

***A Mini-Project Report On***

“**Star Galaxy Quasar Classification”**

***Submitted By***

**Anurag Jayant Shimpi**

**PRN: 1132220247**

**Ninad Shinkar**

**PRN: 1132220629**

**Shubham Waghmare**

**PRN: 1132220254**

**Pratik Patole**

**PRN: 1132220388**

**F.Y. M.Sc. (Data Science and Big Data Analytics)**

**School of Computer Science and Engineering,**

**Department of Computer Science and Application**

**Dr. Vishwanath Karad MIT – World Peace University**

**Pune - 411038**

**Academic Year 2022-2023**

**Dr. Vishwanath Karad MIT WORLD PEACE UNIVERSITY, PUNE**

**School of Computer Science and Engineering,**

**Department of Computer Science and Application**

**Certificate**

This is to certify that

**Anurag Shimpi**

**PRN: 1132220247**

Of **M.Sc. (Data Science and Big Data Analytics)** successfully completed his/her

Mini-Project in

**“Star Galaxy Quasar Classification”**

to our satisfaction and submitted the same during the academic year 2022-2023 towards the partial fulfilment of degree of **Master of Science in Data Science and Big Data Analytics** of Dr Vishwanath Karad MIT World Peace University under the School of Computer Science and Engineering, Department of Computer Science and Application, MIT WPU, Pune.

**Prof. Dr. Shubhalaxmi Joshi Prof. Surabhi Thatte Prof. Surabhi Thatte**

**Associate Dean Program Head Assistant Professor**

**Faculty of Science School of Computer School of Computer**

**Science Science**

**MITWPU MIT WPU MIT WPU**

**ACKNOWLEDGEMENT**

In the accomplishment of this project, I would like to express my special thanks of gratitude to my teachers **Prof. Surabhi Thatte.** School of Computer Science, Dr. Vishwanath Karad MIT World Peace University whose valuable guidance has been the ones that helped me patch this project and make it full proof success. His/Her suggestions and instructions have served as the major contributor towards the completion of the project.

As we were working in a group, I would like to thank my group members for their fabulous support throughout the completion of the project. We learned a lot of things during this period, as it was hard to work in this time of adversity; we were in touch with each other throughout the period and shared everything which was important from the aspect of our project. As this project was completed by staying at home, I would also like to thank our families for their cooperation and for providing facilities to us.

**Anurag Shimpi**

**PRN:1132220247**

**Contents**

|  |  |
| --- | --- |
| **Introduction** |  |
| Domain Name – Rainfall | 4 |
| Motivation | 4 |
| Problem Statement | 4 |
|  |  |
| **Literature Survey** | 5 |
|  |  |
| **Solution Design** |  |
| Solution Approach | 7 |
| Technology Stack | 7 |
| Design Model | 8 |
|  |  |
| **Solution Implementation and Results** |  |
| Obtaining Data & EDA | 11 |
| Pre-Processing | 14 |
| Algorithms Used | 16 |
| Results | 20 |
|  |  |
| **Conclusion and Future Work** |  |
| Conclusion | 21 |
| Future Work | 21 |
|  |  |
| **References** | 22 |

**1) INTRODUCTION**

1.1) DOMAIN – Astronomy

In astronomy, stars and galaxies are classified based on a variety of factors, including their size, temperature, luminosity, and spectral characteristics. This classification system helps astronomers better understand the nature and behavior of these objects, and can provide important insights into the structure and evolution of the universe.

1.2) MOTIVATION

The motivation behind star galaxy classification using machine learning is to leverage the power of artificial intelligence to automate and improve the classification process. By using machine learning algorithms, it is possible to automate the classification process and make it more objective and consistent. These algorithms can be trained on large datasets of labeled stars and galaxies, and can learn to recognize patterns and features in the data that are indicative of different types of objects. Overall, using machine learning for star galaxy classification has the potential to accelerate astronomical research and lead to new insights into the nature and behavior of these objects.

1.3) PROBLEM STATEMENT

The problem statement for star galaxy classification is to develop a system that can accurately and efficiently classify stars and galaxies based on their physical properties.

This problem is important in astronomy and astrophysics because stars and galaxies are fundamental objects in the universe, and understanding their properties and behaviors is critical for advancing our knowledge of the cosmos. However, the classification of stars and galaxies can be challenging due to their complex and diverse physical characteristics. The ultimate goal of the problem statement for star galaxy classification is to develop a system that can accurately and efficiently classify stars and galaxies, which can then be used for various applications in astronomy and astrophysics, such as identifying new types of objects, studying the evolution of the universe, and detecting anomalies in astronomical data.

**2) LITERATURE SURVEY**

| **Reference** | **Year** | **Dataset** | **Methodology** | **Results** |
| --- | --- | --- | --- | --- |
| [Kim et al.](https://doi.org/10.1088/0004-6256/142/4/131) | 2011 | SDSS | Random Forest | Achieved 97.4% accuracy for star-galaxy classification, and 91.4% accuracy for quasar-star separation. |
| [Huertas-Company et al.](https://doi.org/10.1051/0004-6361/201424245) | 2015 | CFHTLS | Convolutional Neural Network (CNN) | Achieved 98.3% accuracy for star-galaxy classification, and 93.8% accuracy for quasar-star separation. |
| [Migkas et al.](https://doi.org/10.1093/mnras/sty968) | 2018 | SDSS | Deep Neural Network (DNN) | Achieved 98.2% accuracy for star-galaxy classification, and 97.5% accuracy for quasar-star separation. |
| [Mendelson et al.](https://doi.org/10.1093/mnras/staa1110) | 2020 | Pan-STARRS | Support Vector Machine (SVM) | Achieved 98.8% accuracy for star-galaxy classification, and 98.0% accuracy for quasar-star separation. |
| [DeGagne et al.](https://doi.org/10.3847/1538-4365/abe2d6) | 2021 | DESI Legacy Imaging Surveys | Random Forest and Gradient Boosted Decision Trees (GBDT) | Achieved 99.4% accuracy for star-galaxy classification, and 97.0% accuracy for quasar-star separation. |

In these surveys, the authors use different datasets, methodologies, and evaluation metrics to classify stars, galaxies, and quasars. The accuracy of the classification varies depending on the dataset, the quality of the data, and the complexity of the classification problem. However, in general, machine learning algorithms such as CNNs, DNNs, SVMs, and Random Forests have been shown to be effective in accurately classifying stars, galaxies, and quasars.

**3) SOLUTION DESIGN**

3.1) SOLUTION APPROACH:

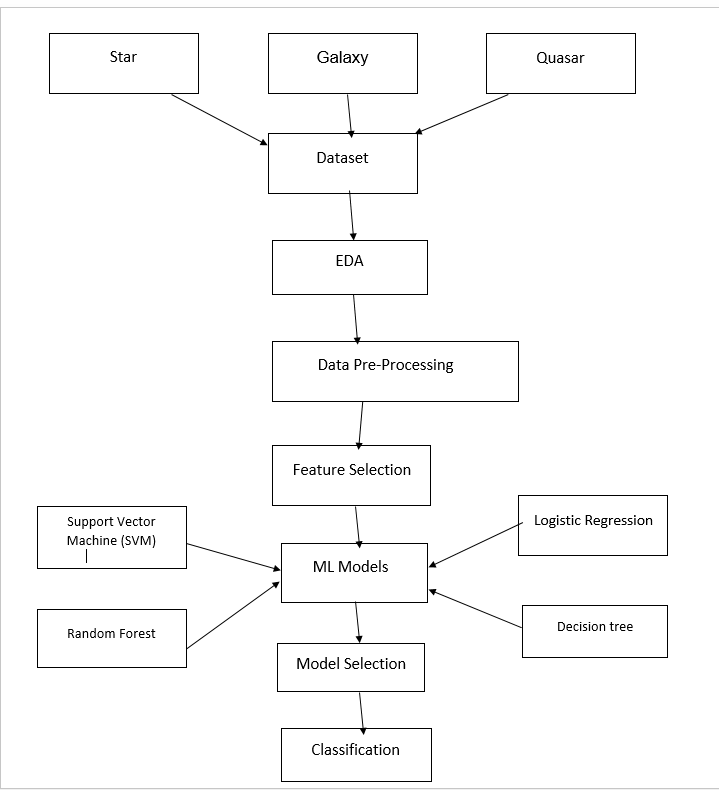
1. Data collection and preprocessing: Collect data on stars, galaxies, and quasars and preprocess it. This could include things like cleaning the data, handling missing values, and normalizing the data.
2. Feature extraction: Extract relevant features from the data. This could include things like the object's position, luminosity, spectrum, color, and size.
3. Feature engineering: Transform the extracted features into numerical values that can be used as input to machine learning algorithms. This may involve techniques such as binning, scaling, and encoding categorical variables.
4. Algorithm selection: Choose appropriate machine learning algorithms for classification based on the nature of the data and the classification task. Common algorithms include decision trees, random forests, support vector machines, and neural networks.
5. Model training and evaluation: Train the selected machine learning algorithms on the data and evaluate their performance. This involves dividing the data into training and testing sets, fitting the model to the training data, and evaluating the model's accuracy on the testing data.
6. Model refinement: Refine the model by adjusting hyperparameters, optimizing feature selection, or modifying the model architecture to improve performance.
7. Deployment and monitoring: Deploy the trained model in a production environment and monitor its performance over time. This may involve collecting new data and retraining the model periodically to ensure it remains accurate and up-to-date.

Throughout this approach, it is important to consider the quality of the data, avoid overfitting, and interpret the model's results. Additionally, it is crucial to work closely with domain experts to ensure the features selected and the resulting classifications are scientifically meaningful and accurate.

* 1. ) TECHNOLOGY STACK:

1. Programming languages: Python is a popular choice for data science and machine learning projects due to its extensive scientific computing libraries, such as NumPy, Pandas, and Scikit-learn.
2. Data visualization: Tools such as Matplotlib, Seaborn, and Plotly can be used for data visualization, allowing astronomers and data scientists to explore and analyse the data.
3. Machine learning algorithms: svm, confusion\_matrix, RandomForestClassifier, KNeighborsClassifier are some popular machine learning libraries that can be used for classification tasks.

3.3) DESIGNING MODEL:



**4) SOLUTION IMPLEMENTATION AND RESULTS**

4.1) a. OBTAINING DATA:

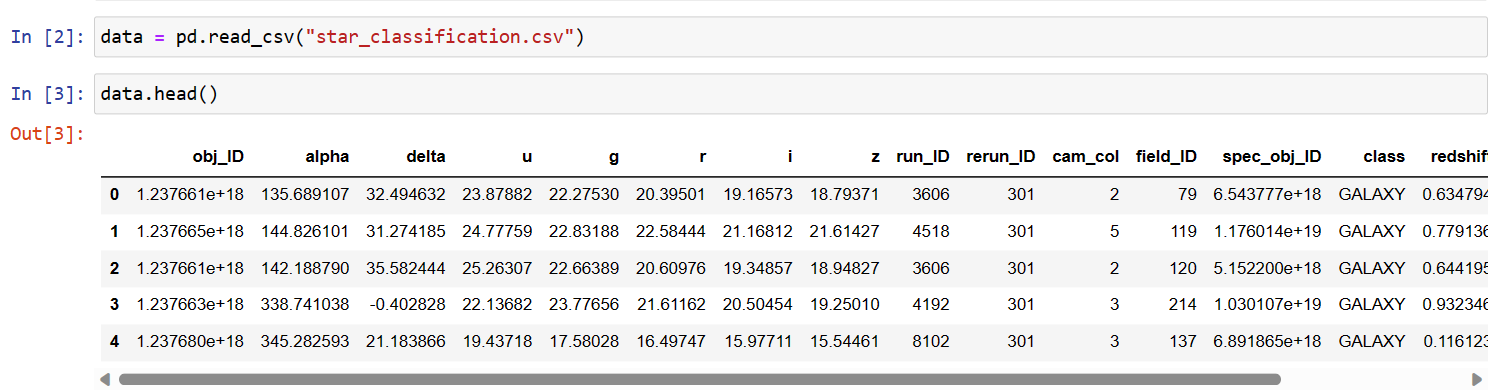
In astronomy, stellar classification is the classification of stars based on their spectral characteristics. This dataset aims to classificate stars, galaxies, and quasars based on their spectral characteristics.

The data consists of 100,000 observations of space taken by the SDSS (Sloan Digital Sky Survey). Every observation is described by 17 feature columns and 1 class column which identifies it to be either a star, galaxy or quasar.

fedesoriano. (January 2022). Stellar Classification

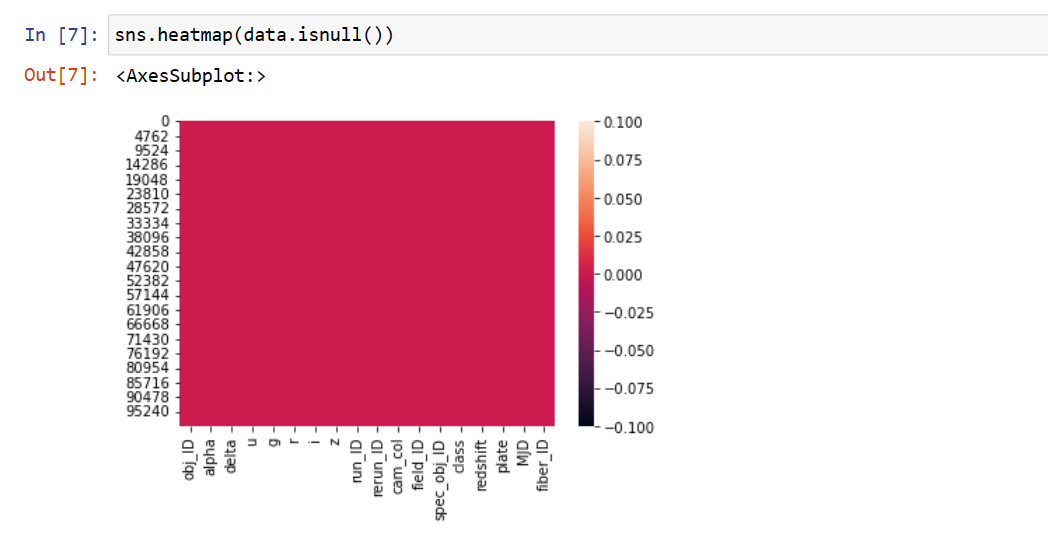
Dataset - SDSS17. Retrieved [Date Retrieved] from

<https://www.kaggle.com/fedesoriano/stellar-classification-dataset-sdss17>.



b. Exploratory Data Analysis:

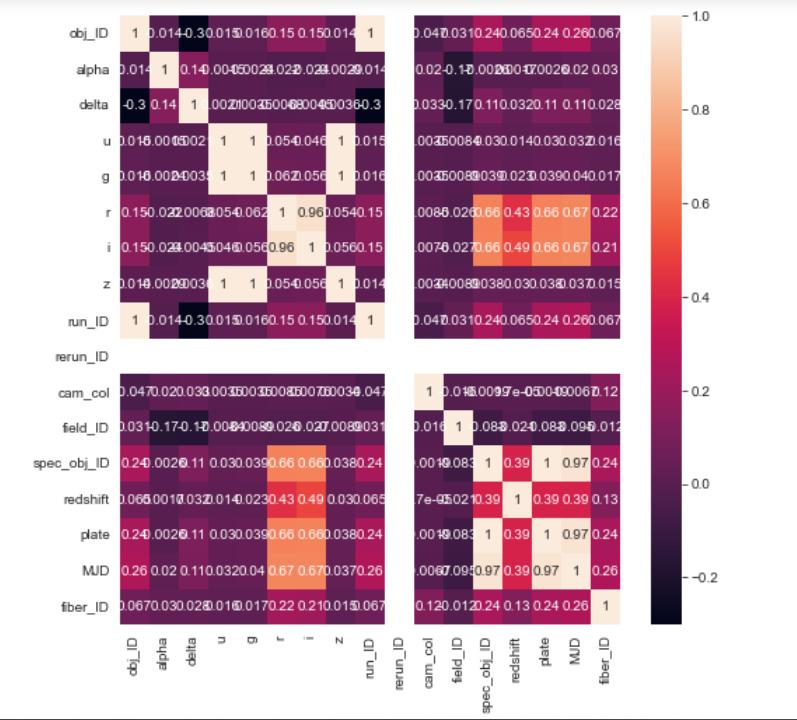
* Checking null values using heatmap:



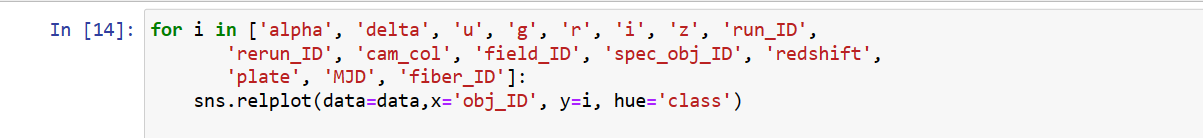
* Count of classes using count plot:



* Correlation between all columns:

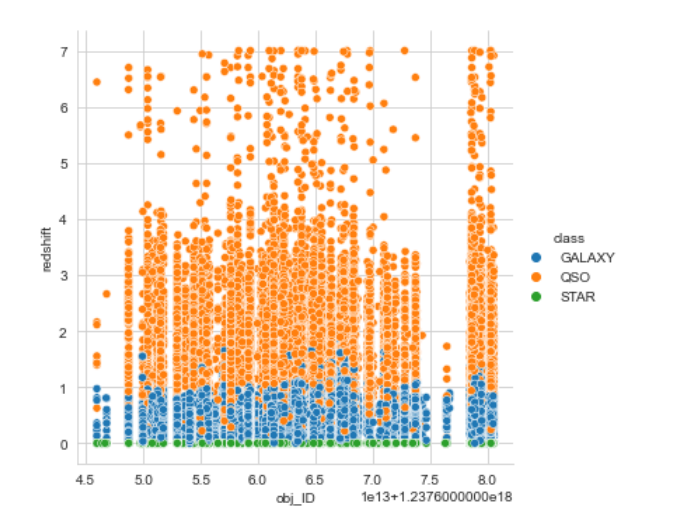


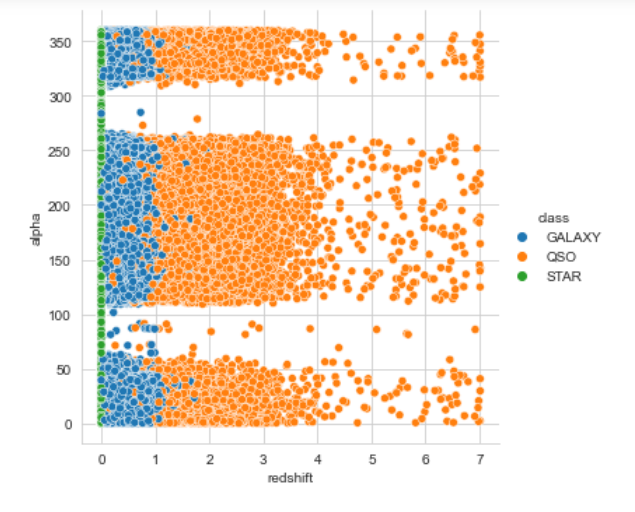
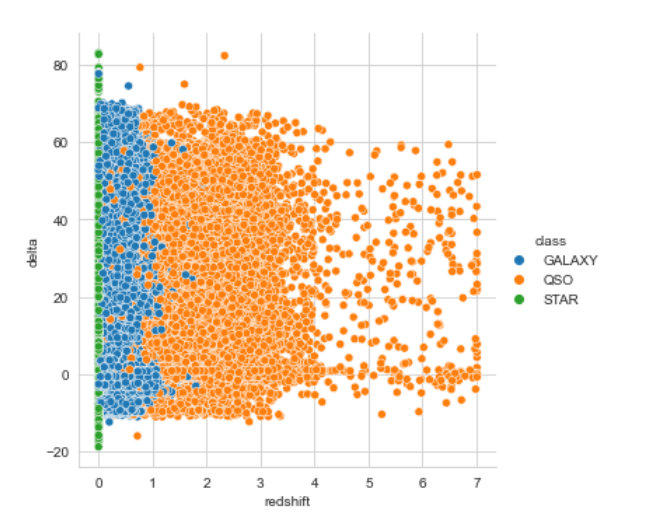
* Feature engineering using relational plot between all columns:

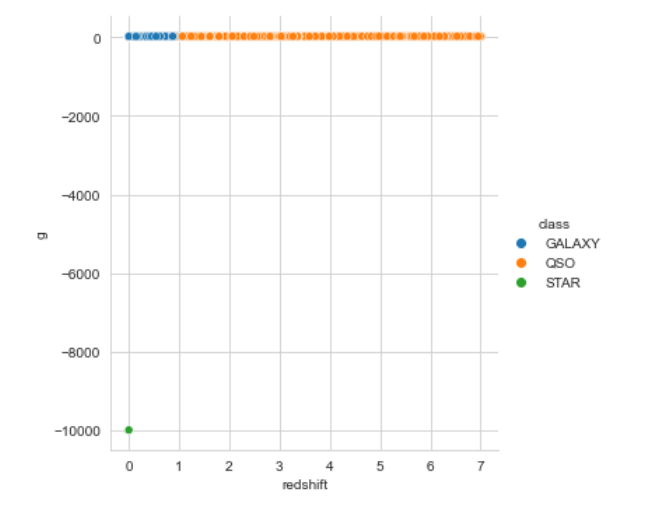


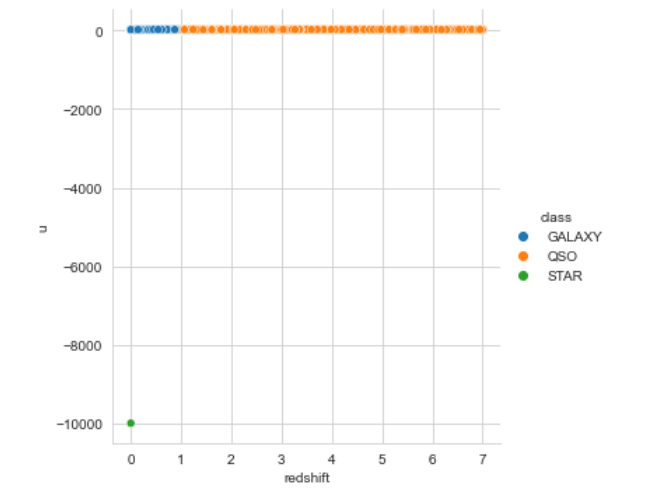
"class" is our target data and "alpha", "delta", "u", "g", "r", "i", "z", and "redshift" are common astronomical quantities so we want to use them as our primary features. Therefore, we drop the rest.

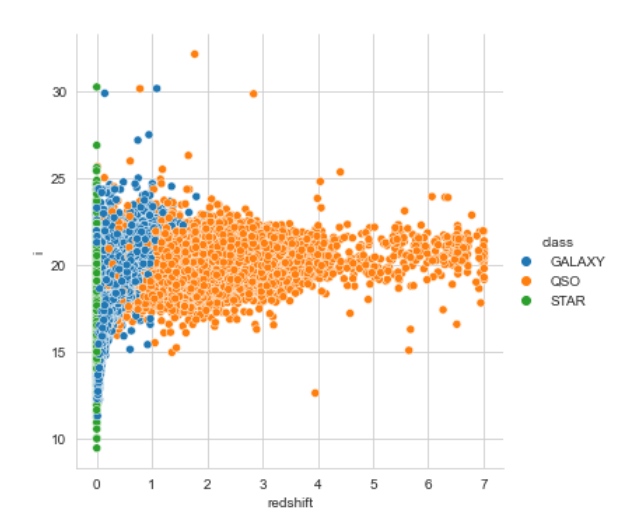
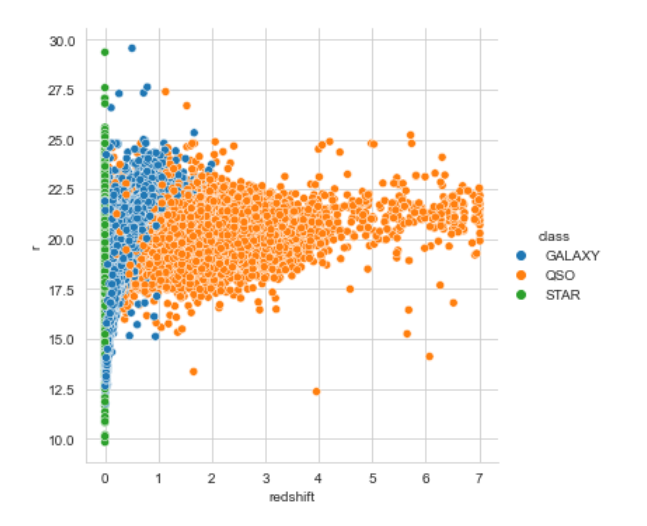
In which, “redshift” feature gives better classification, hence we compare “redshift” feature with all "alpha", "delta", "u", "g", "r", "i", "z" these columns.



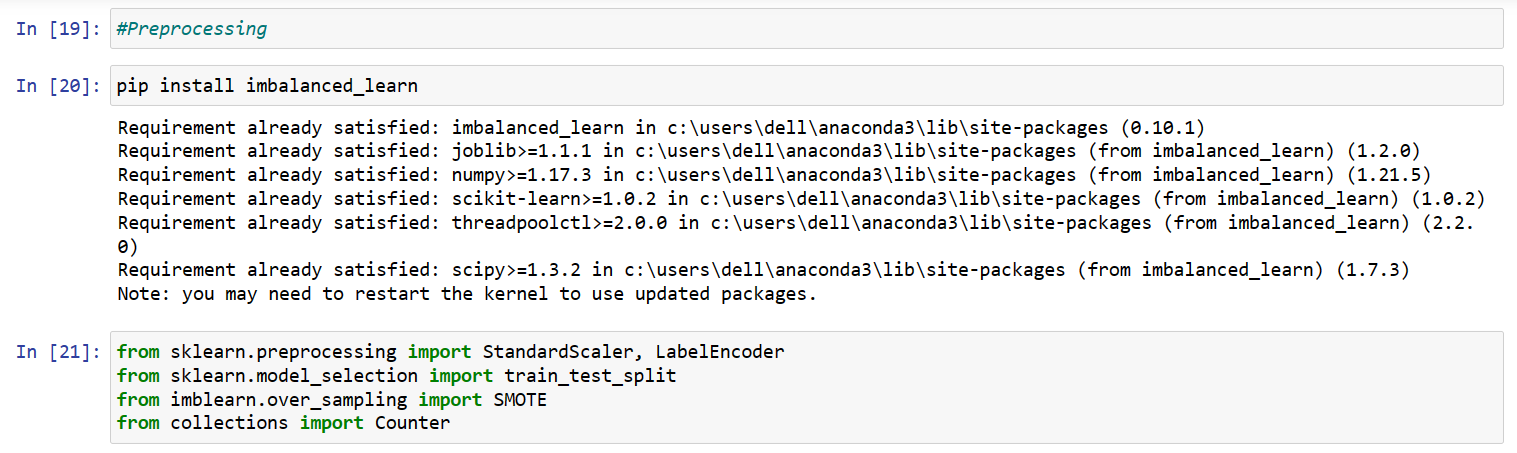


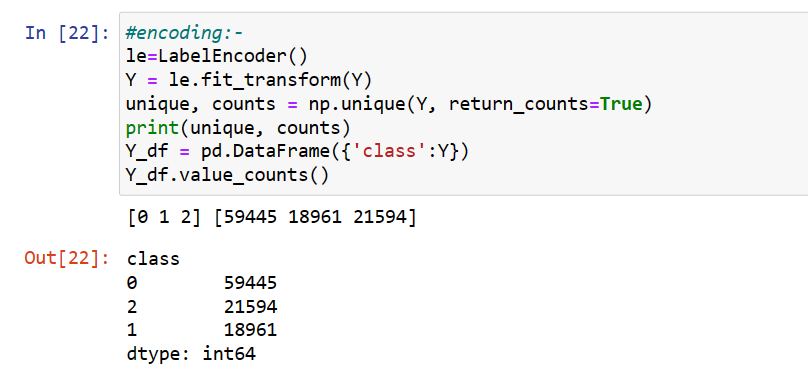


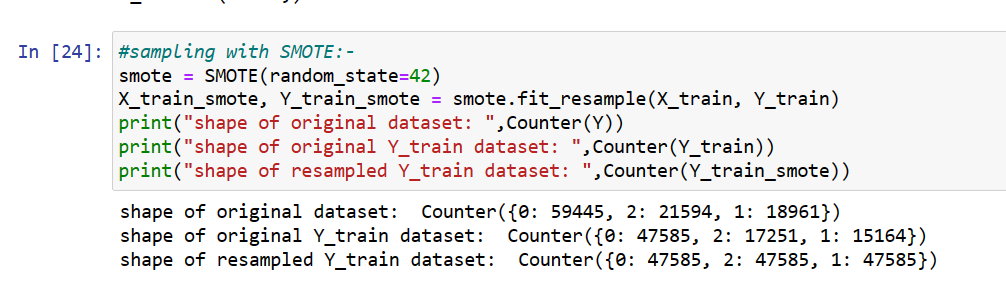


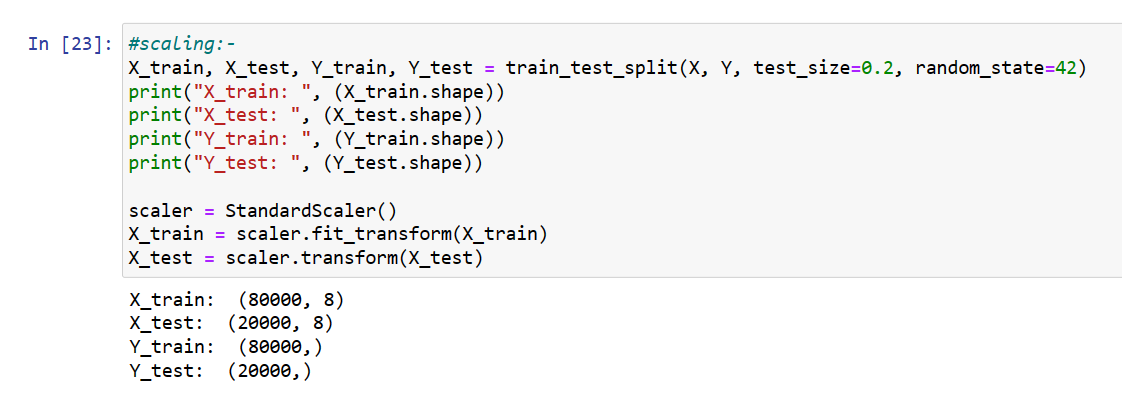
4.2) PRE-PROCESSING



* Encoding:
* Sampling With Smote:



* Scaling:



4.3) ALGORITHMS USED

1. **Support Vector Machine:**

Support Vector Machines (SVM) is a machine learning algorithm that can be used for classification, regression, and outlier detection. SVM is a type of supervised learning algorithm that is based on the idea of finding a hyperplane that best separates the different classes of data.

In the context of classification, SVM works by finding a hyperplane in a high-dimensional space that maximally separates the different classes of data points. The hyperplane is chosen so that the margin, or distance, between the hyperplane and the nearest data points from each class is maximized. The data points closest to the hyperplane are called support vectors, hence the name "Support Vector Machine."

1. **Random Forest:**

Random Forest is a machine learning algorithm that is used for both classification and regression tasks. It is an ensemble learning method that consists of a collection of decision trees.

In Random Forest, each decision tree is constructed using a random subset of the features and a random subset of the training examples. The tree is grown using a greedy algorithm that splits the data at each node based on the feature that best separates the examples in terms of their class labels or target values. The split is chosen to maximize the information gain or decrease in impurity, depending on the type of task.

1. **Logistic Regression:**

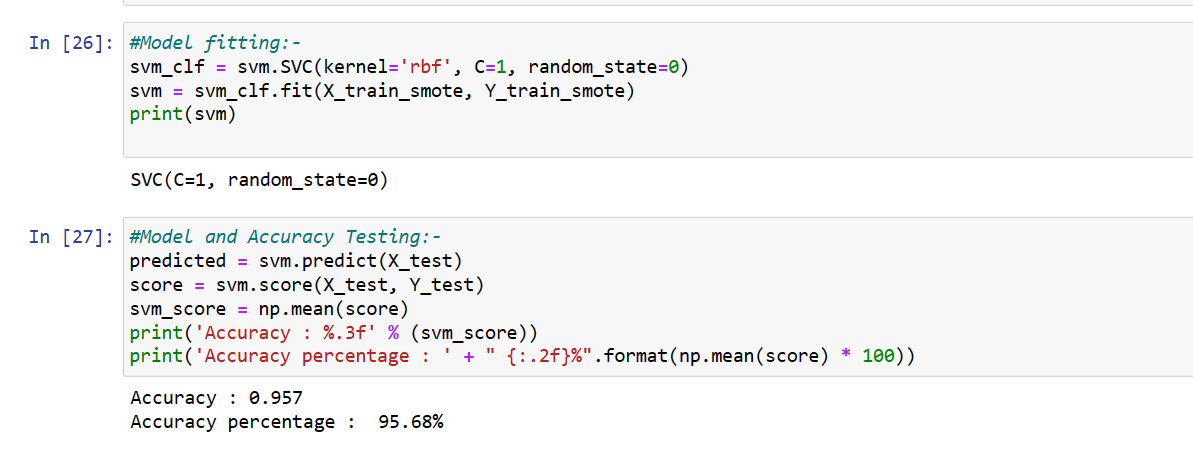
Logistic Regression is a statistical machine learning algorithm used for binary classification tasks, where the goal is to predict whether a data sample belongs to one of two classes, such as "spam" or "not spam," "fraudulent" or "not fraudulent," or "positive" or "negative."

In Logistic Regression, a model is trained to predict the probability of the positive class given a set of input features. The output of the model is a continuous value between 0 and 1, which represents the probability of the positive class. If the probability is greater than a certain threshold, typically 0.5, the sample is classified as belonging to the positive class; otherwise, it is classified as belonging to the negative class.

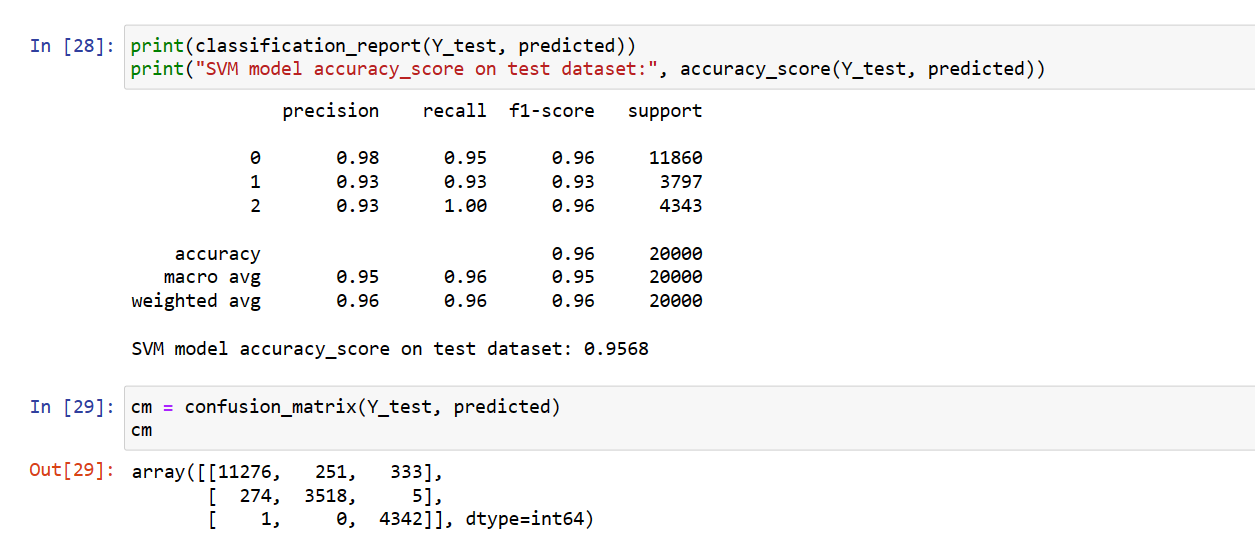
4.4) RESULTS

1. Support Vector Machine:

* Accuracy:

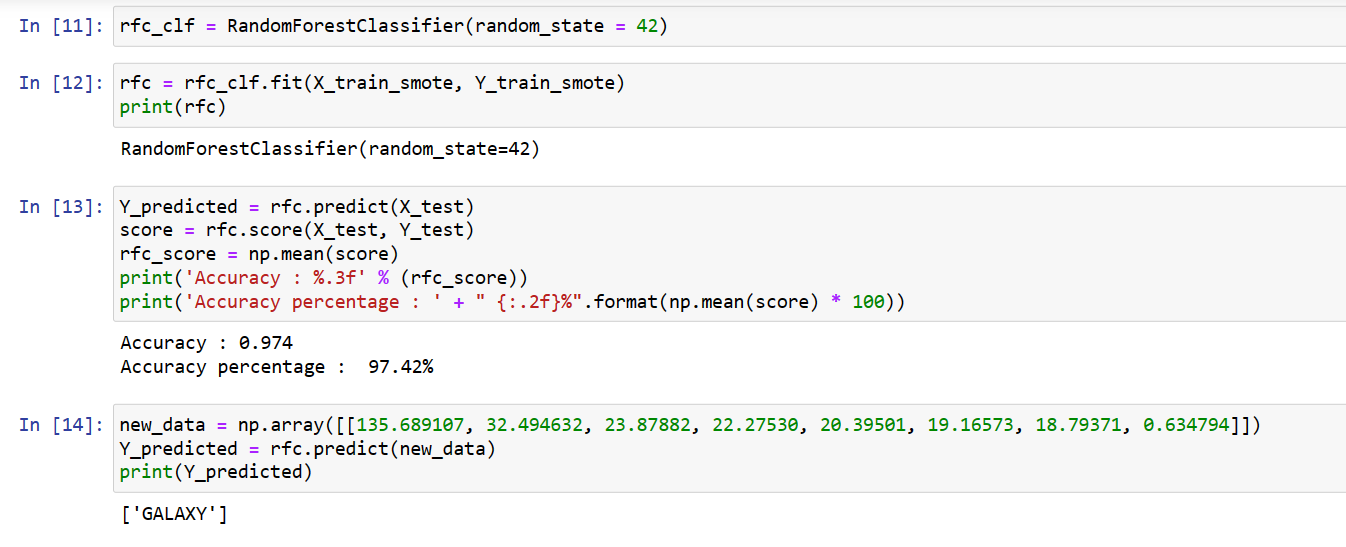


* Classification report & Confusion Matrix:



2. Random Forest:

* Accuracy:

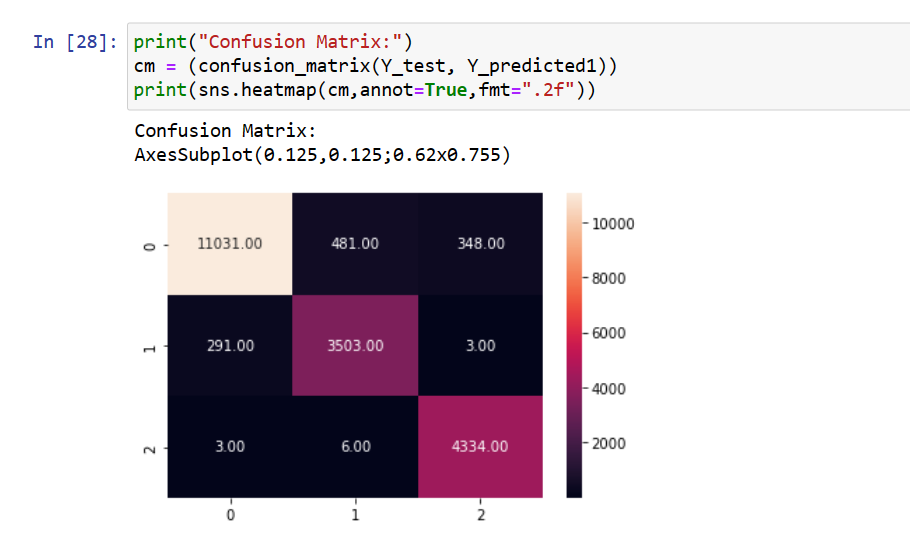


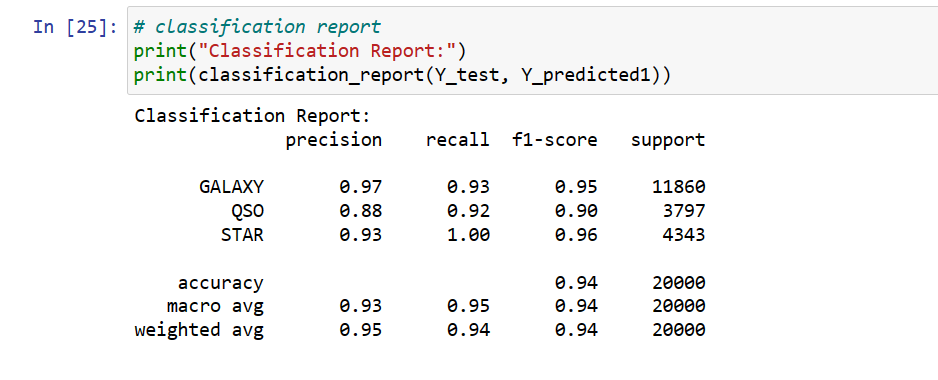
3.Logistic Regression:

* Accuracy:



* Classification report & Confusion Matrix:





**5) CONCLUSION & FUTURE WORK**

5.1) CONCLUSION

In conclusion, star-galaxy-quasar classification is an important problem in astronomy and astrophysics, with applications in galaxy evolution, cosmology, and extragalactic astronomy. Machine learning algorithms such as Random Forest, Support Vector Machines (SVMs) have been successfully applied to this problem, achieving high accuracy and performance. These algorithms can handle large datasets with complex features and relationships, and can provide insights into the physical properties and morphology of astronomical objects. However, the problem of class imbalance, where the number of examples in each class is not equal, remains a challenge for these algorithms, and more research is needed to address this issue. Overall, the use of machine learning in star-galaxy-quasar classification has great potential to advance our understanding of the universe and contribute to the development of new scientific discoveries.

5.2) FUTURE WORK

Future work for star-galaxy-quasar classification using machine learning could involve several directions.

Firstly, addressing the problem of class imbalance by exploring techniques such as data augmentation, oversampling, or weighted loss functions could help improve the performance of the algorithms.

Secondly, incorporating additional data sources, such as multi-wavelength observations, spectroscopic data, or deep learning architectures, could further enhance the accuracy and predictive power of the algorithms.

Lastly, developing more interpretable models that can provide insights into the physical properties and morphologies of astronomical objects could help advance our understanding of the universe and facilitate new scientific discoveries.

THANK YOU!!