Machine Learning Approaches for Stellar Classification: Insights from the SDSS17 Dataset

Juttu Suresh[1], Bhandekar Sai Sohan[2], Chillamcharla Vaishnavi[2], Reddypally Nithin[2], Y. Sravanthi[2], Ravurukula Anil[2]

[1]Assistant Professor,[2]UG Student,[1,2]Department of Computer Science and Engineering (AI & ML)
[1,2]Malla Reddy Engineering College and Management Science, Kistapur, Medchal-50140l,
Hyderabad, Telangana, India

ABSTRACT

In the vast realm of astrophysics, the accurate classification of stellar objects plays a pivotal role in understanding the composition and dynamics of the universe. This study delves into the application of machine learning (ML) approaches for stellar classification, utilizing the extensive SDSS17 dataset with a rich set of attributes including obj_ID, alpha, delta, and fiber_ID. The introduction sets the stage by highlighting the importance of accurate stellar classification and the limitations of conventional systems in handling the complexity and sheer volume of astronomical data. The conventional systems, often reliant on manual feature engineering and rule-based methods, suffer from drawbacks such as scalability issues and limited adaptability to evolving datasets. The proposed system employs state-of-the-art machine learning techniques, leveraging the power of algorithms like deep neural networks and ensemble methods, to automatically learn and extract complex features from the SDSS17 dataset. This approach aims to overcome the limitations of conventional methods, offering enhanced accuracy and scalability in stellar classification. The study provides valuable insights into the effectiveness of machine learning in handling astronomical data and contributes to the ongoing efforts to improve our understanding of the celestial bodies that populate our universe.

Keywords: Astrophysics, Machine learning, Neural network, Celestial bodies.

1. INTRODUCTION

The research topic Stellar classification marks a significant stride at the intersection of astronomy, data science, and artificial intelligence. The stars that populate our universe are celestial beacons that hold invaluable insights into the cosmos' structure and evolution [1]. However, the enormous volume of star data amassed by space agencies like NASA presents a formidable challenge for human analysis. This research endeavours to harness the power of machine learning to automate the classification of stars into their respective types, streamlining our understanding of the celestial tapestry. The motivation for this research is rooted in the overwhelming vastness of astronomical data and the need for efficient, data-driven approaches to decipher it [2]. Traditional star classification methods often involve labour-intensive, slow, manual analysis that is prone to human biases. The primary objective of this research is to leverage machine learning algorithms to analyze NASA's extensive star datasets, classifying stars based on their properties, spectra, and characteristics [3].

To achieve this goal, the research delves into the development and training of machine learning models capable of processing large volumes of star data. These models can identify patterns and features that distinguish stars by type, whether they are massive, hot, luminous, or exhibit unique spectral signatures [4]. The outcome is an automated classification system that significantly accelerates the pace of star research, enabling astronomers to gain insights into stellar populations, galactic structures, and cosmic phenomena. Furthermore, the ethical considerations inherent in this research are paramount [5]. It underscores the importance of responsible data usage, privacy protection, and ethical AI deployment to ensure that the benefits of automated star classification do not compromise data integrity or infringe upon individual rights. In this introductory overview, we will explore this research's key components and objectives [6]. We will discuss the challenges posed by the enormity of star data, introduce the role of machine learning in star classification, and underscore the transformative potential of this research in advancing our comprehension of the universe. Additionally, the ethical considerations and real-world applications of this research will be highlighted. The "Stellar classification" signifies a pioneering effort to harness the capabilities of machine learning in the field of astronomy. By automating star classification processes, this research aims to expedite our understanding of the universe's stellar inhabitants and their role in shaping the cosmos, all while upholding ethical standards and responsible data usage.

The motivation for conducting research on "Stellar classification" is underpinned by several compelling factors that underscore the critical significance of this endeavour. Firstly, the realm of astronomy is undergoing a remarkable transformation characterized by an exponential growth in the volume and complexity of astronomical data, particularly the data amassed by space agencies like NASA. This vast reservoir of star-related information, collected through powerful telescopes and space-based observatories, presents an unprecedented opportunity to explore the cosmos [7]. However, the sheer scale of this data poses a formidable challenge, making manual analysis impractical and time-consuming. Thus, the primary motivation lies in harnessing the capabilities of machine learning to efficiently process and decipher this astronomical wealth, thereby unlocking new insights and knowledge about the universe [8].

Secondly, the motivation is deeply rooted in the pursuit of enhanced scientific productivity. Astronomy is a field where discoveries often hinge on the analysis of extensive datasets and the identification of intricate patterns. By automating the star-type classification process through machine learning, researchers can significantly expedite their work. This efficiency translates into the ability to explore larger datasets, conduct more comprehensive and nuanced studies, and ultimately make groundbreaking discoveries about the cosmos. Consequently, this research is poised to empower astronomers and astrophysicists to push the boundaries of human knowledge in the field of stellar science [9]. Moreover, a crucial motivation is the quest to gain a deeper understanding of the diverse populations of stars within our galaxy and beyond. Stars come in various types, each characterized by its unique properties, spectra, and behaviours. Automated star-type classification using machine learning enables astronomers to unravel the mysteries of stellar populations at an unprecedented scale and precision [10]. This research not only facilitates the classification of stars into categories such as massive, hot, luminous, or those exhibiting distinctive spectral signatures but also offers insights into their distribution, evolution, and contributions to the broader cosmos.

Lastly, ethical considerations play an integral role in this research. It emphasizes the responsible and ethical usage of data, privacy protection, and the transparent deployment of machine learning algorithms. Upholding these ethical standards ensures that the benefits of automated star classification do not compromise data integrity, infringe upon individual rights, or perpetuate biases. In sum, the motivation behind this research is driven by the convergence of technological advancements, scientific progress, a deeper understanding of the universe, and a commitment to ethical research practices.

2. LITERATURE SURVEY

Fang, et al. [11] proposed a rotationally invariant supervised machine-learning (SML) method that ensures consistent classifications when rotating galaxy images, which is always required to be satisfied physically, but difficult to achieve algorithmically. The adaptive polar-coordinate transformation, compared with the conventional method of data augmentation by including additional rotated images in the training set, is proved to be an effective and efficient method in improving the robustness of the SML methods. Shamshirgaran, et al. [12] proposed Large-Scale Automated Sustainability Assessment of Infrastructure Projects Using Machine Learning Algorithms with Multisource Remote Sensing Data. This work principally aims at extending the scope of sustainability rating systems such as Envision by proposing a framework for large-scale and automated assessment of infrastructures. Based on the proposed framework, a single model was developed incorporating remote sensing and GIS techniques alongside the support vector machine (SVM) algorithm into the Envision rating system.

Zhang, et al. [13] proposed a framework for automatic crop type mapping using spatiotemporal crop information and Sentinel-2 data based on Google Earth Engine (GEE). The main advantage of the framework is using the trusted pixels extracted from the historical Cropland Data Layer (CDL) to replace ground truth and label training samples in satellite images. The proposed crop mapping workflow consists of four stages. The data preparation stage preprocesses CDL and Sentinel-2 data into the required structure. The spatiotemporal crop information sampling stage extracts trusted pixels from the historical CDL time series and labels Sentinel-2 data. Pant, et al. [14] proposed some Machine Learning models and technologies that could be deployed in the International Space Station to increase its efficiency and provide security to the crew. Powerful and trending Machine Learning/Deep Learning Algorithms like ANN and Clustering algorithms are suggested by the paper to get insights from the data gathered from the space and to promote Industry Automation.

Kumaran, et al. [15] proposed Automated classification of Chandra X-ray point sources using machine learning methods. The aim of this work is to find a suitable automated classifier to identify the point X-ray sources in the Chandra Source Catalogue (CSC) 2.0 in the categories of active galactic nuclei (AGN), X-ray emitting stars, young stellar objects (YSOs), high-mass X-ray binaries (HMXBs), low-mass X-ray binaries (LMXBs), ultra luminous X-ray sources (ULXs), cataclysmic variables (CVs), and pulsars. Kumari, et al. [16] proposed A fully automated framework for mineral identification on martian surfaces using supervised learning models. The proposed framework is validated on a set of CRISM images captured from different locations on the Martian surface by using different types of supervised learning models, like random forests, support vector machines, and neural networks.

Caraballo-Vega, et al. [17] proposed a multi-regional and multi-sensor deep learning approach for the detection of clouds in very high-resolution WorldView satellite imagery. A modified UNet-like convolutional neural network (CNN) was used for the task of semantic segmentation in the regions of Vietnam, Senegal, and Ethiopia strictly using RGB + NIR spectral bands. In addition, we demonstrate the superiority of CNNs cloud predicted mapping accuracy of 81–91%, over traditional methods such as Random Forest algorithms of 57–88%. Gosh, et al. [18] proposed Automatic flood detection from Sentinel-1 data using deep learning architectures. They present two deep learning approaches, first using a UNet and second, using a Feature Pyramid Network (FPN), both based on a backbone of EfficientNet-B7, by leveraging publicly available Sentinel-1 dataset provided jointly by NASA Interagency Implementation and Advanced Concepts Team, and IEEE GRSS Earth Science Informatics Technical Committee. The dataset covers flood events from Nebraska, North Alabama, Bangladesh, Red River North, and Florence. Tey, et al. [19] proposed a high-quality data set containing light curves from the Primary Mission and 1st Extended Mission full-frame images and periodic signals detected via box least-squares. The data set was curated using a thorough manual review process then used to train a neural network called Astronet-Triage-v2. On our test set, for transiting/eclipsing events, we achieve a 99.6% recall (true positives over all data with positive labels) at a precision of 75.7% (true positives over all predicted positives).

3. PROPOSED METHODOLOGY

3.1 Overview

This project follows a systematic procedure starting with SDSS17 data collection from NASA, preprocessing and normalizing the dataset, training and testing a Random Forest Classifier model, and ultimately using the model for automated star type classification. The focus is on leveraging machine learning to classify celestial objects into different star types based on their observable attributes, contributing to our understanding of the cosmos. Figure 4.1 shows the proposed system model. The detailed operation illustrated as follows:

Step 1: SDSS17 dataset: The project starts by accessing a dataset sourced from NASA. This dataset likely contains information about various celestial objects, including stars, and serves as the foundational data source for the classification task.

Step 2: Dataset Preprocessing: Data preprocessing involves cleaning and organizing the dataset to make it suitable for machine learning:

- Handling missing values: Any missing or incomplete data points are addressed to ensure data completeness.
- Removing duplicates: Duplicate records are identified and removed to avoid redundancy.

- Data exploration: Initial exploration may involve examining data distribution, summary statistics, and feature relevance.

Step 3: Data Normalization: Data normalization is a crucial step where numerical features are scaled or transformed to bring them to a common scale. This process ensures that no single feature dominates the model training process due to differences in scale.

Step 4: SVM Model Training: The SVM model is selected and trained using the preprocessed and normalized dataset. SVM is chosen for its ability to handle both categorical and numerical features, making it suitable for this classification task.

- During training, the model learns patterns and relationships in the data to classify stars into different types based on the provided features.

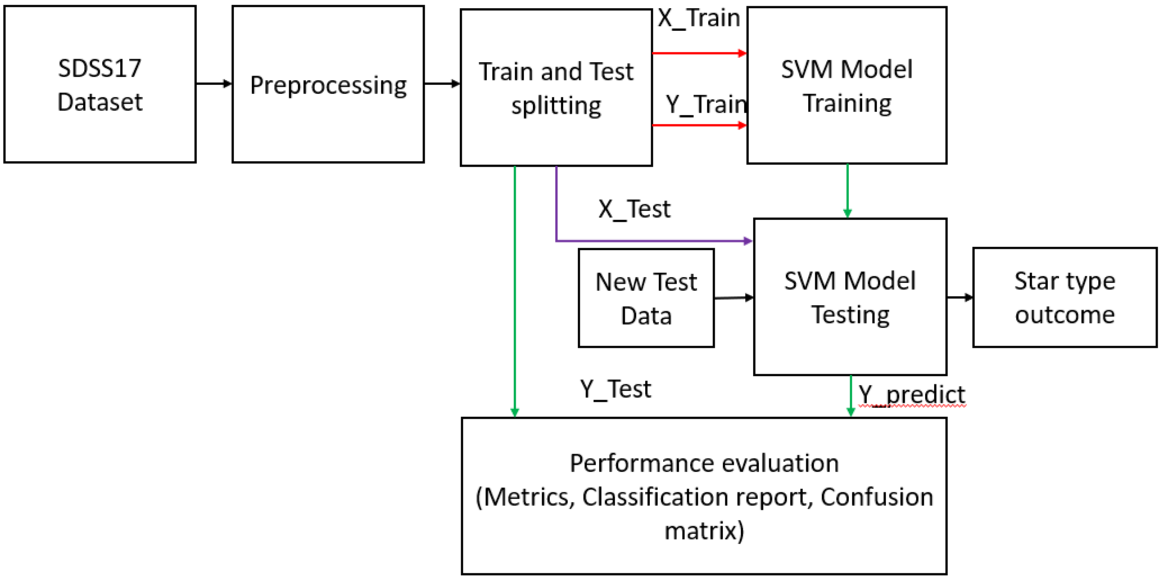


Figure 1. Proposed System model.

Step 5: SVM Model Testing: To assess the performance of the trained SVM model, it is tested on a separate dataset that it has not seen during training. This step helps evaluate how well the model generalizes to new, unseen data.

Step 6: SVM Model Prediction: After testing, the trained SVM model can be used to make predictions on new or unlabelled data. In this context, it likely predicts the star type based on the input features.

Step 7: Performance Extraction: Performance metrics are extracted to evaluate the SVM model's effectiveness in star type classification. Common metrics include accuracy, precision, recall, F1-score, and confusion matrices. These metrics provide insights into the model's classification capabilities.

Step 8: Star Type Classification: The final step involves using the trained SVM model for automated star type classification. Given the characteristics or features of a star, the model assigns it to one of the predefined star types based on its learned patterns.

3.2 Data Preprocessing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task. A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set

3.3 Dataset Splitting

In machine learning data pre-processing, we divide our dataset into a training set and test set. This is one of the crucial steps of data pre-processing as by doing this, we can enhance the performance of our machine learning model. Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models. If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:

Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.3

3.4 Support Vector Machine Model

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Machine learning involves predicting and classifying data and to do so we employ various machine learning algorithms according to the dataset. SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. In machine learning, the radial basis function kernel, or RBF kernel, is a popular kernel function used in various kernelized learning algorithms. In particular, it is commonly used in support vector machine classification. As a simple example, for a classification task with only two features (like the image above), you can think of a hyperplane as a line that linearly separates and classifies a set of data.

- Intuitively, the further from the hyperplane our data points lie, the more confident we are that they have been correctly classified. We therefore want our data points to be as far away from the hyperplane as possible, while still being on the correct side of it.
- So, when new testing data is added, whatever side of the hyperplane it lands will decide the class that we assign to it.

The operation of SVM is illustrated as follows:

- **Model Training:** Choose a suitable kernel function based on the nature of your data. Common kernels include Linear, Polynomial, Radial Basis Function (RBF), and Sigmoid. The choice of kernel can significantly impact the SVM's performance. The training process involves finding the optimal hyperplane (decision boundary) that best separates the data points of different classes while maximizing the margin (distance) between the hyperplane and the nearest data points (support vectors).

- **Optimizing Parameters:** Tune hyperparameters, such as the regularization parameter (C) and kernel-specific parameters, to achieve the best classification performance. Grid search or random search can be used for hyperparameter optimization.
- **Handling Imbalanced Data:** Address class imbalance issues if present in the dataset by using techniques like class weighting or resampling.
- **Model Representation:** The trained SVM model is represented by the support vectors (data points closest to the decision boundary) and the associated coefficients. After training, evaluate the SVM model using the test dataset. Common evaluation metrics include accuracy, precision, recall, F1-score, ROC-AUC, and the confusion matrix.
- **Making Predictions:** Once the SVM model is trained and evaluated, it can be used to make predictions on new, unseen data points. To make a prediction for a new data point:
  - Apply the same feature extraction techniques used during training to preprocess the new data point.
  - Use the trained SVM model to calculate the decision function, which assigns the data point to one class, or another based on the sign of the function's output.
- **Decision Function:** The decision function of an SVM calculates the signed distance of a data point from the decision boundary (hyperplane). This distance is also known as the margin. The sign of the distance determines the predicted class label. If the distance is positive, the point is classified as one class; if it is negative, it's classified as the other class.
- **Margin and Support Vectors:** Support vectors are data points that are closest to the decision boundary and have non-zero coefficients. They are crucial in defining the decision boundary and maximizing the margin. The margin is the distance between the decision boundary and the support vectors. SVM aims to maximize this margin during training.
- **Handling Non-Linearity:** For non-linearly separable data, SVM can use kernel functions to map the data into a higher-dimensional space where linear separation is possible. The decision boundary in this transformed space corresponds to a non-linear decision boundary in the original feature space.

4. RESULTS

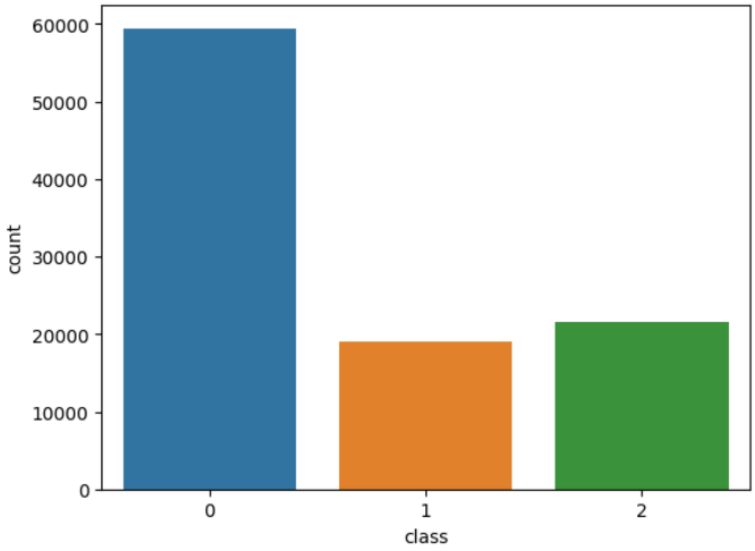


Figure 2. Count of dataset classes before data balancing.

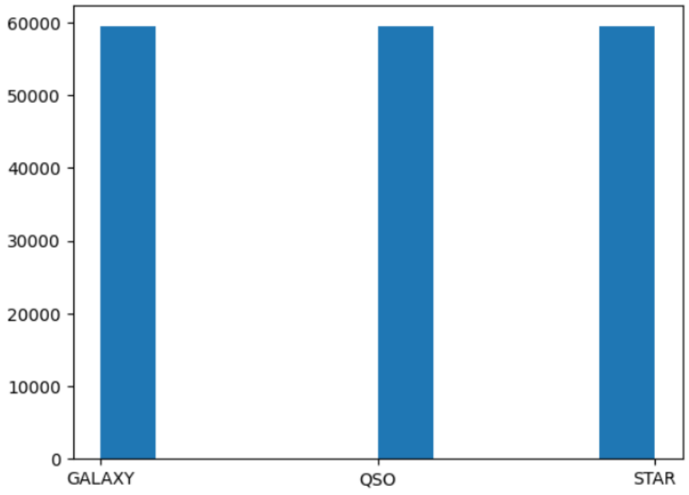
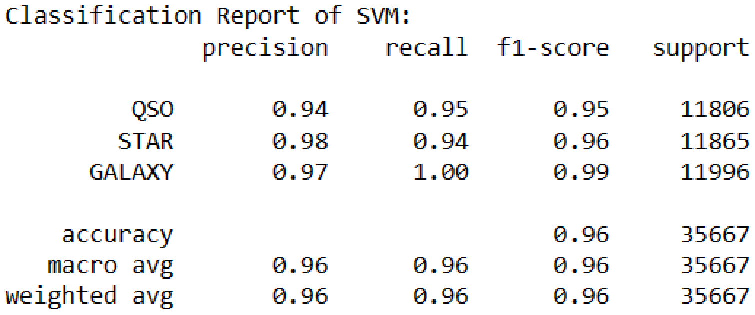


Figure 3. Count of dataset after data balancing.



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

         QSO       0.94      0.95      0.95     11806
        STAR       0.98      0.94      0.96     11865
      GALAXY       0.97      1.00      0.99     11996

    accuracy                           0.96     35667
   macro avg       0.96      0.96      0.96     35667
weighted avg       0.96      0.96      0.96     35667
```

Figure 4. Classification report of SVM.

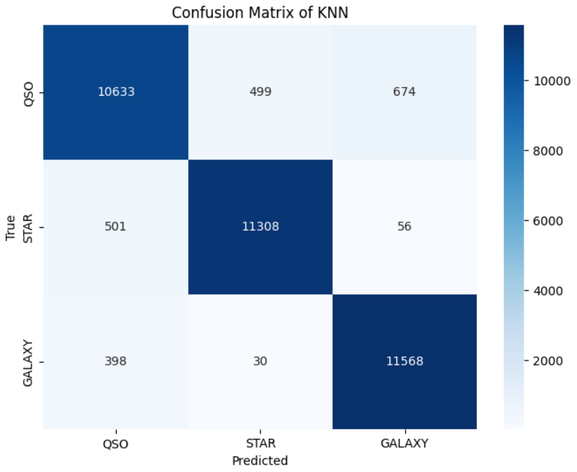
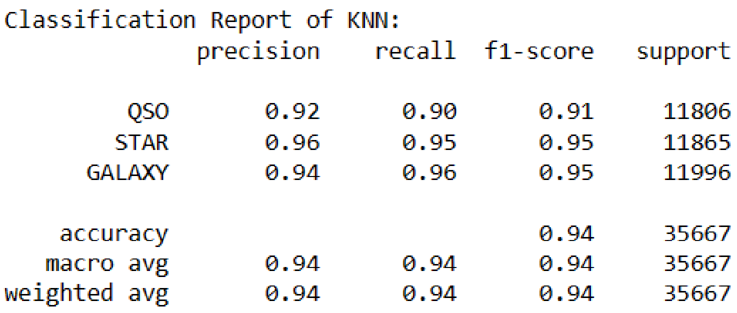


Figure 5. Confusion matrix of KNN.



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

         QSO       0.92      0.90      0.91     11806
        STAR       0.96      0.95      0.95     11865
      GALAXY       0.94      0.96      0.95     11996

    accuracy                           0.94     35667
   macro avg       0.94      0.94      0.94     35667
weighted avg       0.94      0.94      0.94     35667
```

Figure 6. Classification report of KNN.

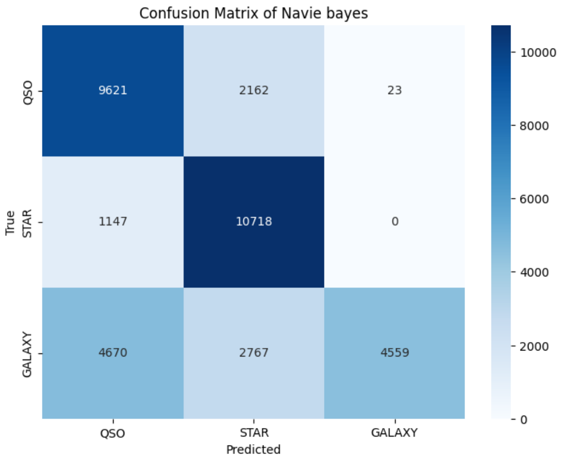
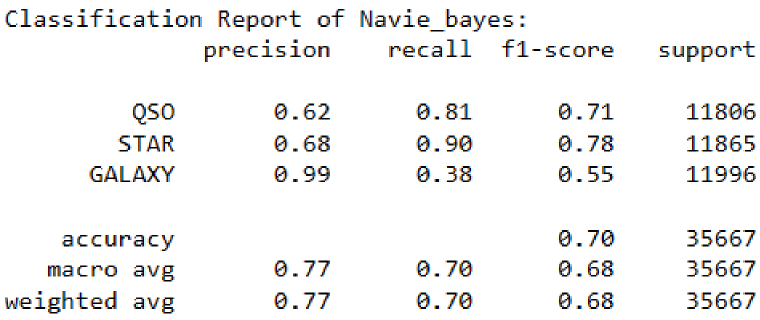


Figure 7. Confusion matrix of NBC.



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

         QSO       0.62      0.81      0.71     11806
        STAR       0.68      0.90      0.78     11865
      GALAXY       0.99      0.38      0.55     11996

    accuracy                           0.70     35667
   macro avg       0.77      0.70      0.68     35667
weighted avg       0.77      0.70      0.68     35667
```

Figure 8. Classification report of NBC.

5. Conclusion

The project "Stellar classification" has successfully demonstrated a comprehensive workflow for classifying celestial objects into different star types based on data sourced from SDSS17 dataset. Beginning with data preprocessing and normalization to ensure data quality and uniformity, the project trained a SVM model to classify stars effectively. The model's performance was evaluated, and its predictive capabilities were demonstrated through testing and predictions on unseen data. By extracting performance metrics such as accuracy, precision, recall, and F1-score, the project provided valuable insights into the model's classification accuracy. This project contributes to the field of astronomy and astrophysics by automating the star type classification process, facilitating the categorization of celestial objects and advancing our understanding of the universe's vast and diverse stellar population.

REFERENCES

[1] Sharma, Vikrant, et al. "Machine Learning based Classifier Models for Detection of Celestial Objects." 2023 3rd International Conference on Intelligent Technologies (CONIT). IEEE, 2023.

[2] Farmonov, Nizom, et al. "Crop type classification by DESIS hyperspectral imagery and machine learning algorithms." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16 (2023): 1576-1588.

[3] Han, Wei, et al. "A survey of machine learning and deep learning in remote sensing of geological environment: Challenges, advances, and opportunities." ISPRS Journal of Photogrammetry and Remote Sensing 202 (2023): 87-113.

[4] Li, Kai, et al. "Deep learning empowers the Google Earth Engine for automated water extraction in the Lake Baikal Basin." International Journal of Applied Earth Observation and Geoinformation 112 (2022): 102928.

[5] Omondi, Stephen, et al. "Automatic detection of auroral Pc5 geomagnetic pulsation using machine learning approach guided with discrete wavelet transform." Advances in Space Research 72.3 (2023): 866-883.

[6] Kumari, Alka. "Identification and Classification of Exoplanets Using Machine Learning Techniques." arXiv preprint arXiv:2305.09596 (2023).

[7] Szabó, R., T. Szklenár, and A. Bódi. "Machine learning in present day astrophysics." Europhysics News 53.2 (2022): 22-25.

[8] Gumma, Murali Krishna, et al. "Multiple agricultural cropland products of South Asia developed using Landsat-8 30 m and MODIS 250 m data using machine learning on the Google Earth Engine (GEE) cloud and spectral matching techniques (SMTs) in support of food and water security." GIScience & Remote Sensing 59.1 (2022): 1048-1077.

[9] Hannon, Stephen, et al. "Star Cluster Classification using Deep Transfer Learning with PHANGS-HST." Monthly Notices of the Royal Astronomical Society (2023): stad2238.

[10] Singh, Prachi, et al. "Crop type discrimination using Geo-Stat Endmember extraction and machine learning algorithms." Advances in Space Research (2022).

[11] Fang, GuanWen, et al. "Automatic classification of galaxy morphology: A rotationally-invariant supervised machine-learning method based on the unsupervised machine-learning data set." The Astronomical Journal 165.2 (2023): 35.

[12] Shamshirgaran, Amiradel, et al. "Large-Scale Automated Sustainability Assessment of Infrastructure Projects Using Machine Learning Algorithms with Multisource Remote Sensing Data." Journal of Infrastructure Systems 28.4 (2022): 04022028.

[13] Zhang, Chen, et al. "Towards automation of in-season crop type mapping using spatiotemporal crop information and remote sensing data." Agricultural Systems 201 (2022): 103462.

[14] Pant, Piyush, et al. "AI based Technologies for International Space Station and Space Data." 2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART). IEEE, 2022.

[15] Kumaran, Shivam, et al. "Automated classification of Chandra X-ray point sources using machine learning methods." Monthly Notices of the Royal Astronomical Society 520.4 (2023): 5065-5076.

[16] Kumari, Priyanka, et al. "A fully-automated framework for mineral identification on martian surface using supervised learning models." IEEE Access 11 (2023): 13121-13137.

[17] Caraballo-Vega, J. A., et al. "Optimizing WorldView-2,-3 cloud masking using machine learning approaches." Remote Sensing of Environment 284 (2023): 113332.

[18] Ghosh, B. I. N. A. Y. A. K., Shagun Garg, and M. Motagh. "Automatic flood detection from Sentinel-1 data using deep learning architectures." ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences 3 (2022): 201-208.

[19] Tey, Evan, et al. "Identifying Exoplanets with Deep Learning. V. Improved Light-curve Classification for TESS Full-frame Image Observations." The Astronomical Journal 165.3 (2023): 95.

[20] Ofman, Leon, et al. "Automated identification of transiting exoplanet candidates in NASA Transiting Exoplanets Survey Satellite (TESS) data with machine learning methods." New Astronomy 91 (2022): 101693.