**Report**

**Stellar Classification**

*By*

**Group No: 2**

**Utkarsh Pise [BT23CSA032]**

**Divwardhan Agrawal [BT23CSA023]**

**Vishal Raykhere [BT23CSA019]**

**Anurag Gaikwad [BT23CSA008]**

**Semester: 3 Branch: CSA**

**Course Name: AIML Workshop-1 [CSP203]**

*Under the guidance of*

**Ms. Madhuri Dubey**



**Department of Computer Science and Engineering Indian Institute of Information Technology Nagpur - 441108 (India)**

**Dec – 2024**

**© Indian Institute of Information Technology (IIIT) 2024**

**Report**

**Stellar Classification**

**Introduction**

The classification of astronomical objects—such as stars, galaxies, and nebulae—is a foundational aspect of astronomical research, with far-reaching implications in understanding the universe. Distinguishing between different types of celestial bodies provides insights into the processes that govern their formation, evolution, and distribution. For instance, categorizing galaxies by shape and structure aids in examining their developmental pathways and contributes to theories on the universe's expansion and structure. Similarly, analyzing star types by temperature, luminosity, and spectral lines informs our knowledge of stellar lifecycles, ultimately revealing the lifecycle of elements essential for planetary systems and life.

As astronomical data has grown exponentially with advancements in telescope technology and space missions, traditional classification methods have become time-consuming and challenging. Modern surveys produce vast amounts of data—often beyond what can be analyzed manually. Thus, automating the classification process through machine learning (ML) algorithms has become not only practical but essential. Automation enables astronomers to rapidly sift through enormous datasets, identify patterns, and make classifications with improved accuracy, freeing time for in-depth analyses and new discoveries. Such efficiency is especially beneficial in identifying rare objects that might otherwise go unnoticed in manual reviews.

Beyond its scientific and operational benefits, this project offers educational value. Applying machine learning to astronomical classification presents a unique interdisciplinary challenge, combining principles from both astronomy and data science. For students and researchers, this project serves as a real-world application of machine learning algorithms, fostering skill development in data preprocessing, feature engineering, and model training. It offers hands-on experience with astronomical datasets, promoting a deeper understanding of both machine learning techniques and the complex nature of celestial classification. Consequently, this project serves as an essential learning opportunity in the rapidly evolving field of data-driven astronomy, bridging the gap between theoretical knowledge and practical application.

* obj\_ID = Object Identifier, the unique value that identifies the object in the image catalog used by the CAS
* alpha = Right Ascension angle (at J2000 epoch)
* delta = Declination angle (at J2000 epoch)
* u = Ultraviolet filter in the photometric system
* g = Green filter in the photometric system
* r = Red filter in the photometric system
* i = Near Infrared filter in the photometric system
* z = Infrared filter in the photometric system
* run\_ID = Run Number used to identify the specific scan
* rereun\_ID = Rerun Number to specify how the image was processed
* cam\_col = Camera column to identify the scanline within the run
* field\_ID = Field number to identify each field
* spec\_obj\_ID = Unique ID used for optical spectroscopic objects (this means that 2 different observations with the same spec\_obj\_ID must share the output class)
* class = object class (galaxy, star or quasar object)
* redshift = redshift value based on the increase in wavelength
* plate = plate ID, identifies each plate in SDSS
* MJD = Modified Julian Date, used to indicate when a given piece of SDSS data was taken
* fiber\_ID = fiber ID that identifies the fiber that pointed the light at the focal plane in each observation

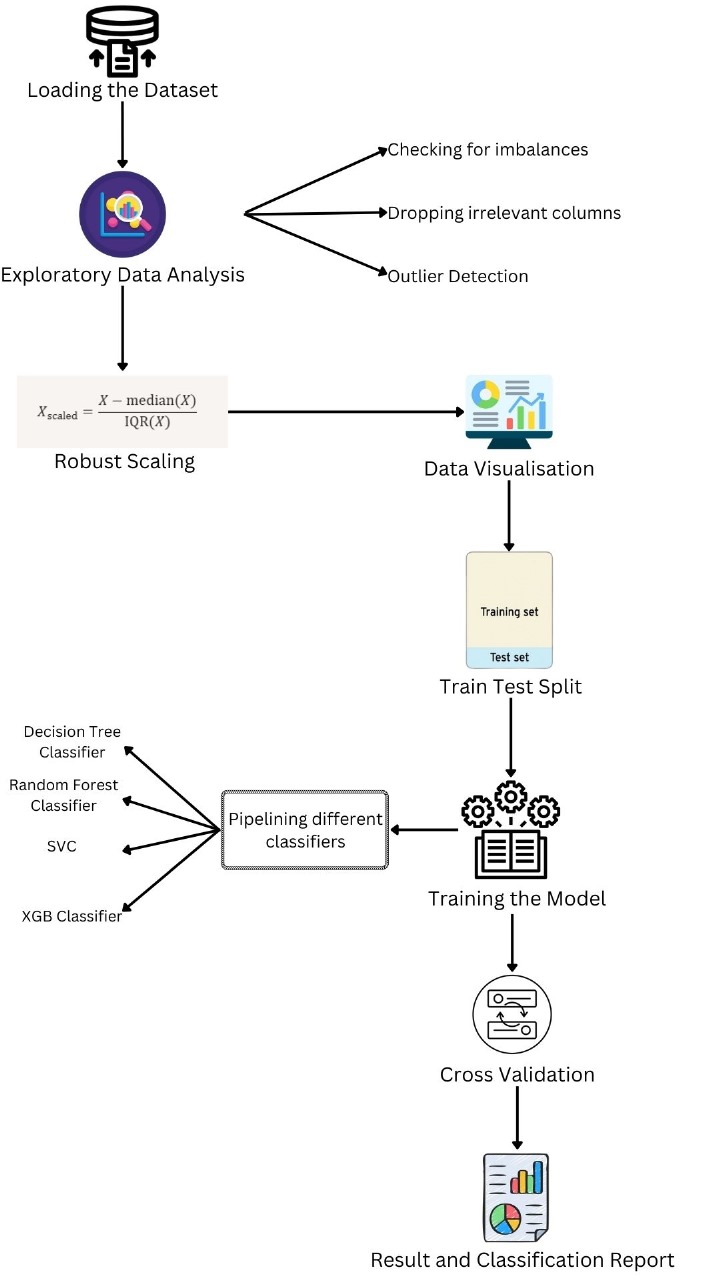
**Literature Survey**

The classification of celestial objects such as stars, galaxies, and quasars has been an area of significant research, advancing alongside astronomical discoveries and computational methods. In early approaches, classification relied heavily on human observation and manual cataloging. However, as astronomical data volumes increased, the need for automated classification systems became essential. Recent studies, such as those by Xiao-Qing & Jin-Meng (2021), have applied machine learning (ML) techniques to classify celestial objects using data from sources like the LAMOST (Large Sky Area Multi-Object Fiber Spectroscopic Telescope) survey. Their research utilized ML models to classify stars, galaxies, and quasars, demonstrating the effectiveness of ML in handling large datasets and complex patterns that traditional methods struggle to interpret effectively. Moreover, research in machine learning for astronomical applications has expanded to include various algorithms such as Decision Trees, Random Forests, and Support Vector Machines (SVM). For instance, Decision Trees, which are well-suited for classification tasks, have been utilized by researchers due to their interpretability and ease of implementation .Random Forest, an ensemble method that builds multiple decision trees, has proven particularly effective in achieving high accuracy with astronomical datasets. Studies have shown that Random Forest classifiers can handle class imbalances and perform well across different celestial object classes. The SVM model, known for its capacity to handle non-linear relationships, has also been adapted for classifying celestial objects with promising results, especially in distinguishing subtle differences between classes (IBM, n.d.). Overall, the literature highlights that ML-driven classification offers significant advantages in processing vast astronomical datasets, providing accurate and rapid classifications that enhance the understanding of celestial objects' nature and formation.

**Related Work:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| S.no. | Research paper | Dataset Source | Models Used | Accuracy  score | precision  score | Recall  score | F1 score |
| 1 | Qi, Zhuliang. "Stellar Classification by Machine Learning." *SHS Web of Conferences*. Vol. 144. EDP Sciences, 2022. | Sloan Digital Sky Survey | Decision Tree | 0.97 | 0.97 | 0.96 | 0.95 |
| Random Forest | 0.98 | 0.96 | 0.97 | 0.98 |
| Support Vector Machine | 0.97 | 0.97 | 0.95 | 0.97 |
| 2 | Villarreal, Jesus Tamez, and Sophia Barton. "Stellar Classification based on Various Star Characteristics using Machine Learning Algorithms." *Journal of Student Research* 12.1 (2023). | Kaggle | Random  forest | 0.9375 | 0.9188 | 0.9370 | 0.9259 |
| Decision Tree | 0.9375 | 0.9531 | 0.9375 | 0.9341 |
| Support Vector Classifier | 0.8959 | 0.8699 | 0.7292 | 0.7858 |
| 3 | Subramani, K., Pushpavalli, K., Renuka, G. B., & Subburaj, M. S. (2024). Supervised Machine Learning Algorithm for Stellar Classification. *Nanotechnology Perceptions*, 499-509. | Sloan Digital Sky Survey | Random forest | 0.9771 | 0.98 | 0.99 | 0.98 |
| Multilayer  perceptron | 0.9635 | 0.95 | 0.98 | 0.97 |
| Decision tree | 0.9612 | 0.97 | 0.97 | 0.97 |

**Methodology*:***



1 **Data Loading and Exploration**

* Load dataset ()
* Display initial data insights

2 **Data Preprocessing**

* **Duplicate Removal**: Identify and remove duplicate rows
* **Unique Values Analysis**: Visualize the count of unique values per feature
* **Class Distribution Analysis**: Plot the distribution of classes (bar and pie chart)
* **Outlier Detection**: Use IQR method to identify outliers and visualize with box plots
* **Feature Scaling**: Apply RobustScaler to handle outliers

3 **Feature Engineering and Transformation**

* Drop irrelevant columns (obj\_ID, run\_ID, etc.)
* Encode categorical target variable (class) using LabelEncoder
* Create new features if needed (e.g., u\_g, g\_r)

4 **Exploratory Data Analysis (EDA)**

* **Class-Based Visualizations**:
  + KDE plots for each feature by class
  + Average of each class for u, g, r, i, z
  + Histograms for alpha, delta, and redshift by class
  + Scatter plot of alpha vs. delta by class
* **Correlation Analysis**: Display heatmaps for correlations

5 **Model Training and Cross-Validation**

* Define features (X) and target (y)
* Split data into train and test sets (train\_test\_split)
* Create pipelines for models:
  + **Random Forest Classifier**
  + **Decision Tree Classifier**
  + **XGBoost Classifier**
  + **Support Vector Classifier**
* Perform 10-fold cross-validation for each model and evaluate accuracy

6 **Model Evaluation**

* Fit the best-performing model to the training set
* Predict on the test set
* Print classification\_report for detailed performance analysis across classes A graph of blue bars

  Description automatically generated with medium confidence

A bar chart with different colored squares

Description automatically generated

A group of graphs showing different sizes of data

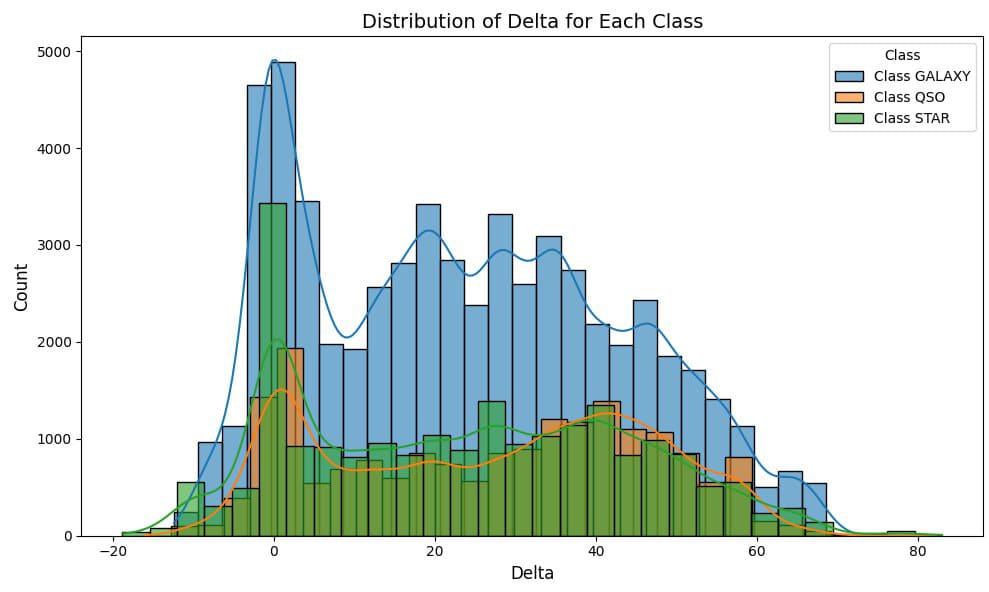
Description automatically generated with medium confidence

KDE Plots of u,g,r,i and z

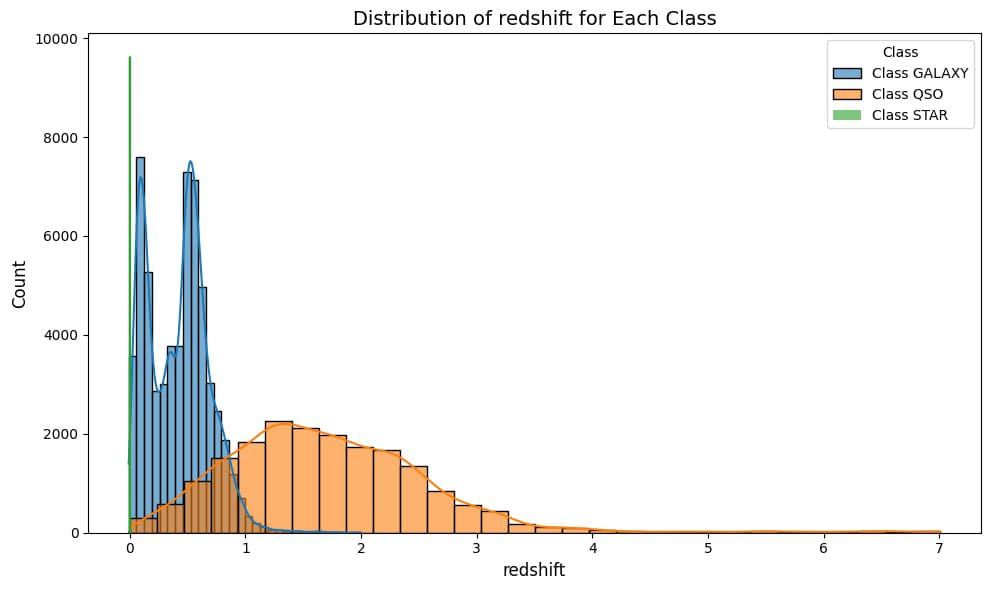
A graph of different colored bars

Description automatically generated

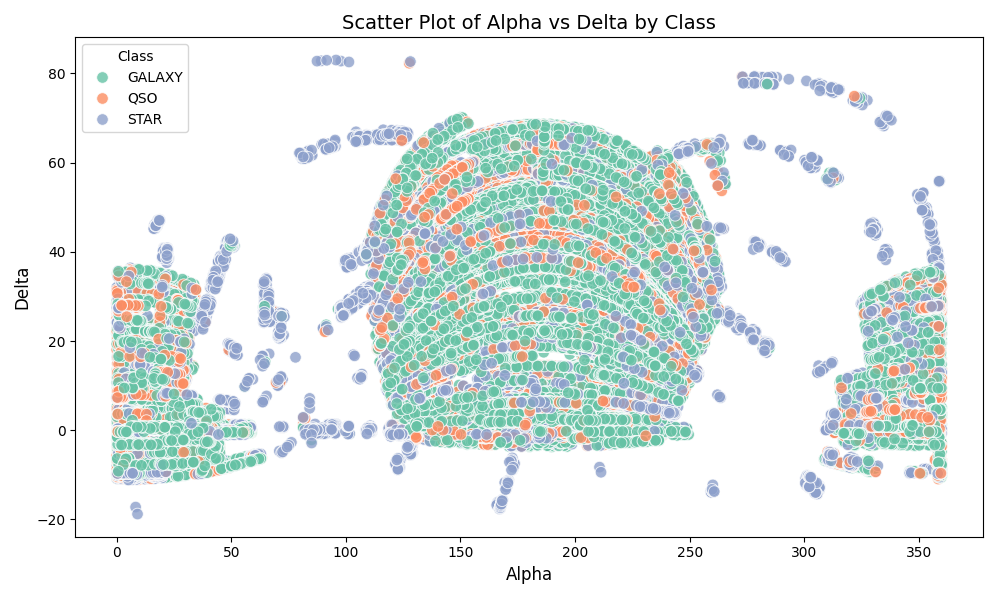
Distribution of Alpha for each class



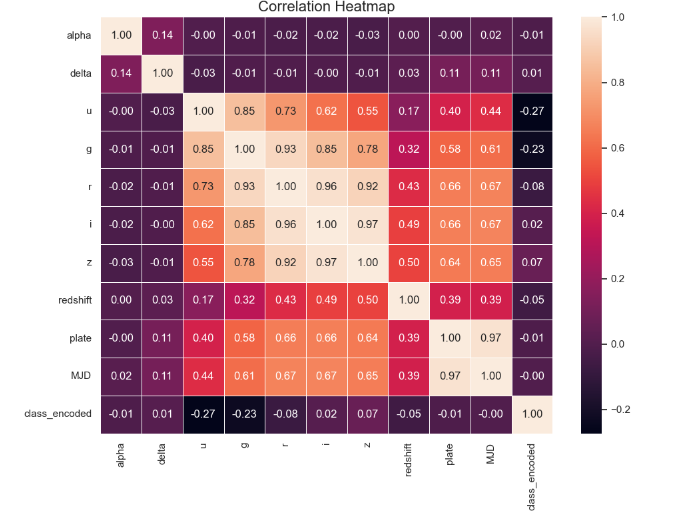
Distribution of Delta for each class



Distribution of redshift for each class



Scatter Plot of Alpha vs Delta by class



Correlation Heatmap

**Result and discussion:**

**Model Evaluation:**

The models are evaluated using cross-validation, with accuracy as the scoring metric. The models' performance is printed with their mean and standard deviation scores.

**Model Cross-Validation Scores**

1. **Random Forest**: Mean score ≈ 0.9777 ± 0.0018
2. **Decision Tree**: Mean score ≈ 0.9666 ± 0.0020
3. **XGBoost**: Mean score ≈ 0.9769 ± 0.0023
4. **Support Vector Classifier**: Mean score ≈ 0.9655 ± 0.0028

**Classification Report for Three Classes (GALAXY, QSO, STAR)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Used |  | Accuracy | F1 score | Recall | Precision |
| Random Forest | GALAXY | 0.98 | 0.98 | 0.98 | 0.98 |
| QSO | 0.94 | 0.92 | 0.96 |
| STAR | 0.99 | 0.99 | 0.99 |
| Decision Tree | GALAXY | 0.96 | 0.97 | 0.97 | 0.97 |
| QSO | 0.92 | 0.91 | 0.91 |
| STAR | 0.99 | 0.99 | 0.99 |
| SVC | GALAXY | 0.96 | 0.97 | 0.96 | 0.98 |
| QSO | 0.94 | 0.93 | 0.94 |
| STAR | 0.97 | 0.99 | 0.94 |
| XGboost | GALAXY | 0.96 | 0.97 | 0.96 | 0.98 |
| QSO | 0.95 | 0.94 | 0.95 |
| STAR | 0.97 | 0.99 | 0.95 |

These metrics indicate that all models perform well, with high precision, recall, and F1-scores, especially for the STAR and GALAXY classes. The RandomForest model achieves the highest mean cross-validation score with a small standard deviation, suggesting consistency across folds.

The results suggest that:

* The models are highly accurate and consistent, particularly RandomForest and XGBoost.
* There is minimal overfitting or underfitting, as indicated by high cross-validation scores and low standard deviations.
* The models are effective at identifying all three classes, with a slight room for improvement in recalling QSO instances.

**Conclusion :**

The machine learning classification model for distinguishing between GALAXY, QSO, and STAR classes demonstrates excellent performance. With an overall accuracy of 98% and high precision, recall, and F1-scores across all classes, particularly for STAR and GALAXY, the model effectively handles this multi-class classification task. The RandomForest model, in particular, shows both high accuracy and stability, as evidenced by its low variance in cross-validation scores. This suggests that the model generalizes well to new data and is robust to different data splits. The high F1-score, especially in the weighted average, indicates a balanced performance across classes, making this model highly reliable for practical use in classifying astronomical objects.

**Future Scope**

1. **Hyperparameter Tuning**:
   * Fine-tuning the hyperparameters for each model, particularly for RandomForest and XGBoost, could yield marginal improvements in performance. Advanced techniques like Bayesian optimization or genetic algorithms can be explored for more optimized configurations.
2. **Incorporating Additional Features**:
   * If available, including more discriminative features could improve classification performance. For example, features that capture spectral or spatial information about each object might enhance the model’s ability to distinguish between classes.
3. **Ensemble Methods**:
   * Combining multiple models (e.g., RandomForest, XGBoost, and SVC) through an ensemble approach could potentially boost performance further. Techniques like stacking, where predictions from individual models are used as inputs to a final classifier, could leverage the strengths of each model.
4. **Handling Class Imbalance**:
   * Although the model performs well, slight imbalance in recall (especially for the QSO class) can be addressed by using techniques like oversampling or undersampling, or by experimenting with different class weights to ensure even better recall for the QSO class.
5. **Model Deployment and Real-Time Classification**:
   * Deploying the model as a web or cloud-based application could enable real-time classification of astronomical objects. This would allow researchers and analysts to input data and get instant predictions, enhancing the model’s utility in practical applications.
6. **Continuous Learning and Updates**:
   * Implementing a continuous learning pipeline where the model can periodically retrain on new labeled data can ensure that it stays up-to-date with changing data patterns or new classes. This is especially useful in fields like astronomy where new discoveries and data are constantly emerging.
7. **Explainability and Interpretability**:
   * For further research, adding interpretability techniques like SHAP (Shapley Additive explanations) or LIME (Local Interpretable Model-agnostic Explanations) can help researchers understand which features influence predictions the most. This could be beneficial in fields like astrophysics where understanding the basis of predictions can lead to further scientific insights.

By pursuing these enhancements, this classification model could become an even more powerful and insightful tool in the study of celestial objects, aiding both scientific discovery and practical applications in astronomy.

**Reference:**

[1] Encyclopædia Britannica, inc. (n.d.). Stellar classification. Encyclopædia Britannica. Retrieved February

18, 2023, from http://www.britannica.com/science/stellar-classification

[2] Xiao-Qing, W., &amp; Jin-Meng, Y. (2021). Classification of star/galaxy/QSO and star spectral types

from LAMOST data release 5 with Machine Learning Approaches. Chinese Journal of Physics, 69, 303–311.

https://doi.org/10.1016/j.cjph.2020.03.008

[3] Paulycyclic&gt;. (n.d.). Paul Bacher: Notebooks expert. Kaggle. Retrieved February 18, 2023, from

http://www.kaggle.com/paulbacher/code

[4] Sklearn.tree.decisiontreeclassifier. scikit. (n.d.). Retrieved February 18, 2023, from https://scikit-

learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[5] https://medium.com/@e.woods.business/distinguishing-galaxies-from-stars-using-machine-learning-in-r-python-c4e62b2cab63

[6] Decision tree algorithm in Machine Learning - Javatpoint. www.javatpoint.com. (n.d.). Retrieved

February 19, 2023, from http://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[7] Sklearn.ensemble.randomforestclassifier. scikit. (n.d.). Retrieved February 19, 2023, from https://scikit-

learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[8] Meltzer, R. (2021, July 15). What is Random Forest? [beginner's guide + examples]. CareerFoundry.

Retrieved February 19, 2023, from https://careerfoundry.com/en/blog/data-analytics/what-is-random-forest/

[9] Chauhan, A. (2021, February 23). Random Forest classifier and its hyperparameters. Medium. Retrieved

February 19, 2023, from https://medium.com/analytics-vidhya/random-forest-classifier-and-its-

hyperparameters-8467bec755f6

[10] Sklearn.svm.SVC. scikit. (n.d.). Retrieved February 19, 2023, from https://scikit-

learn.org/stable/modules/generated/sklearn.svm.SVC.html

[11] How SVM works. (n.d.). Retrieved February 19, 2023, from http://www.ibm.com/docs/en/spss-

modeler/saas?topic=models-how-svm-works