IT496: Introduction to Data Mining

Course Project - 01

[Deadline: 10th September 2023, Sunday 11:59 PM]

Course Project - 01: Data Preprocessing, EDA and Regression Analysis

Welcome to the course project on Exploratory Data Analysis, Data Preprocessing, and Regression Model Building! This project will provide you with hands-on experience in understanding, cleaning, and preprocessing data, as well as building and fine-tuning regression models. These skills are fundamental in the field of machine learning and data analysis.

Project Duration: 3-4 weeks

Please adhere to the lab policies mentioned in the google classroom

Cite resources and give credit where it's due. If you discuss the questions with your peers, please mention your collaborators in your submission.

Acts of plagiarism will not be tolerated and result in a straight ZERO for this project.

Assignments submitted up to 24 hrs late will be given a 40% penalty. Teams who don't submit their assignment by 11th September 2022, will get ZERO.

There are 3 tasks in this course project, which are discussed in the further section.

In evaluation, your answers for the tasks and the code will be reviewed. Every team member must know their submission's concepts and code. Any member can be asked anything about their role and contributions in submission.

There are 6 datasets, and each dataset is assigned to 5 Teams, as we have total of 30 teams finalized. so you have to perform analysis on your own assigned dataset. No queries would be entertained regarding changes in the dataset.

Submission should have a summary of the dataset and features, EDA findings with visualizations, Explanation of preprocessing steps taken and reasons for each, and a comparison of different models, Submit your file in a well-documented manner. The platform and format for the submission will be shared later.

IT496: Introduction to Data Mining

Course Project - 01

[Deadline: 10th September 2023, Sunday 11:59 PM]

Project tasks

Exploratory Data Analysis (EDA) & Preprocessing:

T1. Explore the dataset assigned to your team and provide:

a. A summary of the dataset (should include information columns present, attribute types, null values, and a summary of each attribute).

b. Data Visualization, summarizing insights about the dataset through EDA.

Regression Analysis:

T2. Identify and list regression problems on your assigned dataset. Which one does seem the most interesting to you and why?

T3. Build an end-to-end Machine Learning pipeline for your assigned dataset for the aforementioned most interesting regression problems found in T2. Your pipeline should include components for dataset preprocessing, transformation, regression model building hyperparameter tuning, grid search or optimization, and evaluation. Report results on the regression models with hyperparameter tuning, and report the best hyperparameter values. Report results using at least two relevant evaluation metrics like RMSE,MAE. Compare results for different models and give the reasoning for that.