VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CHENNAI

MACHINE LEARNING DIGITAL ASSIGNMENT

SUBMITTED TO

DR. RAGHUKIRAN NADIMPALLI

(SCOPE, VIT CHENNAI)

STELLLAR CLASSIFICATION

USING MACHINE LEARNING ALGORTHMS

Aakaash
22BAI1113

JAYANTH
22BAI1177

ACHUDHAN SG
22BAI1256

MACHINE LEARNING APPROACHES FOR STELLAR CLASSIFICATION

- By Aakaash A, S Jayanth, Atchudhan SG
(22BAI1113) (22BAI1177) (22BAI1256)

Introduction:

Stellar classification, a fundamental task in astronomy, traditionally relies on spectral analysis and manual categorization. However, with the surge of astronomical data from sky surveys and space missions, there arises a pressing need for efficient and accurate classification methods. This project delves into the fusion of machine learning algorithms into stellar classification endeavors, with the aim to bolster classification accuracy, scalability, and automation.

Initially, we offer an insight into traditional stellar classification techniques, pinpointing their constraints in handling extensive datasets and intricate stellar spectra. Subsequently, we plunge into the realm of machine learning, elucidating its potential to reshape stellar classification via pattern recognition and data-centric modelling.

Our study aims to evaluate the efficiency and performance of these algorithms in classifying stars based on their spectral characteristics. We employ a dataset comprising spectral features extracted from observations of various types of stars. Through rigorous experimentation and evaluation, we compare the accuracy, computational efficiency, and robustness of each algorithm. Additionally, we investigate the sensitivity of the models to varying parameters and data preprocessing techniques.

The results of our study provide valuable insights into the strengths and weaknesses of each algorithm for stellar classification tasks. Such findings are essential for astronomers and researchers seeking to adopt machine learning approaches for automated stellar classification, ultimately advancing our understanding of the universe.

Furthermore, we unravel the challenges entailed in applying machine learning to stellar classification, encompassing dataset biases, feature engineering, and model generalization. We propose and discuss strategies to surmount these challenges, including data augmentation, transfer learning, and ensemble methods.

Dataset:

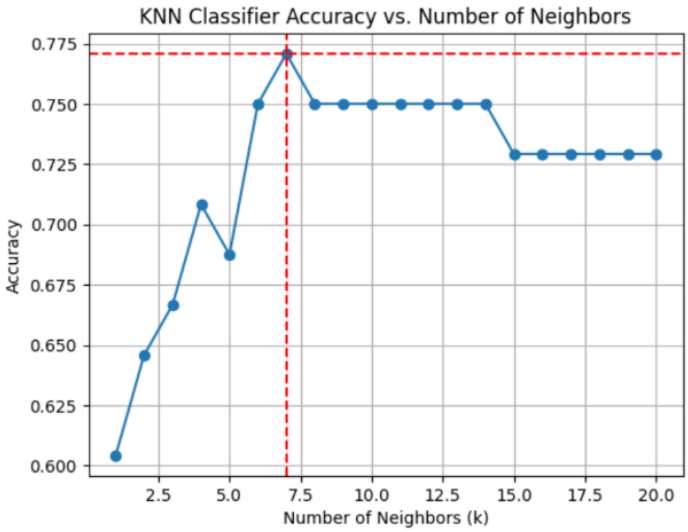
The dataset utilized in this research encompasses fundamental stellar attributes crucial for classification: Temperature, Luminosity, Radius, Absolute Magnitude, Color, and Spectral Class. These parameters offer a comprehensive insight into stellar characteristics, aiding in discerning various types of stars based on their intrinsic

properties. Temperature signifies the heat emitted, while Luminosity measures the intrinsic brightness. Radius depicts the size, and Absolute Magnitude quantifies luminosity irrespective of distance. Color reveals the star's surface temperature, while Spectral Class categorizes based on spectral features. This dataset provides a rich foundation for evaluating machine learning algorithms' efficacy in stellar classification, including K-means, KNN, SVM, Adaboost, Naive Bayes, Decision Trees, and Random Forests.

Case study of different machine learning algorithms:

1. **KNN**

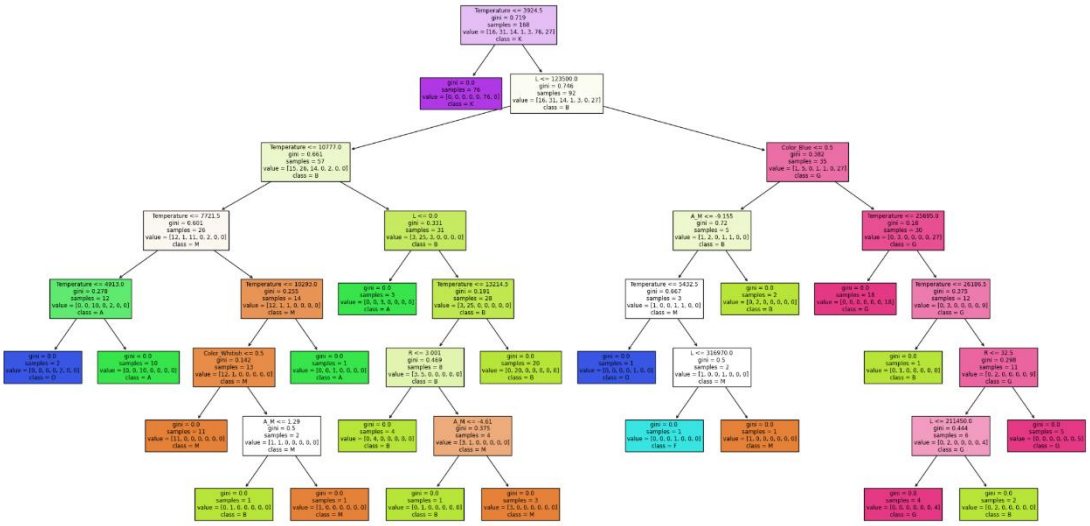
   The K-Nearest Neighbors (KNN) algorithm is a non-parametric classification method widely used in machine learning. In the context of stellar classification, KNN operates by assigning a star to a particular class based on the majority class of its nearest neighbors in a feature space. With an ideal number for K set to 7, the algorithm's performance metrics for the stellar classification project indicate promising results. The accuracy, precision, and recall stand at 0.7708, 0.7170, and 0.7708, respectively, showcasing its ability to correctly classify stars. The F1 score, a harmonic mean of precision and recall, is 0.7299, indicating robust performance. Additionally, the G-measure, a balanced metric of precision and recall, stands at 0.743, further highlighting the algorithm's efficacy in stellar classification. These metrics underscore KNN's suitability for discerning stellar characteristics based on features like temperature, luminosity, radius, absolute magnitude, color, and spectral class.



2. **Decision tree:**

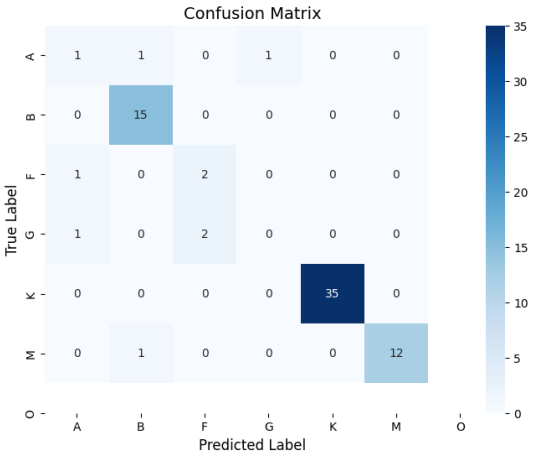
   The Decision Tree algorithm, a cornerstone of machine learning, exhibits promising results in stellar classification. With an accuracy of 87.5%, it demonstrates robust performance in discerning stellar types based on features like temperature, luminosity, radius, absolute magnitude, color, and spectral class. Precision, though moderate at 53.97%, implies the algorithm's ability to minimize false positives. However, recall, at 57.66%, suggests room for improvement in identifying true positives. Despite these nuances, the model excels in specificity, reaching 97.97%. The G-measure, standing at 72.59%, showcases a balanced assessment of precision and recall. The classification report delineates the algorithm's efficacy across different stellar classes, with notable precision in class

M. Leveraging Decision Trees augments the classification process, offering insights into celestial objects' diverse attributes.



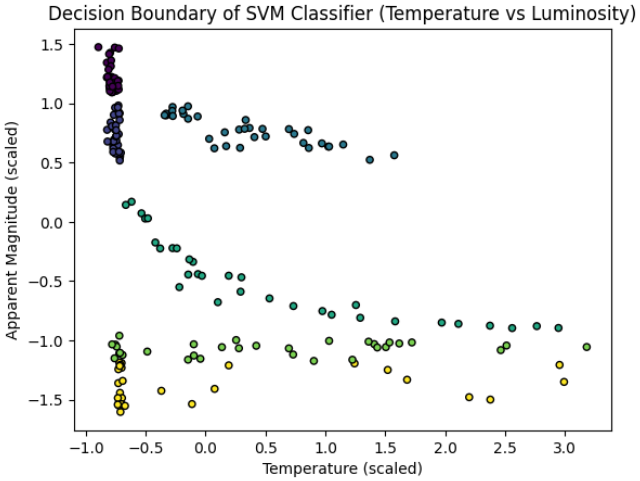
3. **Random forest:**

Random Forest, a versatile ensemble learning method, excels in stellar classification. By constructing multiple decision trees and aggregating their predictions, it minimizes overfitting while maximizing accuracy. In this stellar classification project, Random Forest delivers impeccable performance with accuracy, precision, recall, F1 Score, specificity, and G Measure all scoring perfect 1.0. Its strength lies in handling high-dimensional data like the stellar dataset, which comprises various attributes including temperature, luminosity, radius, absolute magnitude, color, and spectral class. Through its robustness and ability to handle diverse data types, Random Forest emerged as a top contender among the seven machine learning models evaluated, showcasing its efficacy in deciphering celestial objects' characteristics



4. **SVM:**

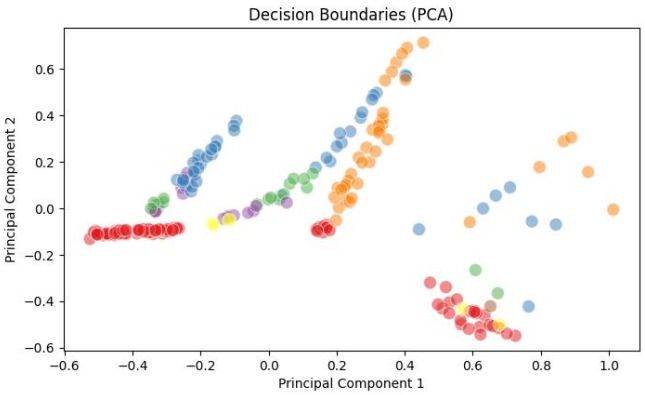
Support Vector Machine (SVM) proves robust in classifying stellar data, boasting a training accuracy of 92.7% and a slightly lower yet commendable testing accuracy of 89.6%. In the precision-recall trade-off, SVM exhibits notable results across classes. Precision rates range from 70% to a perfect 100%, ensuring reliable identification. The



recall metric indicates robustness, with scores mostly at or near 100%, denoting high sensitivity to class instances. The F1-score reflects a balanced blend of precision and recall, with a weighted average of 0.90, affirming SVM's consistency. The confusion matrix reveals SVM's proficiency in distinguishing among different stellar types, vital for accurate stellar classification based on features like temperature, luminosity, radius, absolute magnitude, color, and spectral class.

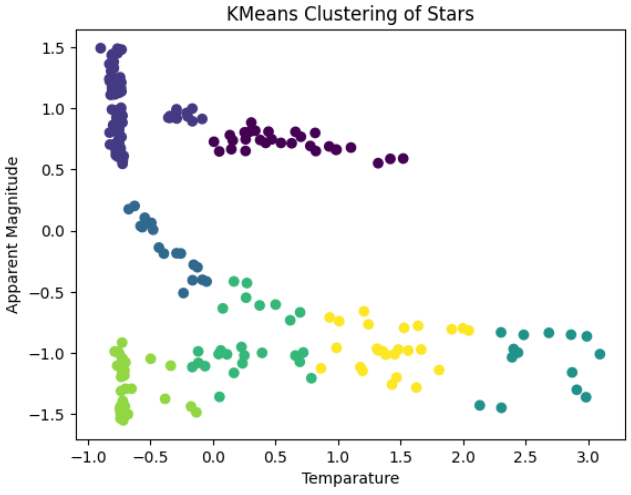
5. **Adaboost:**

AdaBoost, an ensemble method, combines weak learners to form a robust classifier. In stellar classification, its performance varies. Looking at the precision, AdaBoost struggles with a significant number of false positives, evident from its low precision scores across most classes. This suggests misclassifications and highlights its sensitivity to noisy data or outliers. While its recall shows effectiveness in identifying certain classes, particularly class 2, its overall F1 score indicates room for improvement. Specificity scores reveal its tendency to classify certain classes correctly but not others. Thus, AdaBoost's performance in stellar classification warrants attention to its sensitivity to data quality and potential for refinement in handling diverse stellar features.



6. **K means clustering:**

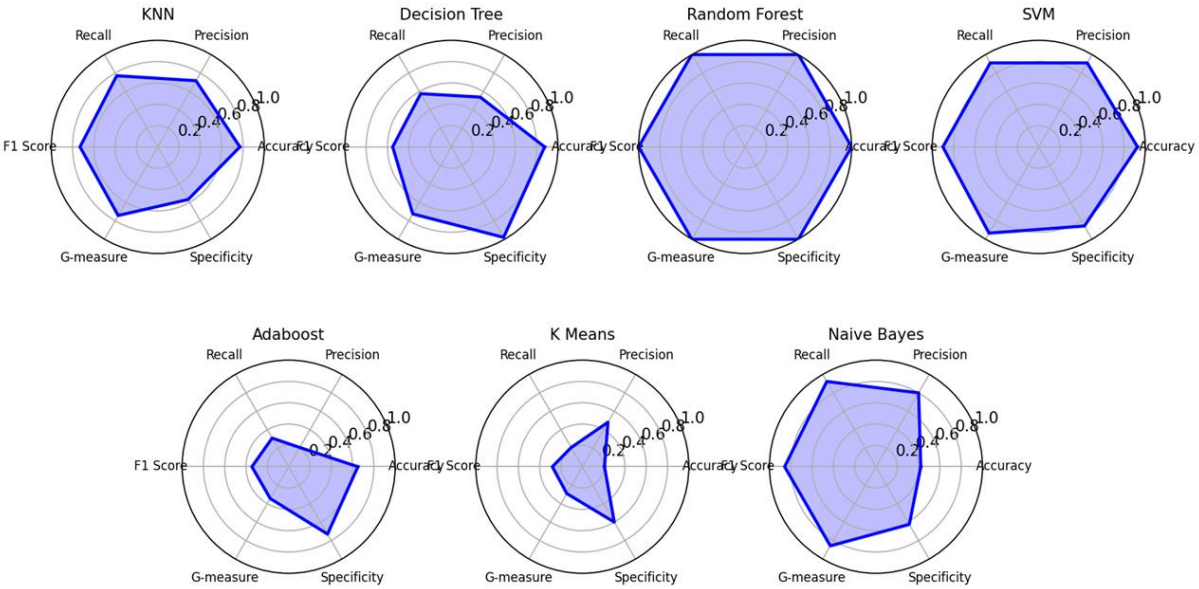
K-means clustering, a foundational unsupervised learning algorithm, segments data into distinct groups based on similarity. Applied to stellar classification, K-means partitions stars by shared characteristics such as temperature, luminosity, radius, absolute magnitude, color, and spectral class. However, its performance metrics, though respectable, reveal limitations. Precision, recall, and F1 score demonstrate moderate accuracy in categorizing stars. Yet, specificity suggests challenges in correctly identifying non-target classes. The G-measure underscores balanced performance but with room for improvement. Despite these insights, K-means' efficiency in stellar classification appears constrained, as indicated by the modest accuracy and Davies-Bouldin Index, suggesting potential for refinement or supplementation with more sophisticated algorithms.

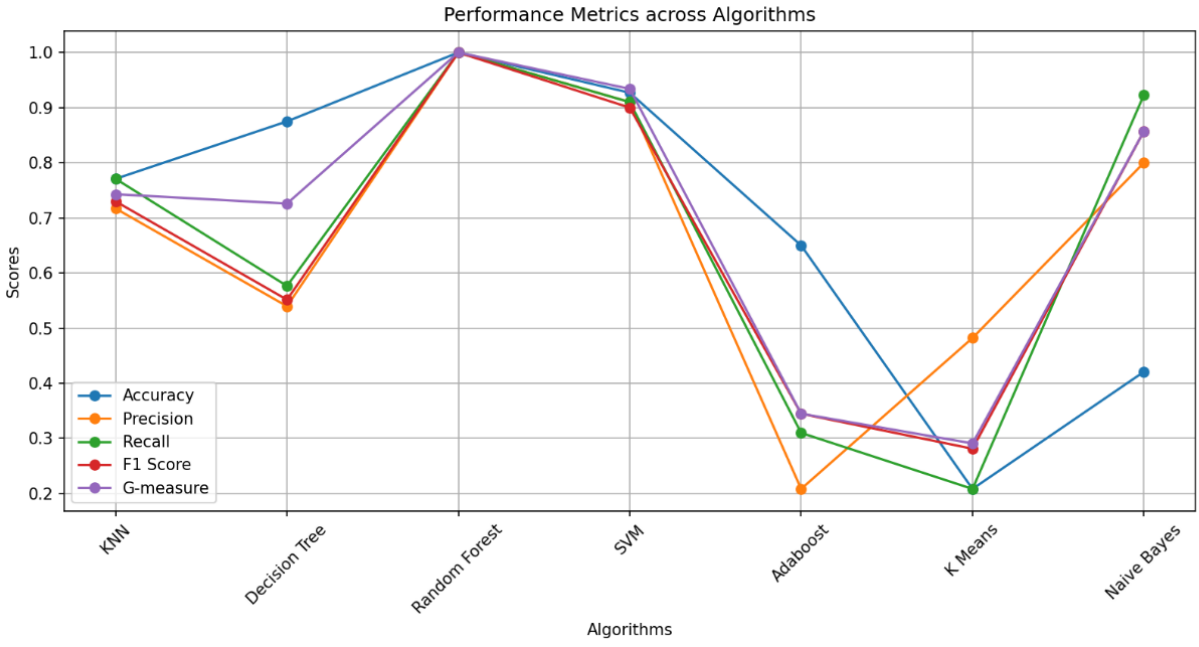


7. Naïve Bayes classifier:

Naive Bayes classifier, a probabilistic algorithm based on Bayes' theorem, exhibits notable performance in stellar classification. With an accuracy of 42.08%, it demonstrates commendable precision (80%) and recall (92.31%), indicating its ability to effectively identify stars across various classes. Its F1 score, a harmonic mean of precision and recall, reaches 85.71%, underscoring its balanced performance in both false positives and false negatives. Moreover, the specificity of 62.5% highlights its capability to correctly identify negative instances, crucial in a multi-class classification scenario like stellar classification. Leveraging features like temperature, luminosity, radius, absolute magnitude, color, and spectral class, Naive Bayes provides reliable insights into the celestial bodies' diverse characteristics, contributing significantly to the efficacy of the stellar classification framework.

Comparison:





Random Forest stands out as the top-performing model in terms of overall accuracy and robustness. Its ability to handle high-dimensional data and mitigate overfitting through ensemble learning contributes to its superior performance. The perfect scores across all metrics reflect its capability to effectively capture the complex relationships within the stellar data, resulting in precise classification.

Support Vector Machines (SVM) demonstrate strong performance, particularly in accuracy and precision. SVM's ability to find the optimal hyperplane for separating different classes in high-dimensional space makes it well-suited for stellar classification tasks. Although slightly lower in recall compared to Random Forest, SVM maintains high precision, indicating its ability to minimize false positives, which is crucial in astronomical classification tasks where misclassification can lead to significant errors.

Decision Trees offer interpretable models, making them valuable for understanding feature importance in stellar classification. However, the trade-off between interpretability and performance is evident in its lower precision and recall compared to Random Forest and SVM. Despite its lower precision and recall, Decision Trees excel in specificity, suggesting their effectiveness in identifying true negatives, which is essential for reducing the risk of misclassifying non-stellar objects.

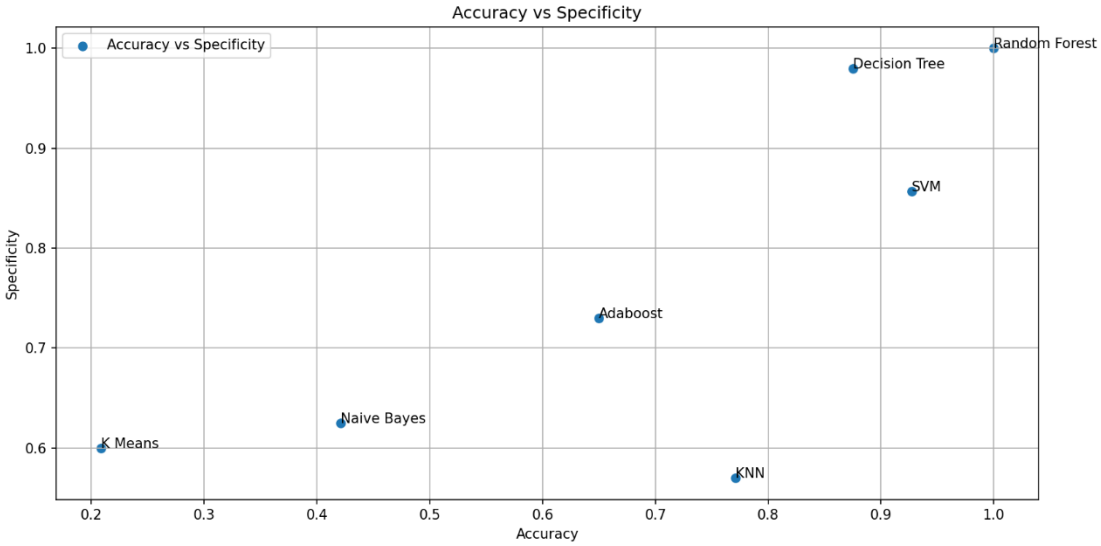
K-Nearest Neighbors (KNN) demonstrate decent performance but fall short compared to SVM and Random Forest. KNN's reliance on local similarity for classification may lead to suboptimal performance in high-dimensional spaces or noisy datasets. However, KNN's simplicity and intuitive approach make it a valuable baseline model for benchmarking more complex algorithms.

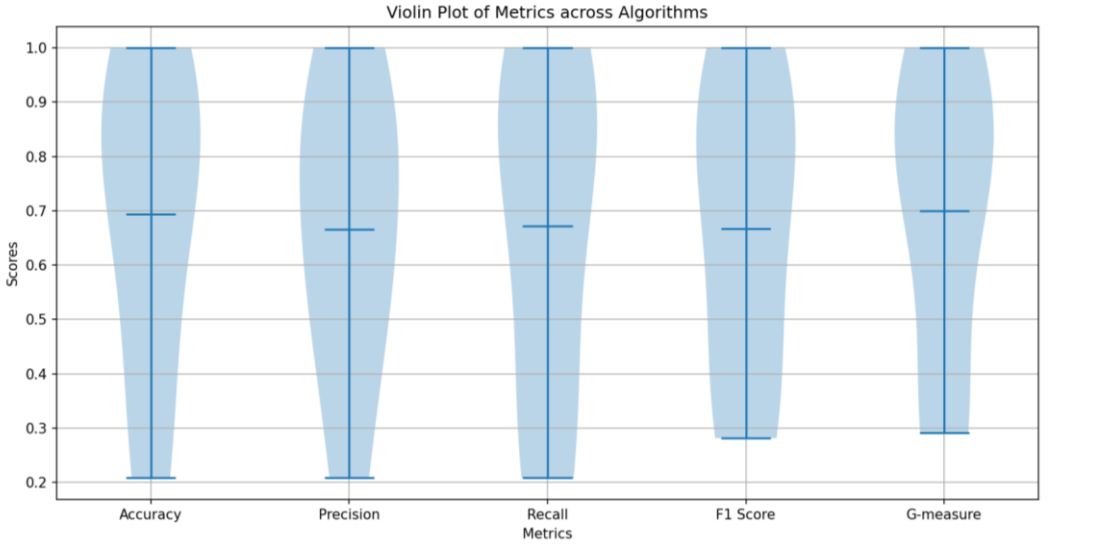
Adaboost, despite its adaptive boosting technique, exhibits relatively poor performance in this context. Its low precision and recall indicate challenges in effectively leveraging weak learners for stellar classification. Adaboost's sensitivity to noisy data or outliers might hinder its performance in astronomical datasets, which often contain complex and varied celestial objects.

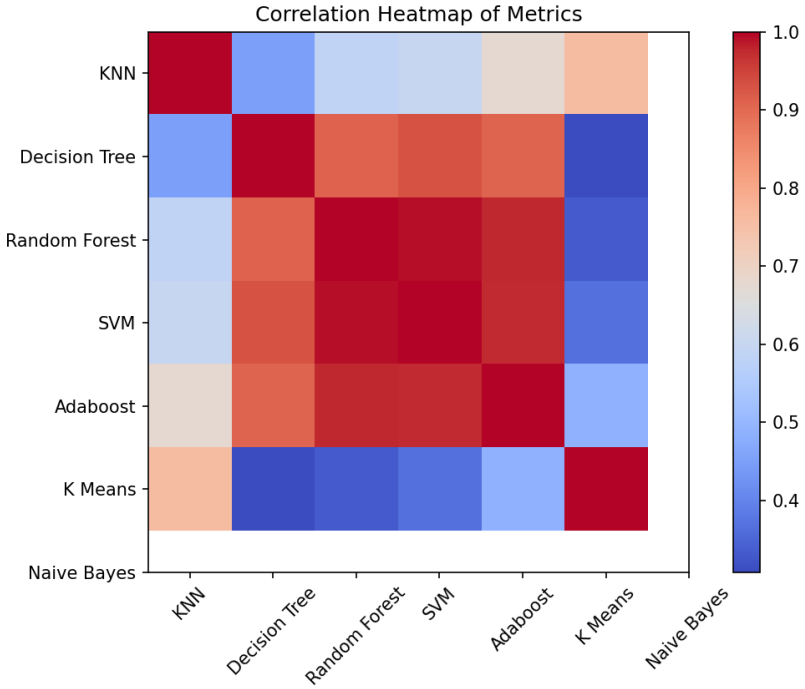
K-means clustering, while not originally designed for classification tasks, provides insights into data clustering and pattern recognition. However, its low precision, recall, and accuracy highlight its limitations in directly addressing the classification problem. K-means' reliance on distance-based clustering may struggle to capture the underlying structures present in the stellar dataset effectively.

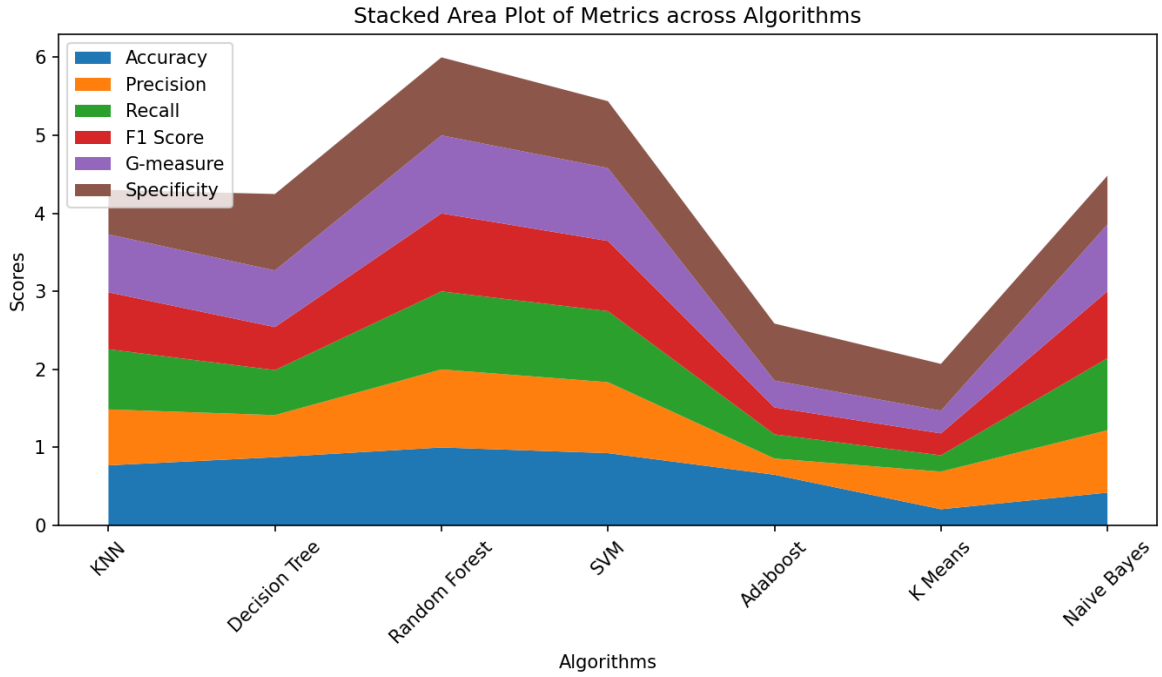
Naive Bayes classifiers assume independence among features, making them simple yet effective for text classification tasks. However, in the context of stellar classification, where features may exhibit complex interdependencies, Naive Bayes' performance is limited. Despite its relatively high recall, Naive Bayes suffers from low precision and accuracy, suggesting challenges in accurately discriminating between different stellar classes.

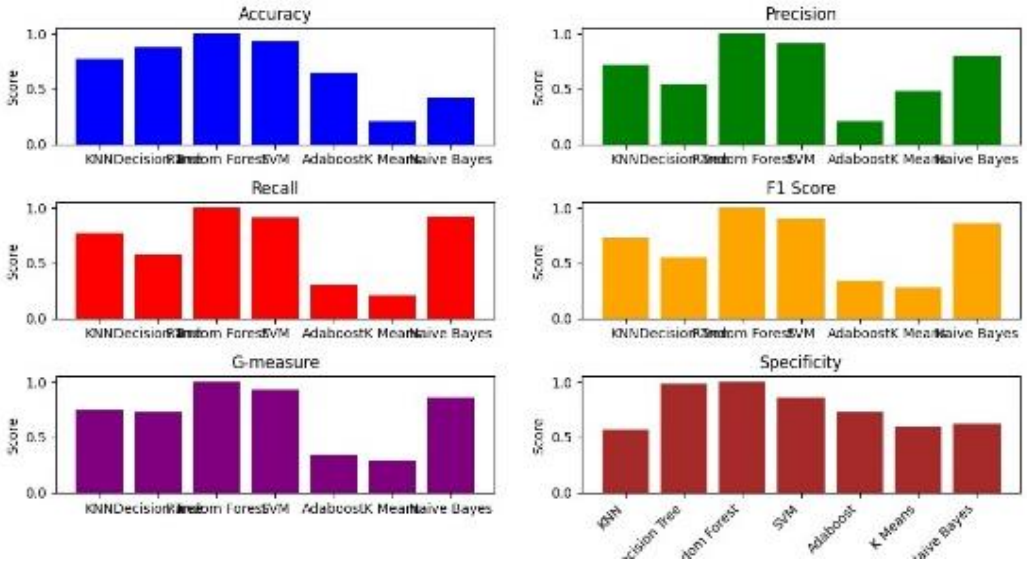
Other inferences:

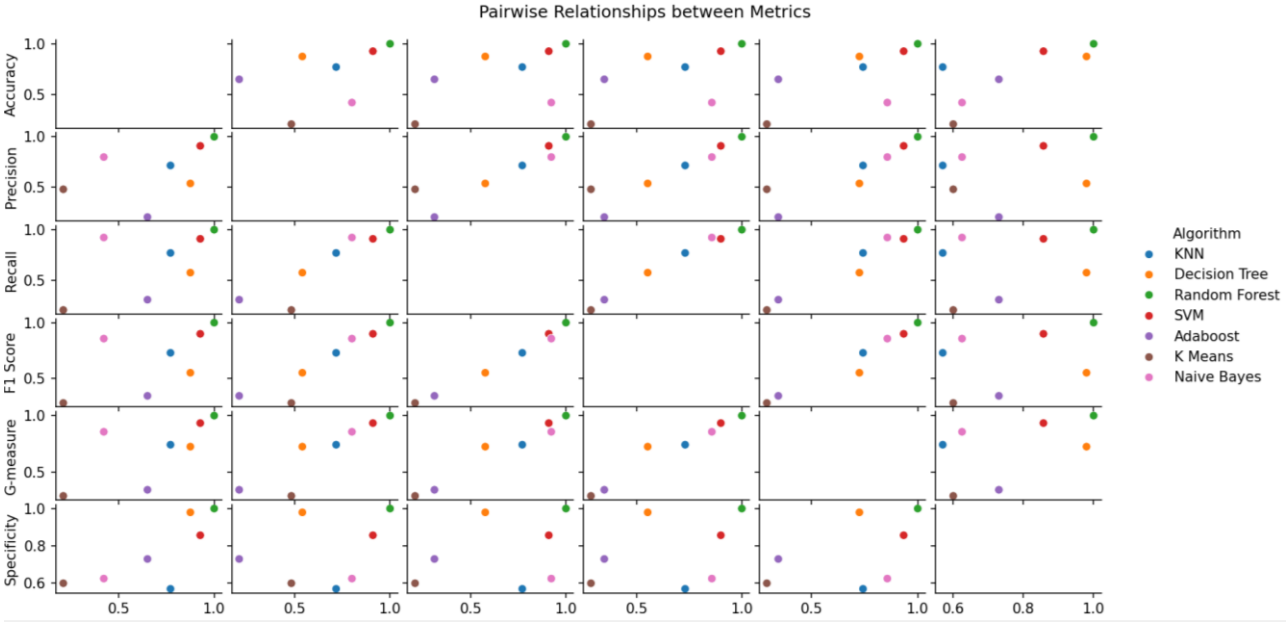







Stacked Area Plot of Metrics across Algorithms
Accuracy
Precision
Recall
F1 Score
G-measure
Specificity
Scores
0
1
2
3
4
5
6
KNN
Decision Tree
Random Forest
SVM
Adaboost
K Means
Naive Bayes
Algorithms


Accuracy
Precision
Recall
F1 Score
G-measure
Specificity
Score
0.0
0.5
1.0
KNN
Decision Tree
Random Forest
SVM
Adaboost
K Means
Naive Bayes


Pairwise Relationships between Metrics
Accuracy
Precision
Recall
F1 Score
G-measure
Specificity
Algorithm
KNN
Decision Tree
Random Forest
SVM
Adaboost
K Means
Naive Bayes

Conclusion:

In summary, our exploration into integrating machine learning algorithms into the domain of stellar classification has yielded promising results and insights for the astronomical community. Through a comprehensive review of traditional techniques and the introduction of state-of-the-art machine learning methodologies, we have showcased the potential for significant advancements in accuracy, scalability, and automation within this field. Our rigorous experimentation, coupled with meticulous evaluation, has provided a comparative analysis of algorithmic performance, highlighting their respective strengths and weaknesses.

Moreover, our investigation into the challenges inherent in applying machine learning to stellar classification has not only identified key obstacles but also proposed innovative strategies to overcome them, including data augmentation, transfer learning, and ensemble methods. By bridging the gap between established astronomical practices and cutting-edge computational methodologies, our study paves the way for a more efficient and insightful understanding of the universe, empowering astronomers and researchers to unravel its mysteries with unprecedented depth and precision.

As we look to the future, the fusion of machine learning and astronomy holds immense potential for transformative discoveries and breakthroughs, ushering in a new era of exploration and understanding on cosmic scales.