, or supplementary analysis.

### **Rules for Writing the Report**

1. **Clarity and Conciseness**
   * Use straightforward language to explain complex concepts.
   * Avoid unnecessary jargon and define technical terms.
2. **Objective Tone**
   * Maintain a formal and unbiased tone throughout the report.
   * Base statements on factual information.
3. **Structure and Formatting**
   * Use headings and subheadings for clear organization.
   * Ensure consistent formatting (font, size, spacing) throughout.
4. **Visual Aids**
   * Include diagrams, tables, and charts for better comprehension.
   * Ensure all visual aids are clearly labeled and referenced in the text.
5. **Proofreading and Editing**
   * Carefully proofread to correct spelling and grammatical errors.
   * Ensure consistency in terminology and style.
6. **Citing Sources**
   * Accurately cite all sources to avoid plagiarism.
   * Use a consistent citation format (e.g., APA, MLA).
7. **Audience Awareness**
   * Tailor content and tone to suit the intended audience.
   * Consider the readers’ level of expertise on the subject.
8. **Timeliness**
   * Adhere to deadlines for submissions and revisions.
   * Ensure information is up-to-date and relevant.
9. **Feedback Incorporation**
   * Seek feedback from peers or mentors before finalizing the report.
   * Be open to constructive criticism and make necessary adjustments.