Implementation of Machine Learning to Identify Deep Space Objects

*By Saarthak Gupta*

*Under the Guidance of Dr. Sarvjeet Herald*

**Abstract**

*The purpose of this study is to test the accuracy and effectiveness of different machine learning algorithms to identify stars and galaxies in a cosmological dataset based on 14 unique features.* *The dataset used for this project is called Sloan Digital Sky Survey DR14. The dataset contains 10,000 total items. However, for the purpose of this program, the data pertaining to Quasars was removed. This left us with 4998 Galaxies and 4152 stars.*

*The dataset has 14 features for each object. (objid, specobjid, rerun and classes were dropped as they do not help us in predicting the type of the object). The dataset has 5 features, which represent different wavelengths of light. (u, g, r, i and z). Redshift is the displacement of the spectrum of an astronomical object toward longer (red) wavelengths. In astronomy, declination (abbreviated dec; symbol δ) is one of the two angles that locate a point on the celestial sphere in the equatorial coordinate system, the other being hour angle. Right ascension (ra) is the angular distance of a particular point measured eastward along the celestial equator from the Sun at the March equinox to the point above the earth in question. Run, rerun, camcol and field are features which describe a field within an image taken by the SDSS. A field is basically a part of the entire image corresponding to 2048 by 1489 pixels. A field can be identified by: run number, which identifies the specific scan, the camera column, or "camcol," a number from 1 to 6, identifying the scanline within the run, and the field number. The field number typically starts at 11 and can be as large as 800 for particularly long runs. An additional number, rerun, specifies how the image was processed. Each spectroscopic exposure employs a large, thin, circular metal plate that positions optical fibers via holes drilled at the locations of the images in the telescope focal plane. Modified Julian Date (mjd), used to indicate the date that a given piece of SDSS data (image or spectrum) was taken. The SDSS spectrograph uses optical fibers to direct the light at the focal plane from individual objects to the slithead. Each object is assigned a corresponding fiberID. The algorithms tested were:*

*1. Decision Tree Classifier: In computer science, Decision tree learning uses a decision tree to go from observations about an item to conclusions about the item's target value. It is one of the predictive modeling approaches used in statistics, data mining and machine learning.*

*2. Logistic Regression: Logistic regression is a classification algorithm used to assign observations to a discrete set of classes.*

*3. Naïve Bayes: In machine learning, Naïve Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.*

*4. Support Vector Machines: The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N = the number of features) that distinctly classifies the data points.*

*5. Random Forest: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees.*

*Table

Description automatically generated*