

**ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING COURSE PROJECT DEPARTMENT OF CS&AI**

**ACADEMIC YEAR(2023-24)**

**THEME:**

**STAR TYPE CLASSIFICATION**

|  |  |
| --- | --- |
| **TEAM MEMBERS** | |
| **ST.NAME** | **HT.NUMBER** |
| **K.NEERADWAJA** | **2203A53006** |
| **A.SAI KIRAN** | **2303A53L01** |
| **B.THANUJ REDDY** | **2203A53016** |

**GUIDED BY**

**MR.D.RAMESH,**

**ASSOCIATE PROFESSOR,SR UNIVERSITY**

**STAR TYPE CLASSIFIACTION BASED ON**

**MACHINE LEARNING**

**ABSTRACT:**

In the realm of astronomy, the classification of stars based on their spectral characteristics plays a pivotal role. These spectral features provide insights into elements, temperature, density, and magnetic fields associated with celestial objects. Traditionally, astronomers have employed manual methods for star classification, which are time-consuming and prone to human error. However, the advent of machine learning, particularly deep learning models, has revolutionized this field.In this context, we explore an end-to-end machine learning solution for stellar classification—a critical problem in astronomy. By leveraging large datasets of labeled stellar spectra, we train convolutional neural networks (CNNs) and recurrent networks to discern star varieties and traits with remarkable precision. These models can detect subtle changes in stellar quality, enhancing our understanding of the cosmos.

Our approach encompasses essential components:

**Data Acquisition:** We utilize data from the Sloan Digital Sky Survey (SDSS), comprising 100,000 observations of space. Each data point is characterized by features such as right ascension, declination, photometric filters (ultraviolet, green, red, near-infrared, and infrared), and redshift.

**Pre-processing:** Feature engineering and data enhancement techniques are crucial for model performance. We preprocess the data to ensure its suitability for training**.**

**Model Selection:** SVM,KNN,descision tree,random forest,GBoost,XGboost and recurrent networks are chosen for their ability to learn intricate patterns from spectral data.

**Evaluation:** Rigorous evaluation metrics validate the model’s accuracy and generalization capabilities.Despite the promise of machine learning, challenges persist. An extensive labeled dataset is essential, and deep learning models require substantial computational resources. Nevertheless, the integration of machine learning into stellar classification holds immense potential. It can enhance efficiency, accuracy, and precision, benefiting both scientific research and our understanding of the universe.

**Introduction**

In the vast expanse of the cosmos, stars twinkle like celestial gems, each harboring its unique spectral signature. Stellar classification, akin to deciphering the cosmic code, unveils the secrets encoded in starlight. Let us embark on a journey through the astral realms, where machine learning and artificial intelligence (AI) converge to unravel the mysteries of distant suns.

**Stellar Classification:** A Machine Learning Odyssey

**Introduction to the Cosmic Symphony:**

Astronomy, the study of celestial objects beyond Earth’s atmosphere, beckons us to explore the enigmatic universe. Among the stars, spectral characteristics hold the key to their elemental composition, temperature, density, and magnetic fields.

Our mission: to classify stars, galaxies, and quasars (luminous supermassive black holes) based on their spectral fingerprints**.**

**The Stellar Dataset:**

Our star atlas comprises 100,000 observations captured by the Sloan Digital Sky Survey (SDSS). Each data point reveals a cosmic snapshot, described by 17 features and classified as a star, galaxy, or quasar.

Features include right ascension, declination, photometric filters (ultraviolet, green, red, near-infrared, and infrared), and redshift.

**Navigating the Celestial Sphere:**

Imagine an imaginary celestial sphere, concentric with Earth, where all celestial objects project their essence. The celestial equator, akin to Earth’s equator, guides our exploration 1.

Ascension and declination angles steer our cosmic ship, mapping the stars’ positions 1.

**Machine Learning Constellations:**

Our AI crew includes SVM,KNN and recurrent networks. Trained on spectral data, they learn to discern star varieties and traits with unprecedented accuracy.

Feature extraction and wavelet transformations enhance our starlight analysis.

**Challenges in the Galactic Voyage:**

Our quest demands extensive labeled data, akin to charting uncharted constellations.

Deep learning constellations require computational supernovae.

The Cosmic Promise:

Integration of machine learning into stellar classification augurs efficiency, precision, and cosmic enlightenment.

As we sail through the cosmic sea, our AI compass guides us toward understanding the universe—one star at a time.

May our algorithms illuminate the cosmic tapestry, revealing the stellar symphony hidden in the fabric of spacetime.

## Dataset

## The dataset consists [241 rows x 7 columns] a set of data created for star classification. I recommend using this data set for educational purposes, The star classification have these the features and Temperature,relative luminosity,relative radius,ABSOLUTE magnitude,Color,Spectral Class,Type.The Output of the dataset is Class . It is Target value from the data set build a machine learning model for star classification with a data set that has clear and concise target variables. The model will then be able to classify star images into their respective categories based on the labels you've provided.

## Methodology

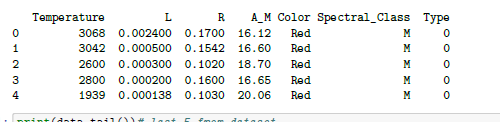
## About Data Set:

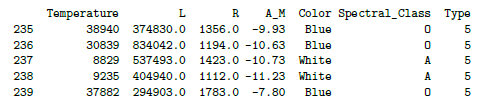
Name of Data set: STAR CLASSIFCATION

Taken From: KAGGLE WEBSITE

Link of Data set: [Star Type Classification / NASA (kaggle.com)](https://www.kaggle.com/datasets/brsdincer/star-type-classification)

DATA SET :





**DATA ANALYSIS :**

Temperature:

The first histogram (top left) shows a decrease in count as temperature increases. This suggests that lower temperatures are more common.

L (Luminosity):

The second histogram (top middle) labeled “L” has an x-axis with large values and shows a sharp decline in count at the beginning. This indicates that lower luminosity values are less frequent.

R (Radius):

The third histogram (top right) labeled “R” also exhibits a sharp decline in count at the beginning. Similar to luminosity, smaller radii are less common.

Color:

The fourth histogram (bottom left) labeled “Color” displays two prominent peaks around values 0 and 10 respectively. This suggests that there are specific color categories that occur frequently.

Spectral Class:

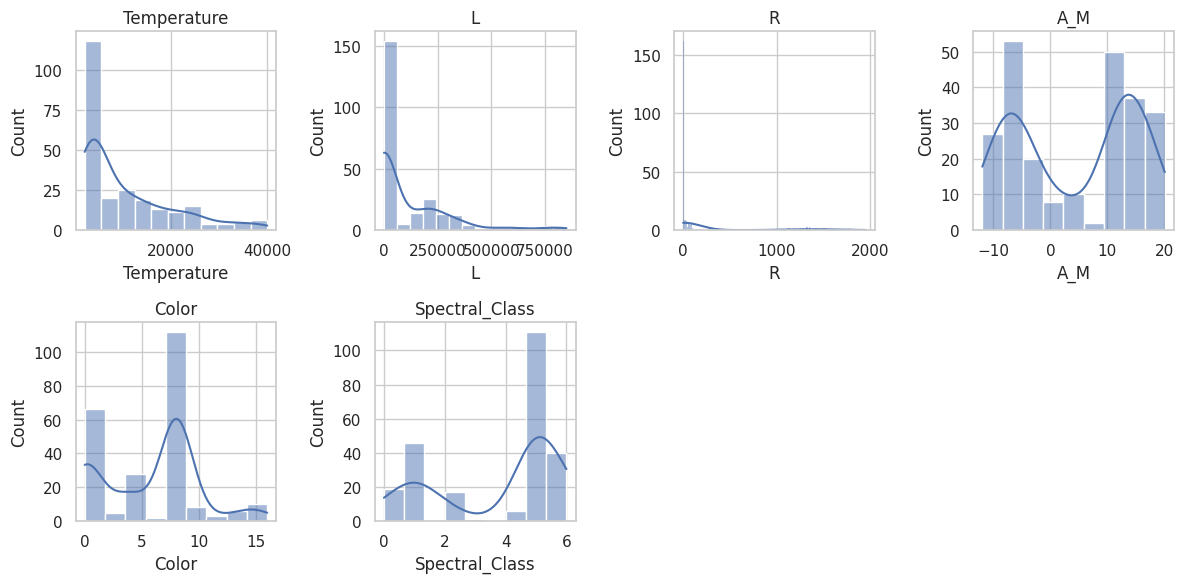
The fifth histogram (bottom middle) labeled “Spectral Class” shows counts for specific classes with peaks at class 2 and class 5. These spectral classes are likely significant in the dataset.

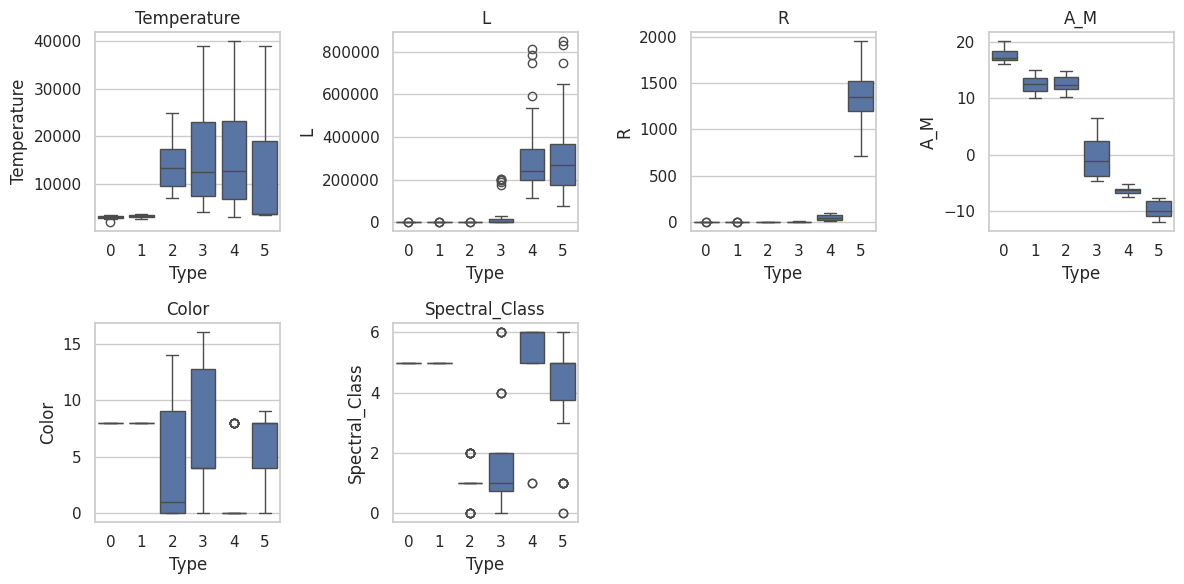
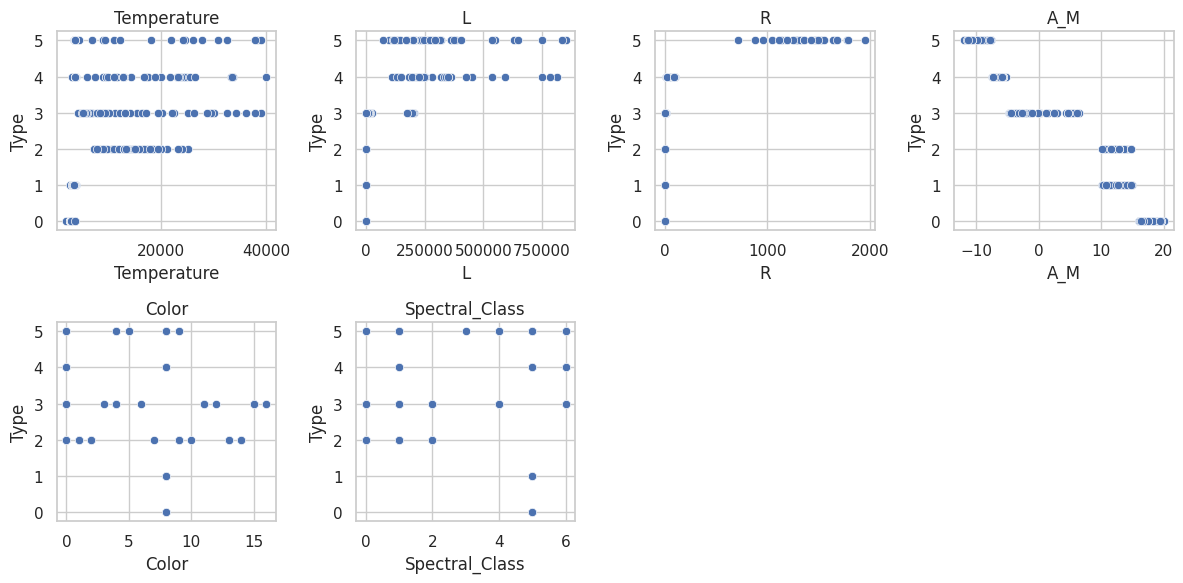
A\_M (Absolute Magnitude):

The sixth histogram (bottom right) depicts a symmetrical distribution with peaks around -5 and 15. This indicates that certain absolute magnitude values are more prevalent.

Please note that without additional context or labels, we can only infer patterns based on the shapes of the histograms. Further analysis would require understanding the specific variables and their significance in the dataset.

For example.





**CORRELATION MATRIX :**

The correlation matrix is a square matrix that shows the linear relationship between pairs of variables in a dataset. Each cell in the matrix represents the correlation coefficient between two variables. Here are some key points about the correlation matrix:

**Interpretation:**

Values range from -1 to 1.

Positive values indicate a positive linear relationship (direct proportionality).

Negative values indicate a negative linear relationship (inverse proportionality).

A value close to 0 indicates little or no linear relationship.

**Diagonal Elements:**

The diagonal elements (top-left to bottom-right) represent the correlation of each variable with itself, which is always 1.

**Off-Diagonal Elements:**

These elements represent the correlation between pairs of different variables.

For example, if the correlation coefficient between variables X and Y is 0.8, it suggests a strong positive linear relationship between X and Y.

**Use Cases:**

Correlation matrices are useful for feature selection, identifying multicollinearity, and understanding relationships between variables.

**Output Image:**

The image displays a symmetric matrix with numerical values.

Each cell represents the correlation coefficient between two variables.

Positive correlations are shown in shades of blue, while negative correlations are in shades of red.

**Example.**

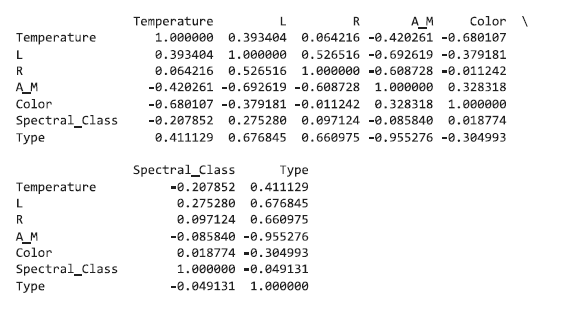


Fig 14

**Priority data processing**

Data preprocessing is an important step in machine learning that involves cleaning, transforming, and organizing the raw data into a format suitable for model training Ensuring the high quality of the data plays an important role in ensuring that machine learning models can identify logical patterns . Here is a detailed explanation of various aspects of data preprocessing.

**1. Data Hygiene:**

Processing missing values: Identifying and dealing with missing data, either by omitting a row or filling it in by methods such as mean, median, or interpolation.This data set does not contain zero values

**2. Data Modification:**

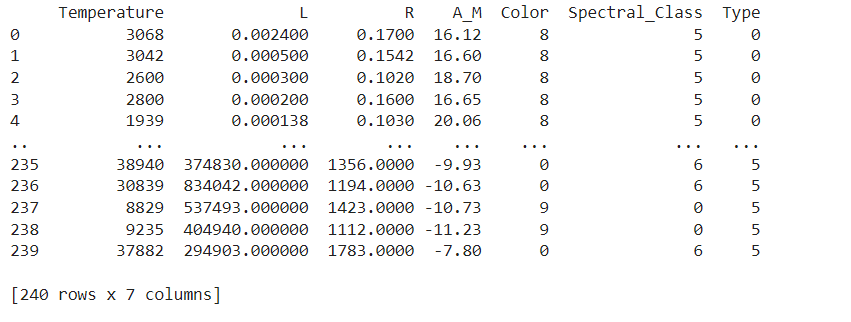
Feature calibration: A mathematical routine or standard used to ensure that features are on the same scale. Common methods include Min-Max scaling and z-score normalization.

**3.Classification of data:**

Train-Validation-Test Split: Split the data set into training, validation, and test sets. The training set is used to train the model, the validation set is used to tune the hyperparameters, and the test set is used to evaluate the performance of the model.

**4. Handling String Data:**

Text preprocessing: In natural language processing tasks, preprocess text data by tokenizing, eliminating stop words, performing stemming or lemmatizing, and converting text to numeric representation (e.g., TF-IDF or word embeddings) so.



## Preprocessing:

## The data set conisits of 241 rows in which the class is the target variable.

## The dataset conists testing and training data set of 0.2 random state of 35.

* **Implementation:**

Star type classification using AIML involves categorizing different types of stars based on various features such as luminosity, temperature, and size. Here's how different machine learning algorithms can be applied to this project:

**1) Logistic Regression:**

In star type classification using AIML, logistic regression can be utilized to classify stars into different types or categories based on their attributes such as temperature, luminosity, and size. By training a logistic regression model on a dataset containing features of various stars and their corresponding types, the algorithm can learn to predict the type of a given star based on its characteristics.

**2)** **Support Vector Machine:**

Support Vector Machine (SVM) can also be employed for star type classification. By utilizing SVM, the project can categorize stars into different types based on their features, similar to logistic regression. SVM can effectively handle non-linear relationships between features and can be trained to classify stars accurately based on their attributes.

**3)** **K Nearest Neighbors:**

K Nearest Neighbors (KNN) with can be applied to star type classification by considering the attributes of neighboring stars to determine the type of a given star. By training a KNN model on a dataset containing features of various stars and their types, the algorithm can classify a new star based on the types of its three nearest neighbors in the feature space.

**GBOOST:**Gradient Boosting is an ensemble learning technique used for both classification and regression tasks. It combines multiple weak models (usually decision trees) to create a strong predictive model. The process involves sequentially training models, where each new model corrects the errors made by the previous one. Gradient Boosting uses gradient descent to minimize the loss function. It is effective in handling non-linear relationships and is widely used in various domains, including star classification.   
  
Accuracy: 0.9561

1. **XGBOOST:**

XGBoost (Extreme Gradient Boosting) is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It has become one of the most popular algorithms due to its ability to handle large datasets and achieve state-of-the-art performance in various tasks, including star type classification.  
  
Accuracy: 0.9561

**RESULT ANALYSIS:**

## **Accuracy of the loggestric regression:**

## Logistic Regression Accuracy:

## Logistic Regression Accuracy: 0.9583333333333334

## Screenshot 2024-05-01 231445

## The classification report for the Logistic Regression model provides a detailed evaluation of its performance on the test data.

## **Accuracy:**

## The overall accuracy of the model is 0.96, which indicates that it correctly predicts the star types for approximately 96% of the instances in the test set.

## **Precision:**

## Precision measures the proportion of true positive predictions among all positive predictions.For each class (star type), precision is calculated. For example:

## Class 0 (Star Type 0): Precision is 0.89, meaning that 89% of the predicted Star Type 0 instances are actually correct.

## Class 1 (Star Type 1): Precision is 0.86.

## Class 2 (Star Type 2): Precision is 1.00.

## …

## Recall (Sensitivity):

## Recall quantifies the proportion of actual positive instances correctly predicted by the model.

## For each class, recall is calculated. For example:

## Class 0 (Star Type 0): Recall is 1.00, indicating that all actual Star Type 0 instances are correctly predicted.

## Class 1 (Star Type 1): Recall is 0.86.

## Class 2 (Star Type 2): Recall is 1.00.

## …

## F1-Score:

## F1-score is the harmonic mean of precision and recall.

## It balances precision and recall, especially useful when dealing with imbalanced datasets.

## For each class, F1-score is calculated. For example:

## Class 0 (Star Type 0): F1-score is 0.94.

## Class 1 (Star Type 1): F1-score is 0.86.

## Class 2 (Star Type 2): F1-score is 1.00.

## …

## **Support:**

## Support represents the number of instances for each class in the test set.

## Quality and Size of Dataset:

## Unfortunately, the image does not provide information about the quality and size of the dataset used for star type classification.

## Feature Selection:

## The code snippet does not explicitly mention feature selection. We need additional context to understand which features were used.

## Non-Linear Relationships:

## The logistic regression model assumes linear relationships between features and the target variable. If there are non-linear relationships, other models (e.g., decision trees, neural networks) might perform better.

## CONFUSION MATRIX:

## The confusion matrix is a visual representation of the performance of a classification model.

## Output :

## The image displays a heatmap representing the confusion matrix.

## The x-axis represents the predicted labels (0, 1, 2, etc.), while the y-axis represents the true labels.

## Each cell in the matrix contains the count of instances falling into that category.

## The diagonal cells (from top left to bottom right) represent correct predictions (TP and TN).

## Off-diagonal cells represent incorrect predictions (FP and FN).

## The color intensity (shades of blue) indicates the number of instances in each category.

## Interpretation:

## For example:

## The cell at row 0, column 0 represents true positives (correctly predicted Star Type 0 instances).

## The cell at row 1, column 0 represents false negatives (actual Star Type 0 instances misclassified as other types).

## The cell at row 0, column 1 represents false positives (predicted as Star Type 0 but actually other types).

## The cell at row 1, column 1 represents true negatives (correctly predicted non-Star Type 0 instances).

## Model Evaluation:

## The confusion matrix helps assess the model’s performance, especially in terms of misclassifications.

## It provides insights into which classes are well-predicted and which need improvement.

## 

## Quality of Model:To evaluate the overall quality of the model, we need additional information such as precision, recall, F1-score, and accuracy.

Classification Report for k-Nearest Neighbors (kNN):

precision recall f1-score support

0 0.70 0.88 0.78 8

1 0.80 0.57 0.67 7

2 0.86 1.00 0.92 6

3 1.00 0.88 0.93 8

4 0.33 0.38 0.35 8

5 0.50 0.45 0.48 11

accuracy 0.67 48

macro avg 0.70 0.69 0.69 48

weighted avg 0.68 0.67 0.67 48

## 

## Classification Report for Logistic Regression:

## precision recall f1-score support

## 0 0.89 1.00 0.94 8

## 1 0.86 0.86 0.86 7

## 2 1.00 1.00 1.00 6

## 3 1.00 0.88 0.93 8

## 4 1.00 1.00 1.00 8

## 5 1.00 1.00 1.00 11

## accuracy 0.96 48

## macro avg 0.96 0.96 0.96 48

## weighted avg 0.96 0.96 0.96 48

## Classification Report for Decision Tree:

## precision recall f1-score support

## 0 1.00 1.00 1.00 8

## 1 1.00 1.00 1.00 7

## 2 1.00 1.00 1.00 6

## 3 1.00 1.00 1.00 8

## 4 1.00 1.00 1.00 8

## 5 1.00 1.00 1.00 11

## accuracy 1.00 48

## macro avg 1.00 1.00 1.00 48

## weighted avg 1.00 1.00 1.00 48

## 

## Classification Report for Random Forest:

## precision recall f1-score support

## 0 1.00 1.00 1.00 8

## 1 1.00 1.00 1.00 7

## 2 1.00 1.00 1.00 6

## 3 1.00 1.00 1.00 8

## 4 1.00 1.00 1.00 8

## 5 1.00 1.00 1.00 11

## accuracy 1.00 48

## macro avg 1.00 1.00 1.00 48

## weighted avg 1.00 1.00 1.00 48

## PERFORMANCE COMPARISION :

## IMG_256

## 

## ACCURACY :

## Fig 24

## **Learning curve for logistic regression:**

## IMG_256

## **Precision-recall curve for logistic regression:**

## IMG_256

## IMG_256

## IMG_256

## IMG_256

## Fig 27

## IMPROVEMENT :

## The success of a star classification system hinges on several critical factors. Firstly, the quality of the dataset significantly impacts model training and accuracy. Selecting an appropriate machine learning framework is equally crucial. Options such as Support Vector Machines (SVMs), logistic regression, k-Nearest Neighbors (k-NN), and regression methods each have their strengths and should be chosen based on the specific requirements of the multi-class classification task. Additionally, domain-specific expertise is essential for fine-tuning the system to the intricacies of star classification, ensuring efficiency, and delivering precise and reliable results.

## CONCLUSION:

## For star classification using SVM, logistic regression, k-NN (k-nearest neighbors), and regression methods, several general conclusions can be drawn:SVM excels in categorizing stars with complex and nonlinear relationships between their attributes, such as luminosity, temperature, and size. It can effectively separate high-dimensional data and identify patterns that may not be linearly separable. SVM is particularly useful when distinguishing between different types of stars with distinct features.Logistic regression serves as a dependable alternative, especially in scenarios where the classification task involves binary outcomes or when interpretability of the results is crucial. It provides easily interpretable coefficients that indicate the influence of each feature on the star's classification.k-NN demonstrates effectiveness in star classification tasks where local structures and relationships play a significant role. By considering the attributes of nearby stars, k-NN can accurately classify stars based on their similarity to neighboring data points. This method is particularly valuable when identifying stars that share similar characteristics with their immediate neighbors.Regression models, on the other hand, can be employed to predict continuous traits associated with stars, such as their brightness or temperature. By analyzing the relationships between various features and the continuous traits of stars, regression methods can provide valuable insights into their properties.The selection of the most appropriate method depends on the characteristics of the star data and the objectives of the classification task. A careful consideration of these methods, along with the specific requirements of the star classification project, can lead to a more accurate and standardized solution for categorizing stars based on their attributes.

## REFERENCES :

## Wissner-Gross, A. "Datasets Over Algorithms". Edge.com. Retrieved 8 January 2016.

## Weiss, G. M.; Provost, F. (1 September 2003). "Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction". Journal of Artificial Intelligence Research. AI Access Foundation. 19: 315–354. doi:10.1613/jair.1199. ISSN 1076-9757. S2CID 2344521.

## Turney, Peter (2000). "Types of cost in inductive concept learning". arXiv:cs/0212034.

## Abney, Steven (17 September 2007). Semisupervised Learning for Computational Linguistics. CRC Press. ISBN 978-1-4200-1080-0.

## Žliobaitė, Indrė; Bifet, Albert; Pfahringer, Bernhard; Holmes, Geoff (2011). "Active Learning with Evolving Streaming Data". Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science. Vol. 6913. Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 597–612. doi:10.1007/978-3-642-23808-6\_39. ISBN 978-3-642-23807-9. ISSN 0302-9743.

## McAuley, Julian; Targett, Christopher; Shi, Qinfeng; Anton van den Hengel (2015). "Image-based Recommendations on Styles and Substitutes". arXiv:1506.04757 [cs.CV].

## "Amazon review data". nijianmo.github.io. Retrieved 8 October 2021.

## Ganesan, Kavita; Zhai, Chengxiang (2012). "Opinion-based entity ranking". Information Retrieval. 15 (2): 116–150. doi:10.1007/s10791-011-9174-8. hdl:2142/15252. S2CID 16258727.

## Lv, Yuanhua, Dimitrios Lymberopoulos, and Qiang Wu. "An exploration of ranking heuristics in mobile local search." Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012.

## Harper, F. Maxwell; Konstan, Joseph A. (2015). "The MovieLens Datasets: History and Context". ACM Transactions on Interactive Intelligent Systems. 5 (4): 19. doi:10.1145/2827872. S2CID 16619709.

## Koenigstein, Noam, Gideon Dror, and Yehuda Koren. "Yahoo! music recommendations: modeling music ratings with temporal dynamics and item taxonomy." Proceedings of the fifth ACM conference on Recommender systems. ACM, 2011.

## McFee, Brian, et al. "The million song dataset challenge." Proceedings of the 21st international conference companion on World Wide Web. ACM, 2012.