Spectral Data

Classifying astronomical objects based on spectrum

And predicting Redshift

**Smrithi (22PD33)**

**Sujan (22PD35)**

Abstract

The project leverages spectral data from the Sloan Digital Sky Survey (SDSS) to classify celestial objects, including stars, quasars, and galaxies, by analyzing their flux values and redshift measurements. The notebook begins with data extraction and transformation, converting raw spectral information into a structured format suitable for machine learning. It includes exploratory data analysis (EDA) steps, such as examining class distributions, redshift patterns, and other spectral characteristics to gain insights into the data.

Through visualizations and statistical summaries, the analysis highlights distinctive spectral features associated with each class of astronomical objects. The notebook also establishes a foundation for machine learning model development by preparing a labelled dataset that can be used for future classification tasks. This project contributes to automated spectral classification efforts, which are vital for efficient processing and analysis of large astronomical datasets.

The Notebook focused on developing and evaluating a machine learning model for regression using various tools, including the astroML package for specialized machine learning functionalities. It begins with data loading, exploration, and visualization to gain insights into the dataset's structure. Preprocessing steps follow, addressing missing values, encoding categorical variables, and scaling features to ensure data quality. The Synthetic Minority Over-sampling Technique (SMOTE) is applied to handle class imbalance, enhancing model performance on underrepresented classes. Multiple regression models, including Linear Regression and Decision Tree Regressor, are implemented with performance optimized through cross-validation. Model evaluation metrics, such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), allow for model comparison, aiding in the identification of the most effective model for accurate predictions.