COMP3162

Project Title: Predictive Analysis and Insights from Real-World Data

Objective:

The objective of this project is for students to gain practical experience in the entire data science workflow, from data acquisition and preprocessing to model training, evaluation, and interpretation. By working on a real-world dataset, students will learn how to apply data science techniques to derive meaningful insights and make predictions.

Project Overview:

Data Acquisition:

Students are tasked with finding an interesting dataset from sources like Kaggle, UCI Machine Learning Repository, or government databases. The dataset should be relevant to a particular domain of interest (e.g., healthcare, finance, social media).

Data Preprocessing:

Once the dataset is selected, students clean and preprocess the data to handle missing values, outliers, and inconsistencies. This step may involve data cleaning, feature engineering, and data transformation.

Exploratory Data Analysis (EDA):

Students perform exploratory data analysis to gain insights into the dataset's characteristics and relationships between variables. They use visualization techniques such as histograms, scatter plots, and heatmaps to explore patterns and correlations.

Feature Selection and Engineering:

Based on insights gained from EDA, students select relevant features and perform feature engineering to create new features or transform existing ones. This step helps improve model performance by focusing on the most informative features.

Model Selection and Training:

Students choose appropriate machine learning algorithms (e.g., regression, classification, clustering) based on the nature of the problem and the dataset. They split the data into training and testing sets and train various models using techniques such as cross-validation.

Model Evaluation and Tuning:

Students evaluate model performance using appropriate evaluation metrics (e.g., accuracy, precision, recall, F1-score, RMSE). They fine-tune hyperparameters using techniques like grid search or random search to optimize model performance.

Predictive Analysis:

Using the trained models, students make predictions on unseen data or perform clustering to identify patterns or groups within the dataset. They interpret the results and analyze the implications for the problem domain.

Presentation of Findings:

Finally, students prepare a presentation to showcase their findings, including insights gained from EDA, model performance, and predictions. They communicate their results effectively to a non-technical audience, highlighting the significance of their analysis and potential real-world applications.

Instructions

This is a group project that should be done in groups of 4-5. You will be asked to give a presentation.

Deliverables:

- Dataset selection and description.
- Data preprocessing and cleaning scripts.
- Exploratory data analysis report with visualizations.
- Model selection and training scripts.
- Model evaluation metrics and performance analysis.
- Final presentation slides summarizing the project findings and insights.