Classification of Cosmic Objects

Using Supervised Machine Learning

Venkata Harshith Nikhil Samudrala
Data Science
University of Colorado Boulder
Boulder, CO, USA
vesa8976@colorado.edu

Sai Vamshi Challamalla
Data Science
University of Colorado Boulder
Boulder,CO,USA
sach4260@colorado.edu

ABSTRACT

The human kind has always been fascinated with the sky and what lies beyond it since time eternity. The technological advancements has enabled us to peek and learn immensely from these cosmic objects. This study aims to employ Supervised Machine Learning to help classify different cosmic objects. The study makes use of the spectral characteristics. The data obtained from Solan Foundation's SSDS and utilizing the data trying to aligning it with astronomical research trend. This study is a detailed study on the characteristics and how Machine Learning can play a pivotal role in classifying the stellar objects. This study not only aims to assist astronomical research but also provide with invaluable insights prompting towards a future collaboration of Data Analysis and Machine Learning and fields like Astronomy.

1 Introduction

The main goal behind the research is to determine the best possible features that will lead to high accuracy in identifying stellar objects based on the spectral data obtained from a telescope. The data from a telescope is very complicated and filled with numbers making it a tedious task to classify them. This research makes use of the Sloan Sky Survey's SSDS Digital Release 17 and Digital Release 14 data in training machine learning models to achieve a better accuracies. There have been many researches into this domain dealing the issue in different ways. This research too aims to classify a few of the stellar objects mainly Stars, Galaxies, Quasars. This research will provide a great help to the astronomers.

Stars are large luminous objects in the universe mainly made up largely of helium and hydrogen emitting light and electro-magnetic radiations. They can be identified due to their large size, luminosity. [1]

Galaxies are huge cosmic entities that form as the fundamental blocks of the universe. They consist of stars, star clusters , black holes etc. They are found in different shapes.[1]

Quasars are highly luminous objects that sometimes outshine the galaxies with high gravity containing super massive black holes in the center.[1]
The classification of these object will help in identifying anomalies, patterns and other phenomenon about the objects. The research aims to use the astronomical data fusing it with modern data analysis and machine learning techniques to contribute to the new bridge between computer science and astronomy.

2 Objective

The main objective of this research is to determine the most important features for exploratory data analysis and training machine learning models to get maximum accuracy for classifying the cosmic objects. This research utilizes the diverse dataset and analyses each feature by analyzing the distribution and the outliers. Performing univariate analysis and other feature determination techniques to make the data ready for model training. The data is trained with traditional machine learning models like Logistic Regression, Support Vector Classifiers to advanced classifiers like Bagging Classifier, KNearest, MLP Classifiers, Gradient Boosting classifiers etc.

3 Data Description

The dataset for this study was accumulated from the Sloan Sky Survey's DR17 and DR14 releases. The data is a rich repository comprising essential columns that illustrate the diverse spectral properties of different cosmic objects. The columns consist a comprehensive array of information information for classifying the distinct characteristics of stars, galaxies, and quasars.[1]

The important columns include 'ra' and 'dec,' denoting the right ascension and declination, respectively, providing the celestial coordinates of the observed objects. The 'run' and 'field' columns offer insights into the observational run and field of view. The 'camcol' and 'field' contribute to the imaging process, identifying the camera column and field within the observed region.

The 'objid' column serves as a unique identifier for each object, facilitating data management and traceability. The 'fiberid' corresponds to the spectroscopic fiber's identification, crucial for extracting spectral data. The 'class' column, representing the classification of cosmic objects, is the target variable for our machine learning classification models.

These columns collectively constitute a detailed framework for understanding the spectral dimensions of the cosmic entities.

Fig 3.1 Sloan Digital Sky Survey (SDSS) Organization

4 Techniques Used

4.1 Class Imbalance

An unequal representation of classes results from class imbalance, which is the result of a dataset's instances being substantially unbalanced among distinct classes. Class imbalance may occur because some cosmic objects are inherently rare in the framework of our study effort, which classifies stars, galaxies, and quasars based on spectral features. Quasars, for instance, might cause a discrepancy in the distribution of classes since they are less common than stars or galaxies. Machine learning models may experience bias towards the dominant class as a result of this imbalance, which could affect their capacity to correctly categorize the minority class. Assuring the model's ability to identify the distinct spectral signatures of every cosmic entity is dependent on mitigating class imbalance, which helps create a more reliable classification.[4][5]

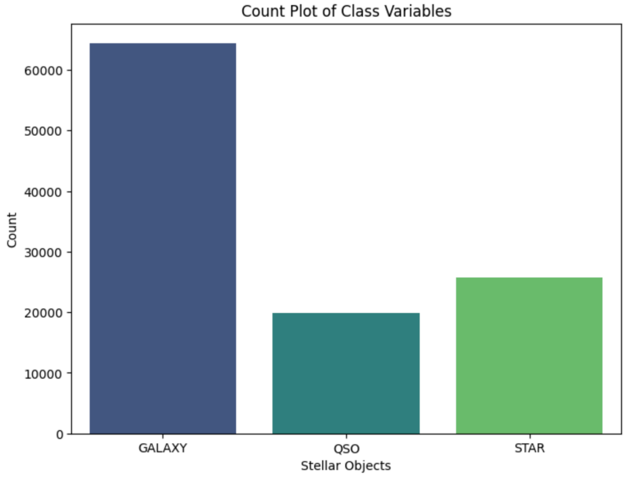


Fig 4.1.1 Class Imbalance in Data

4.2 Over-Sampling

One method used to rectify class imbalance in a dataset—especially in cases where some classes are underrepresented—is oversampling. Oversampling refers to the process of intentionally increasing the number of cases in the minority class or classes in order to obtain a more equal distribution in the context of our research effort that involves identifying stars, galaxies, and quasars. In order to provide a more fair representation of each class for the machine learning model to train on, this is usually accomplished by copying or creating manufactured instances of the minority class. Oversampling helps the model identify and categorize the rarer cosmic objects (like quasars) more accurately by reducing the effect of class imbalance, which leads to a more reliable and objective classification result. Synthetic Minority Over-sampling Technique (SMOTE), random duplication, and other common oversampling techniques.[4][5]

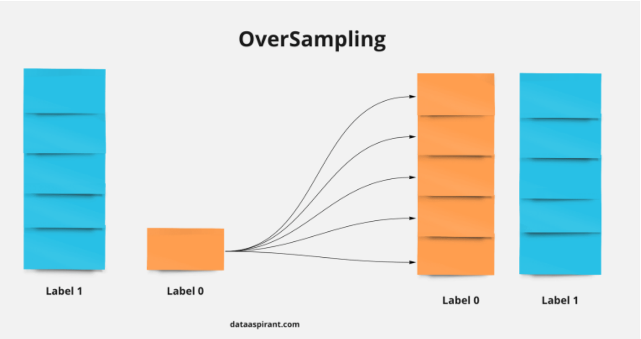


Fig 4.2.1 Over-Sampling Mechanism

4.3 Under-Sampling

One tactic used to address class imbalance in a dataset—especially when some classes are overrepresented—is under sampling. Under sampling in the context of our study project involves lowering the number of instances in the majority class or classes in order to attain a more balanced class distribution for stars, galaxies, and quasars. For the purpose of aligning the dataset with a fairer representation of each class during model training, instances from the majority class are removed at random in this phase. Under sampling ensures a fair evaluation of the spectral features of all cosmic objects by preventing the machine learning model from becoming too biased toward the dominant class. Although under sampling corrects for class imbalance, it may result in data loss from the dominant class and force decision-makers to choose between other re sampling techniques.

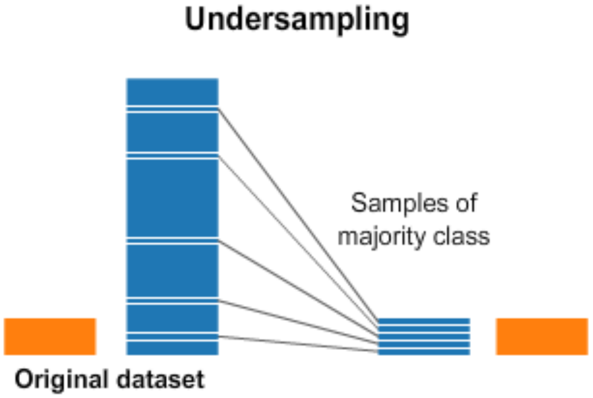


Fig 4.3.1 Under-Sampling Mechanism

4.4 Feature Analysis

In our research project, feature analysis is an essential part of data exploration and preprocessing. It entails a thorough review of the dataset's features in order to derive significant insights and improve the performance of machine learning models. Feature analysis involves evaluating the statistical qualities, distributions, and correlations between various aspects so that stars, galaxies, and quasars can be categorized according to their spectral attributes. In order to determine the distribution of features across different classes and to weed out strongly correlated or unnecessary features, we use techniques like box plots and violin plots. The most important properties are identified with the use of correlation analysis, which guarantees that the data used to train our models is pertinent.[3]

The method of feature engineering makes use of advanced domain knowledge in astronomy to identify trends, aberrations, and possible predictors for the classification of cosmic objects. Our objective is to enhance the dataset by utilizing feature analysis, with a focus on feature quality and relevance. This will help to maximize model performance and further our larger objective of unraveling the cosmic narrative concealed in the spectral data.

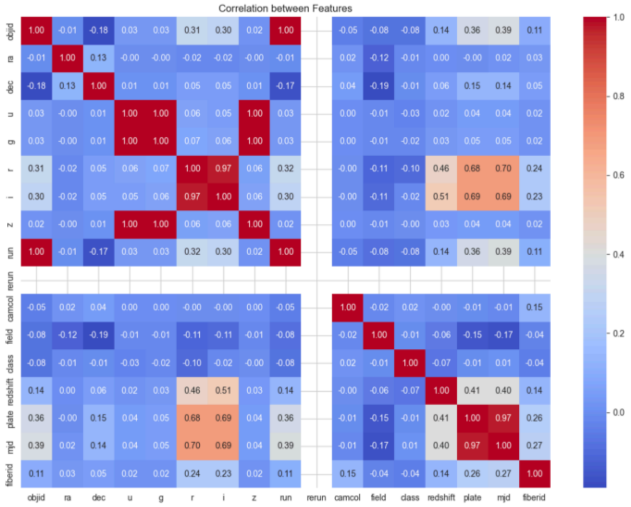
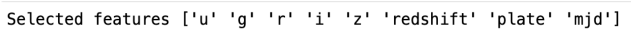


Fig 4.4.1 Correlation between Features

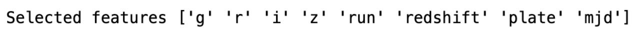
4.5 Univariate Analysis

This research attempts, which focuses on identifying stars, galaxies, and quasars based on spectral features, requires the use of univariate analysis. Using this analytical technique, each variable in the dataset is examined separately in order to determine its distribution, statistical characteristics, and importance with regard to the classification of cosmic objects. In univariate analysis, we visualize the distribution of each feature across many classes using statistical tools like kernel density estimation and histograms. This makes it easier to find trends, variances, and possible outliers among distinct features. To visualize the distribution of the data and get insight into the existence of extreme values and central patterns, we employ box plots. Through feature-by-feature dissection, univariate analysis informs feature selection, engineering, and class imbalance management decisions, among other research considerations. It offers a detailed comprehension of every feature's input to the classification problem, enabling a more focused and knowledgeable method of developing precise machine learning models for the classification of cosmic objects.



```
Selected features ['u' 'g' 'r' 'i' 'z' 'redshift' 'plate' 'mjd']
```

Fig 4.5.1 Univariate Analysis results with f_classif function



```
Selected features ['g' 'r' 'i' 'z' 'run' 'redshift' 'plate' 'mjd']
```

Fig 4.5.1 Univariate Analysis results with mutual_info_classif function

4.6 Histplot

In this study we use the histplot function from the Seaborn library to perform univariate analysis. Specifically, we create histograms, for each feature. These histograms visually display the distribution of values within each feature allowing us to better understand the tendencies and variations in the data. To accomplish this, we iterate through each column in the dataset using a loop and generate histograms for exploration. This technique is particularly helpful in comprehending how stars, galaxies and quasars can be classified based on features. By providing a snapshot of distribution these histograms help us identify potential patterns or variations within each feature. This initial univariate analysis using histplot serves as a starting point for steps such as feature engineering, selection and addressing class imbalance. Ultimately it allows for an informed approach, to classifying cosmic objects. [6]

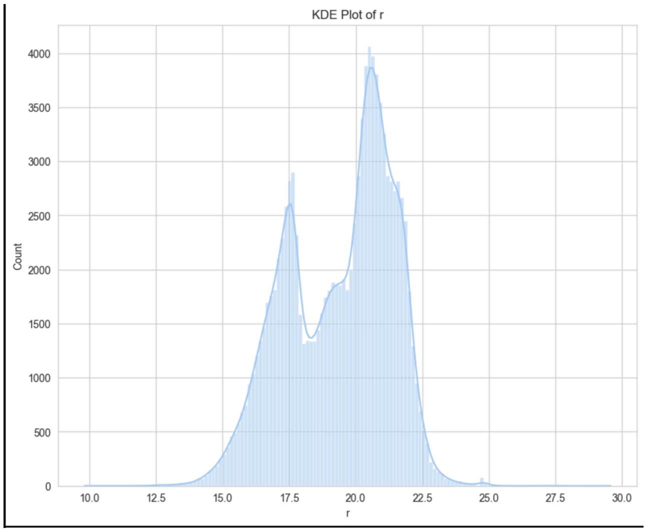


Fig 4.6.1 Histplot of r column

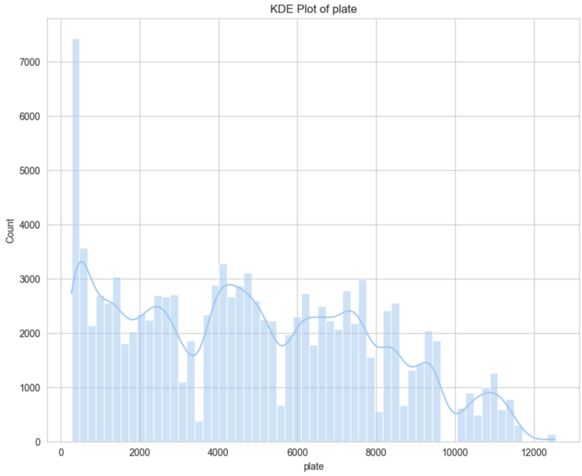


Fig 4.6.2 Histplot of plate column

4.7 Distplot

To visualize the distribution of features we utilized the distplot function from the Seaborn library. This handy tool generates a histogram. Overlays a kernel density estimate giving us a view of how the data is spread out for each feature. In our code we created distribution plots for each column in the dataset. By doing this we can easily assess the tendency spread and possible outliers across classes of cosmic objects. This analysis, with distplot helps us identify any variations in feature distributions among entities. These insights are crucial, for analysis and feature engineering. This visualization method improves the analysis of features by providing a comprehensive view of their characteristics. It helps us better understand the dataset and make decisions during the data preprocessing stage of the object classification project.[3][4]

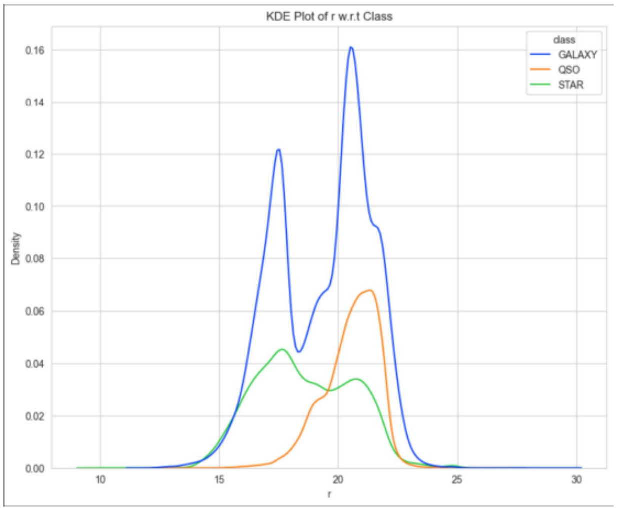


Fig 4.7.1 Distplot

5 Methodology

5.1 Random Under Sampler for Under sampling

We utilized the RandomUnderSampler function from the imbalanced learn library to handle the issue of class imbalance in our dataset. This method involves eliminating instances, from the majority class, which helps equalize the distribution of classes and prevents any bias towards the class during model training. By applying under sampling we were able to achieve a representation of cosmic object classes. This enhancement greatly improved our models ability to effectively learn from all classes. Employing this approach was vital in addressing class imbalance. Ensuring an evaluation of spectral features for precise classification of stars, galaxies and quasars, in our research project. Before using this RandomUnderSampler method the count of Galaxy is more than 60,000 whereas Stellar Objects and Star counts are 20,000 and 28,000 respectively. After performing this method, the count reduced to 20,000 for each class, which balances all the classes and enhances the model ability to learn from all classes effectively.

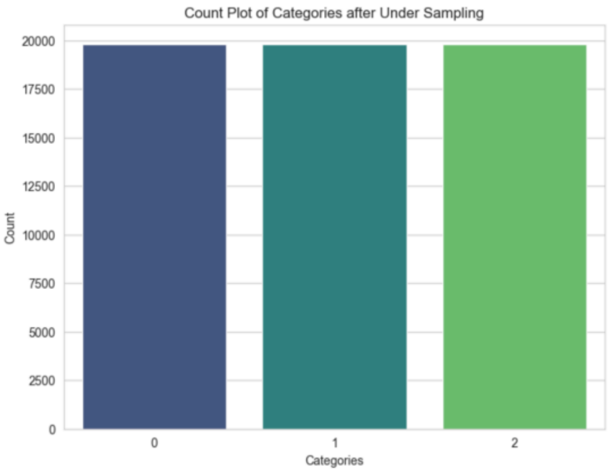


Fig 5.1.1 Count Plot of Categories after Under Sampling

5.2 Feature Analysis through Univariate Analysis and Correlation

We started by analyzing the features and creating a correlation matrix to understand the relationships, between them. To make it easier to interpret we visualized the matrix using a heat map. By examining each features correlation with the target variable 'class' we discovered which features have the influence on the classification task.[5]

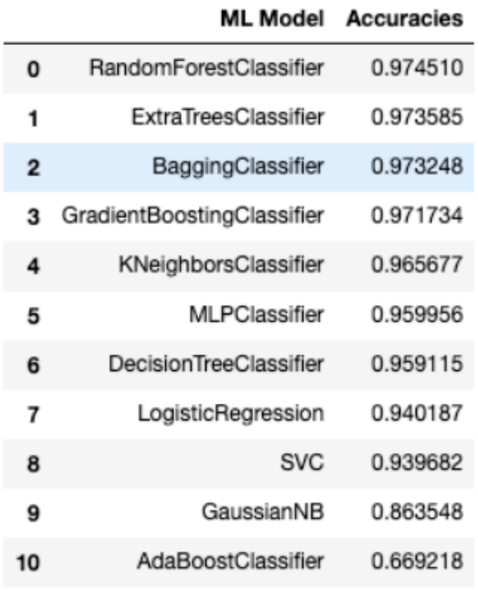
Next we prepared the data for machine learning by applying transformation and addressing any missing values. To choose the features we used the GenericUnivariateSelect method with f_classif and mutual_info_classif score functions selecting the top 7 based on their significance, in predicting the target variable.

Once selected these features were used for both training and testing datasets. This refining process ensures that our dataset is well suited for developing machine learning models that can classify objects based on their spectral properties. The best features that were selected based on the feature analysis were ‘u’, ‘g’,’r’, ‘i’, ‘z’, ‘redshift’, ‘plate’.

5.3 Model Testing

During the evaluation phase of our model we examined different machine learning algorithms to see how well they could classify objects based on their properties. We implemented a range of classifiers, such, as Logistic Regression, Support Vector Machine (SVC) Random Forest, KNeighbors, GaussianNB, DecisionTree,AdaBoost, GradientBoosting, ExtraTrees, Bagging and MLP. Each algorithm was organized within a pipeline to ensure execution. After training the models on a dataset we tested them on a validation set. Calculated their accuracies. The Random Forest Classifier stood out as the performing model with an accuracy of 97.45% compared to other algorithms that were evaluated.

This thorough evaluation offers insights, into how different machine learning models perform for our object classification task. The results help us choose the algorithms for further optimization and fine tuning to enhance our classification model in subsequent project stages.

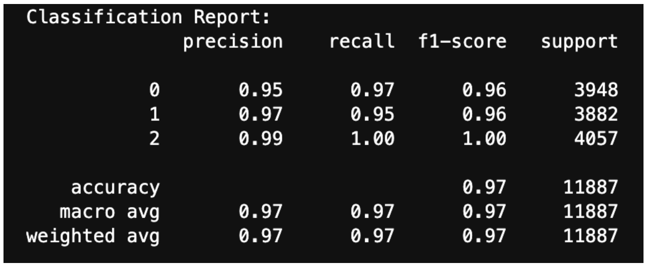


	ML Model	Accuracies
0	RandomForestClassifier	0.974510
1	ExtraTreesClassifier	0.973585
2	BaggingClassifier	0.973248
3	GradientBoostingClassifier	0.971734
4	KNeighborsClassifier	0.965677
5	MLPClassifier	0.959956
6	DecisionTreeClassifier	0.959115
7	LogisticRegression	0.940187
8	SVC	0.939682
9	GaussianNB	0.863548
10	AdaBoostClassifier	0.669218

Fig 5.3.1 ML Models Accuracies

5.4 Increasing the performance

We utilized GridSearchCV to tune the parameters of the Random Forest Classifier for classification of cosmic objects. We defined a parameter grid that explores combinations of 'n_estimators' 'max_depth,' 'min_samples_split,' and 'min_samples_leaf.' By running a 5 cross validation on the training data we exhaustively searched for the parameter set that maximizes model performance. From the results of the grid search we extracted the parameters and their corresponding cross validated score. This provides insights, into the configuration that yielded results during training. Next we evaluated the identified Random Forest model on the test set to assess its generalization performance. This process of hyperparameter tuning ensures that our Random Forest model is calibrated specifically to suit the characteristics of our cosmic object dataset. The reported results, such, as the parameter set and an impressive test accuracy of 97.51% offer guidance for selecting and deploying our model in subsequent phases of our project.



```
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.97      0.96      3948
           1       0.97      0.95      0.96      3882
           2       0.99      1.00      1.00      4057

    accuracy                           0.97     11887
   macro avg       0.97      0.97      0.97     11887
weighted avg       0.97      0.97      0.97     11887
```

Fig 5.4.1 Classification Report of Grid Search

6 Conclusion & Findings

After training the models through the pipeline, the top 5 performing models were Random Forest Classifier, Extra Trees Classifier, Bagging Classifier, Gradient Boosting Classifier, K Nearest Classifier. The accuracy achieved with the test data set for Random forest was 97.45% and for the Extra Trees Classifer is 97.36%.

The Random Forest Classifer was further tuned using the GridSearch CV, to achieve an accuracy of 97.5%.

7 Limitations

There are a few limitations in this research. The original data is slightly skewed toward one class, and the features are all have certain importance. The machine learning model analysis can be done for a different combination of features and check the accuracy. Based on the results from the bivariate analysis and the correlation only certain top features were selected.

The features play a vital role in determining the model training and accuracy. The datasets at hand are from different periods in time, this caused a uneven distribution overall. Although the features were scaled and then the model was 13 trained, having a data that meets all expectations will probably improve the performances of the model.

Deep Learning methods like Artificial Neural Networks (ANNs), Convolutional Neural Networks (CNNs) may also fetch results but the hardware and software limitation at the researching end made it very hard to test. This take a lot of time and computational results. Having better resources can be helpful in testing the model with hundreds of thousands of records that the two hundred thousand records which were used in the research.[8]

8 Future Scope

Computer Science and Data Analysis has always played an important role in astronomy and space research domains. The classification of stellar objects in particular is very growing aspect in the astronomical research, many different perspective of research are already being done by organizations like NASA etc. There are many ways to approach this research like incorporating telescopic images and applying computer vision based techniques to train better machine learning/ deep learning models. There can be long shot applications like identification of life sustainable planets.[]

References

1. D. Omat, J. Otey and A. Al-Mousa, "Stellar Objects Classification Using Supervised Machine Learning Techniques," 2022 International Arab Conference on Information Technology (ACIT), Abu Dhabi, United Arab Emirates, 2022, pp. 1-8, doi: 10.1109/ACIT57182.2022.9994215.

2. Forbes, D. A., & Kroupa, P. (2011). What is a galaxy? Cast your vote here. Publications of the Astronomical Society of Australia, 28(1), 77-82.

3. York, D. G., Adelman, J., Anderson Jr, J. E., Anderson, S. F., Annis, J., Bahcall, N. A., ... & Yasuda, N. (2000). The sloan digital sky survey: Technical summary. The Astronomical Journal, 120(3), 1579.

4. Bhamare, A. R., Baral, A., & Agarwal, S. (2021, June). Analysis of kepler objects of interest using machine learning for exoplanet identification. In 2021 International Conference on Intelligent Technologies (CONIT) (pp. 1-8). IEEE.

5. Petrusevich, D. A. (2020, May). Implementation of machine learning algorithms in the Sloan Digital Sky Survey DR14 analysis. In IOP conference series: materials science and engineering (Vol. 862, No. 4, p. 042005). IOP Publishing.

6. Chuntama, T., Techa-Angkoon, P., Suwannajak, C., Panyangam, B., & Tanakul, N. (2020, December). Multiclass classification of astronomical objects in the galaxy m81 using machine learning techniques. In 2020 24th International Computer Science and Engineering Conference (ICSEC) (pp. 1-6). IEEE.

7. Rony, M. A. T., Reza, D. A., Mostafa, R., & Ullah, M. A. (2021, September). Application of Machine Learning to Interpret Predictability of Different Models: Approach to Classification for SDSS Sources. In 2021 International Conference on Electronics, Communications and Information Technology (ICECIT) (pp. 1-4). IEEE.

8. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

Links

Presentation:
https://www.youtube.com/watch?v=JwmINUfVm28