1 Business Understanding

The High Time Resolution Universe Survey (HTRU) [1] is a large-scale radio astronomy initiative aimed at identifying pulsars—a rare, highly magnetized, rotating neutron star. Due to the immense data volume and the rarity of true pulsars, most candidate detections are noise or RFI (Radio Frequency Interference), requiring significant manual review. Machine learning offers a scalable solution to automate the classification of candidates, reducing time-to-discovery and researcher workload.

This study has two main tasks. The first is to build classification models to predict whether a signal is a pulsar or radio noise. The second task is to utilize clustering techniques to analyze and extract meaningful patterns from features in the pulsar candidate data, with the goal of discovering subgroups of pulsars based on their features.

The project is successful if it enables astronomers and researchers to more effectively identify and investigate true pulsar candidates, reducing the time and effort spent on false positives while ensuring real discoveries are not missed. The project focuses on two actionable metrics to determine the success of modelling efforts. First, the model helps scientists find nearly all of the true pulsars in the dataset. Furthermore, insights from the model or data analysis would contribute to further scientific investigations into the properties of various types of pulsars.

2 Data Understanding

The dataset comprises 17,898 pulsar candidate observations, each described by 8 continuous numerical features and a binary class label, where 0 is spurious noise (noise or RFI) and 1 is a real pulsar. Out of the 17,898 instances, there are 16,259 spurious examples (90.8%) and 1,639 real pulsars (9.2%). The dataset does not contain any missing data, and there are no erroneous values.

The dataset's 8 features are grouped into two sets, with each set capturing distinct signal characteristics. The first set of features, named "Profile" in the dataset, describe the summed signal strength over time across multiple rotations of the pulsar. The feature compresses the time-series data into a shape that emphasizes periodicity and consistency of the pulse. The second set of features, named "DM" in the dataset, is the signal to noise ratio (SNR) across different dispersion measure values. Dispersion is how radio waves are delayed depending on frequency, and it is a key characteristic of pulsar signals due to their journey through space. These features provide insight into how the signal strength behaves under varying astrophysical conditions. For each set of features, the following measures of central tendencies are recorded: mean, standard deviation, excess kurtosis, and skewness. Features such as DM-SNR kurtosis and skewness tend to show stronger separation between classes in visualizations.

3 Data Preparation

The HTRU dataset [1] was loaded using the Pandas python package. Since the dataset does not contain any missing data or erroneous values, there was no need for data cleaning. When the dataset was imported into Python, the feature columns were correctly in float format while the target variable was in integer format. Both minmax and standard normalization were done using the Scikit package, and these data transformations were compared against the original dataset during modeling.

4 Modeling

Four different classification models that were deemed applicable for binary classification were tested: K-nearest-neighbors (KNN), Gaussian Naive Bayes, Decision Tree Classification, and Random Forest. For both KNN and Gaussian Naive Bayes, the results of both the standard and minmax scaled datasets were compared against the original dataset. Hyperparameter tuning was performed for the KNN and decision tree models. For KNN, values of k from 5 to 20 were tested, while for the decision trees, values for the max depth from 3 to 10 were tested.

Metric	**KNN (MinMax, $k = 5$)**	**GNB (Standard)**	**Decision Tree (depth = 4)**	**Random Forest (depth = 10)**
F1 Score	0.8740	0.7340	0.8762	0.8757
Precision	0.9160	0.6451	0.9102	0.9227
Recall	0.8356	0.8514	0.8446	0.8333
Accuracy	0.9801	0.9490	0.9803	0.9804

Table 1: Classification Model Performance

We tested two clustering methods: K-Means and Gaussian Mixture Models (GMM) on the standardized 8 numerical features to explore their natural groupings without using class labels.

For K-Means clustering, we conducted a grid search across cluster values $k = 2$ to $k = 10$, evaluating each model based on the Silhouette Score, which measures how well-separated and internally coherent the clusters are. The results indicated that $k = 2$ yielded the highest silhouette score.

In parallel, we applied Gaussian Mixture Models over the same range of cluster numbers, using both the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to evaluate model quality. These criteria penalize model complexity and help detect the "elbow point" where additional clusters no longer substantially improve the model. For GMM, this elbow occurred at $k = 5$. Since GMM is a soft clustering model, the data points are assigned to the component based on the highest probability.

5 Evaluation

The best models for each classification model were chosen based on the F1 score metric, which takes into account both precision and recall. Since the class labels are unbalanced, accuracy was not chosen as the primary metric for model performance. It is of equal importance to astronomers that the model is able to correctly identify all the true pulsars as well as having the majority of the predicted positive pulsars be true pulsars. The results of the classification models using the optimal parameters can be seen in table 1. The random forest model with a max depth of 10 achieved the highest F1 score of 0.8757, with the confusion matrix seen in figure 3. The random forest model also achieved the highest precision at 0.9227. The Gaussian Naive Bayes model with standard scaling achieved the highest recall at 0.8514.

For clustering methods, $k = 2$ for KMeans suggests that the data naturally separates into two primary groups. This finding aligns well with the binary nature of the ground truth labels (pulsar vs. non-pulsar). Because of the binary classification produced by KMeans, we are able to test the accuracy of the model. We observe that the model achieved an accuracy of 0.9365 and an F1 score of 0.6928. The GMM indicated that a more nuanced segmentation of the data—beyond the binary split—was appropriate. This could reflect subtypes of spurious signals, or variability in pulsar emission patterns not captured by the binary label. However, given the nature of the results from both models, we suggest KMeans as the more appropriate model.

6 Appendix

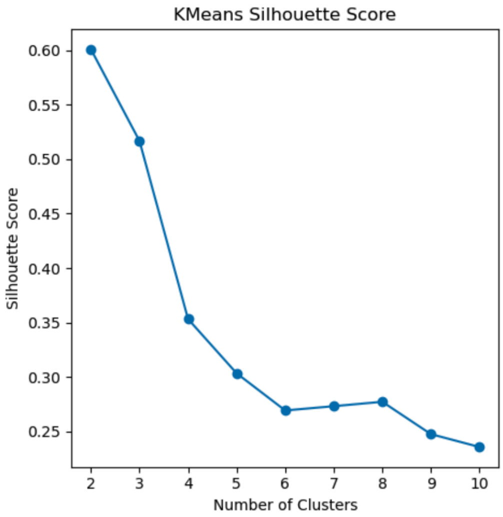
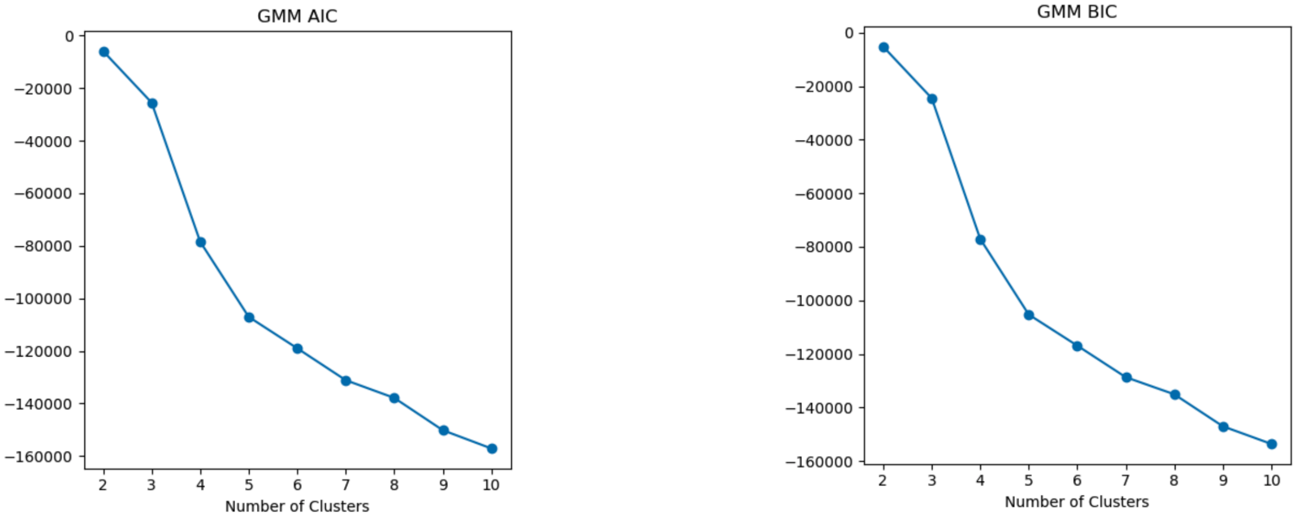


Figure 1: KMeans Silhouette Score for 2 to 10 clusters inclusive.



(a) GMM AIC scores for 2 to 10 clusters inclusive.

(b) GMM BIC scores for 2 to 10 clusters inclusive.

Figure 2: GMM AIC/BIC scores over increasing number of Gaussian components

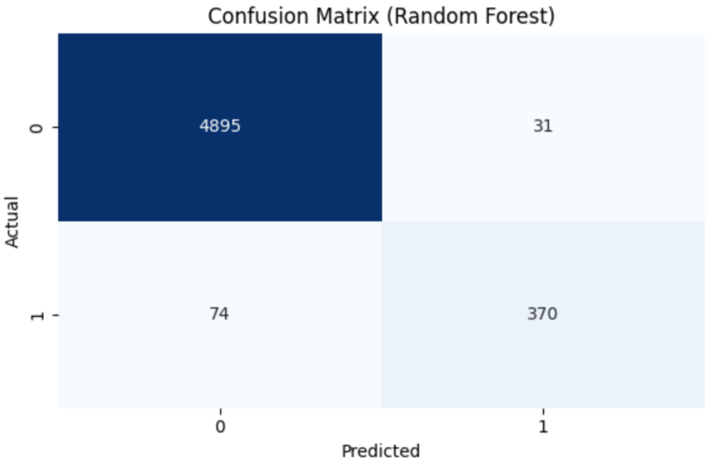


Figure 3: Confusion Matrix of the Random Forest Model

References

[1] Robert J. Lyon, Benjamin W. Stappers, Sally Cooper, J. M. Brooke, and Joshua D. Knowles. Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach. *Monthly Notices of the Royal Astronomical Society*, 459:1104–1123, 2016.