Final project: A Comparative Analysis of ML Algorithms

Deadline: 20th of December

1 Objective

The objective of this assignment is to analyze the your dataset and develop predictive models using four different machine learning algorithms: K-Nearest Neighbors (KNN), Decision Tree, Naive Bayes, and Support Vector Machine (SVM).You will compare the performance of these algorithms .

Please note that the team consists of four to five members.
Submit your team members from here

2 Dataset

The dataset for this assignment can be independently created using web scraping or downloaded from Kaggle. You are required to ensure that your chosen dataset aligns with the objectives of the analysis. The dataset should include comprehensive information about the idea, covering relevant attributes such as demographic details, categorical classifications, numerical measures, or any other features relevant to your chosen context.

You should also provide a detailed description of your dataset, defining each variable, its purpose, and the type of data it represents. This description will serve as a guide to understanding the dataset's structure and how it can be used for analysis.

3 Tasks

In this section, we outline the specific tasks that need to be performed as part of this assignment. These tasks are designed to guide you through the process of analyzing the Titanic dataset and developing predictive models using various machine learning algorithms.

3.1 Data Exploration and Preprocessing

1. Perform exploratory data analysis (EDA) to gain insights into the dataset. Summarize the key characteristics and distributions of the variables.
2. Handle missing values, outliers, and categorical variables appropriately, and any other data preprocessing steps required to the chosen dataset.

3.2 K-Nearest Neighbors (KNN)

1. Implement the KNN algorithm using a suitable library (*e.g.*, scikit-learn) with a range of *k* values.
2. Train the model on the training set and evaluate its performance using appropriate evaluation metrics (*e.g.*, accuracy, precision, recall, F1-score).
3. Experiment with different distance metrics and discuss their impact on the model's performance.
4. Select the best *k* value based on performance metrics and apply the model to the testing set for predictions.

3.3 Naive Bayes

1. Implement the Naive Bayes algorithm using a suitable library (*e.g.*, scikitlearn).
2. Train the model on the training set and evaluate its performance using appropriate evaluation metrics.
3. Apply the trained model to the testing set for predictions.

3.4 Support Vector Machine (SVM)

1. Implement the SVM algorithm using a suitable library (e.g., scikit-learn) with various hyperparameters.
2. Train the model on the training set and evaluate its performance using appropriate evaluation metrics (e.g., accuracy, precision, recall, F1-score).
3. Experiment with different kernel functions (e.g., linear, polynomial, radial basis function) and regularization parameters to assess their impact on the model's performance.
4. Tune the hyperparameters (e.g., C, gamma) using techniques like grid search or randomized search to optimize the model's performance.
5. Select the best combination of hyperparameters based on performance metrics and apply the tuned model to the testing set for predictions.

3.5. Decision Tree

1. Implement the **Decision Tree algorithm** using a suitable library (e.g., scikit-learn).
2. Train the model on the training set and evaluate its performance using appropriate evaluation metrics (e.g., **accuracy**, **precision**, **recall**, **F1-score**).
3. Experiment with different hyperparameters (e.g., **max depth**, **min samples split**, **criterion**) to analyze their impact on performance.
4. Use techniques like **grid search** or **randomized search** to optimize the model's performance.
5. Visualize the decision tree structure and interpret the splits to provide insights.
6. Apply the best-performing Decision Tree model to the testing set for predictions.

4 Comparative Analysis

1. Compare the performance of the four algorithms (KNN, Naive Bayes, SVM, and Decision Tree) based on the evaluation metrics obtained.
2. Discuss the strengths and weaknesses of each algorithm in the context of your dataset.
3. Provide insights into which algorithm performs better in achieving the prediction objective.

5 Deliverables

Provide a printed Jupyter notebook containing the following information:

1. A **comprehensive report** documenting the analysis, implementation, and evaluation of each algorithm.
2. **Visualizations**, **tables**, **confusion matrices**, and **plots** to support the findings.
3. A **discussion section** summarizing the comparative analysis and the insights gained.
4. Submit your work using the following form before your discussion.

Use Markdown cells to provide explanations and discussions.