Introduction

Context of Pulsar Research

Pulsars are a rare and fascinating type of neutron star, the collapsed remnants of massive stars that have undergone a supernova explosion. These celestial bodies emit periodic radio signals as they spin, creating detectable pulses of electromagnetic radiation. First discovered in 1967, pulsars have become subjects of interest in astrophysics for studying fundamental properties of the universe.

Their properties of timing, including some being more accurate than an atomic clock. allows researchers to test theories of space-time, particularly Einstein's theory of general relativity. Pulsars also provide insight into the interstellar medium by analyzing how their signals are distorted as they travel through space. Additionally, they help scientists probe extreme states of matter, such as those found in neutron star cores which are otherwise inaccessible.

HTRU Survey

The High Time Resolution Universe Survey (HTRU) was initiated to expand our understanding of pulsars by systematically identifying new candidates. This survey scans the sky using high-sensitivity radio telescopes and generates vast datasets of potential pulsar signals. However, the challenge lies in distinguishing genuine pulsars from spurious signals caused by radio frequency interference (RFI) and other noise sources.

RFI contaminates the data, producing signals that mimic pulsars but originate from human-made sources or other astrophysical phenomena. This makes manual classification time-consuming and error-prone, necessitating the development of automated classification methods.

Objective

This project aims to apply machine learning techniques to classify pulsar candidates in the HTRU2 dataset, a benchmark dataset derived from the HTRU survey. The goal is to address the binary classification task of distinguishing real pulsars (positive class) from spurious signals (negative class). By leveraging statistical features extracted from the data, machine learning can enhance the efficiency and accuracy of pulsar detection, reducing the reliance on manual verification.

Data

Dataset Overview

The HTRU2 dataset comprises a total of 17,898 instances. These instances are split into two categories:

- **Positive Class**: 1,639 examples of real pulsars.
- **Negative Class**: 16,259 examples of spurious signals.

Each instance is described by 8 continuous features representing statistical properties of the signal. These features are divided into two groups:

1. **Integrated Pulse Profile Features**: Statistical metrics derived from the integrated pulse profile.
2. **DM-SNR Curve Features**: Statistical metrics derived from the dispersion measure (DM) signal-to-noise ratio (SNR) curve.

Feature Description

The eight continuous features include:

1. **Mean**, **Standard Deviation**, **Skewness**, and **Kurtosis** of the integrated pulse profile.
2. **Mean**, **Standard Deviation**, **Skewness**, and **Kurtosis** of the DM-SNR curve.

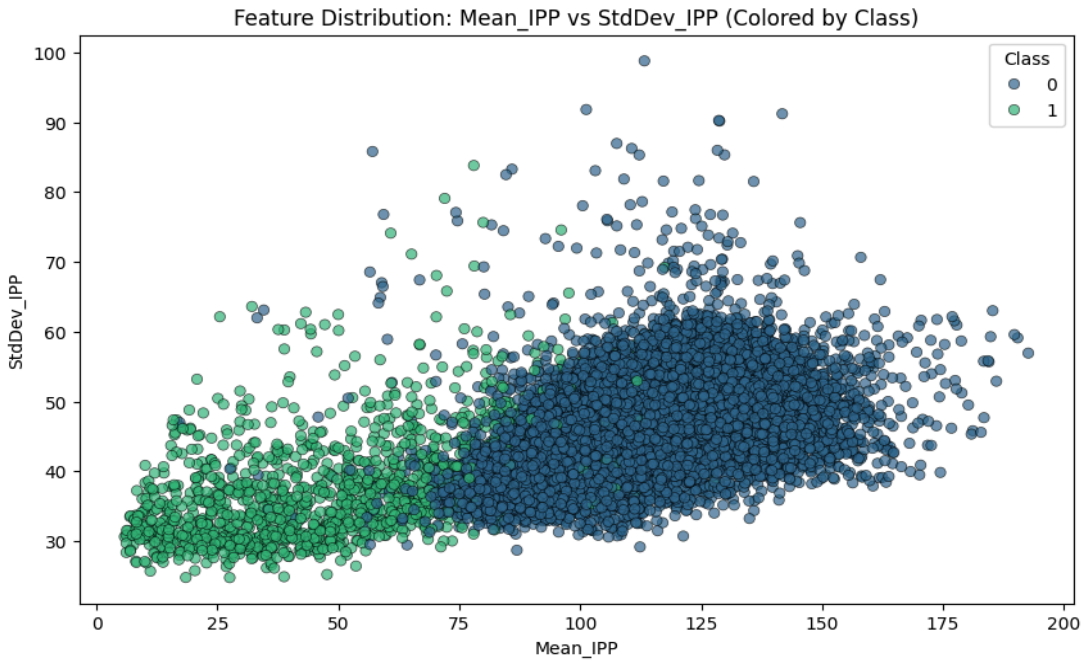
These features provide a statistical summary of the candidate signal's properties which can help distinguish between real pulsars and noise.

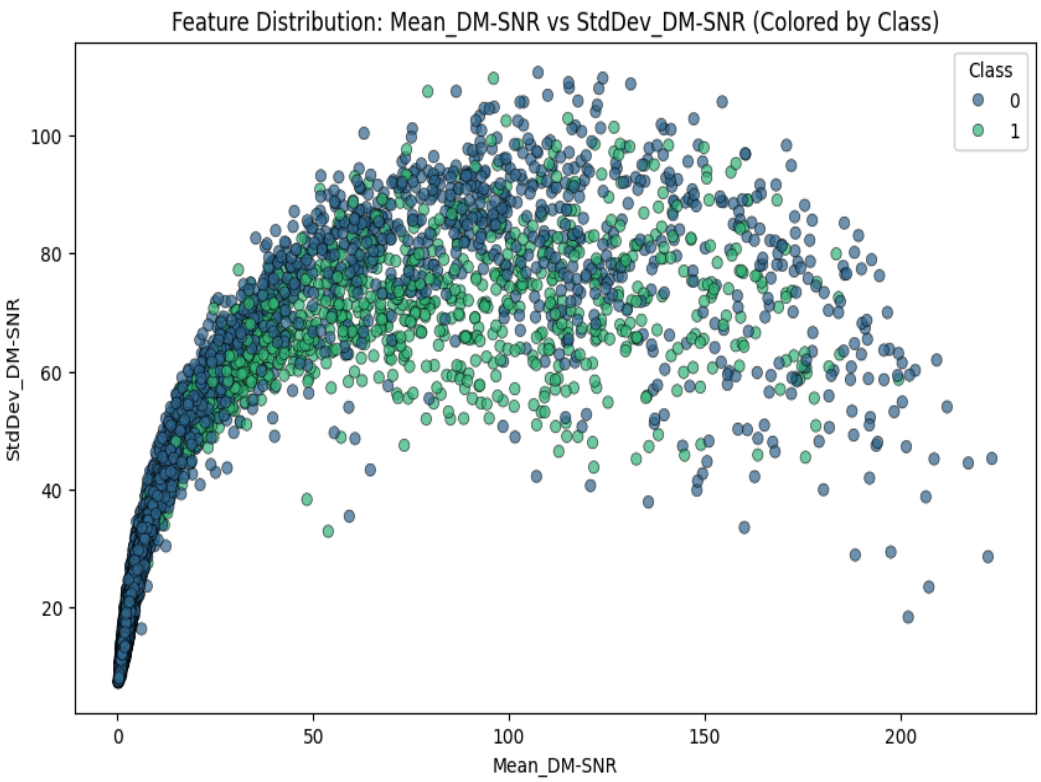
Data Structure

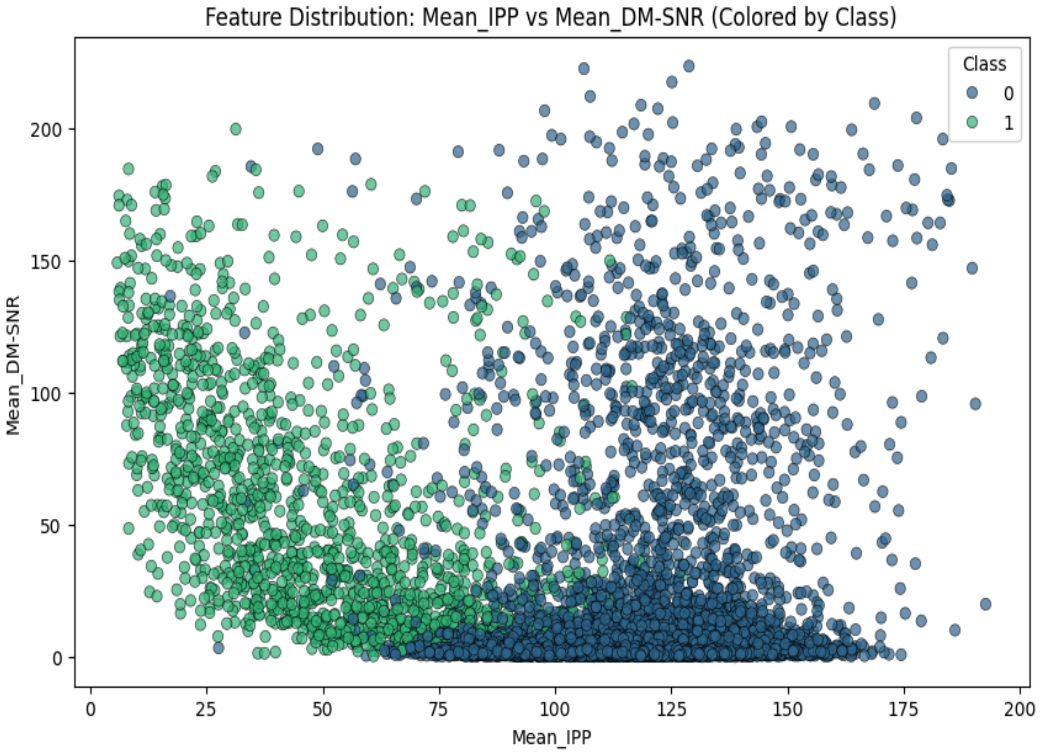
The dataset does not include positional or astrophysical details but focuses on numerical summaries of candidate signals. Each instance is labeled as:

- **"1"**: A real pulsar.
- **"0"**: A spurious signal.

This numerical representation makes the dataset suitable for machine learning algorithms, which can use these features for classification.




Feature Distribution: Mean_DM-SNR vs StdDev_DM-SNR (Colored by Class)
Class
0
1
StdDev_DM-SNR
Mean_DM-SNR
0
50
100
150
200
20
40
60
80
100


Feature Distribution: Mean_IPP vs Mean_DM-SNR (Colored by Class)
Class
0
1
Mean_DM-SNR
Mean_IPP
0
25
50
75
100
125
150
175
200
0
50
100
150
200

Addressing Class Imbalance

One of the primary challenges in this dataset is its imbalance. With 16,259 negative samples and only 1,639 positive samples, the dataset is heavily skewed toward the negative class. This imbalance can cause machine learning models to favor the majority class, leading to poor performance in identifying real pulsars.

To mitigate this, random undersampling is applied:

- The majority class (spurious signals) is reduced to match the size of the minority class (real pulsars).
- This results in a balanced dataset of 3,278 total samples, with 1,639 positive samples and 1,639 negative samples.

Trade-Off of Undersampling

While undersampling ensures that the dataset is balanced, it reduces the total number of samples available for training. This trade-off sacrifices some data diversity in exchange for balanced class representation, which can significantly improve the model's ability to detect the minority class without bias.

Model Design

Preprocessing:

To ensure optimal model performance, I applied several preprocessing steps to the dataset:

- **Feature Standardization:**
  I standardized the features using the StandardScaler function to ensure all input variables are on the same scale. This step is essential for models like Logistic Regression and Support Vector Machines (SVM), where features with large ranges can dominate the learning process.
- **Class Imbalance Handling:**
  Given the class imbalance in the dataset (with pulsars being the minority class), I performed undersampling of the majority class (non-pulsar) to balance the dataset. This helps prevent models from being biased toward the majority class and allows them to better capture patterns in the minority class (the pulsars).

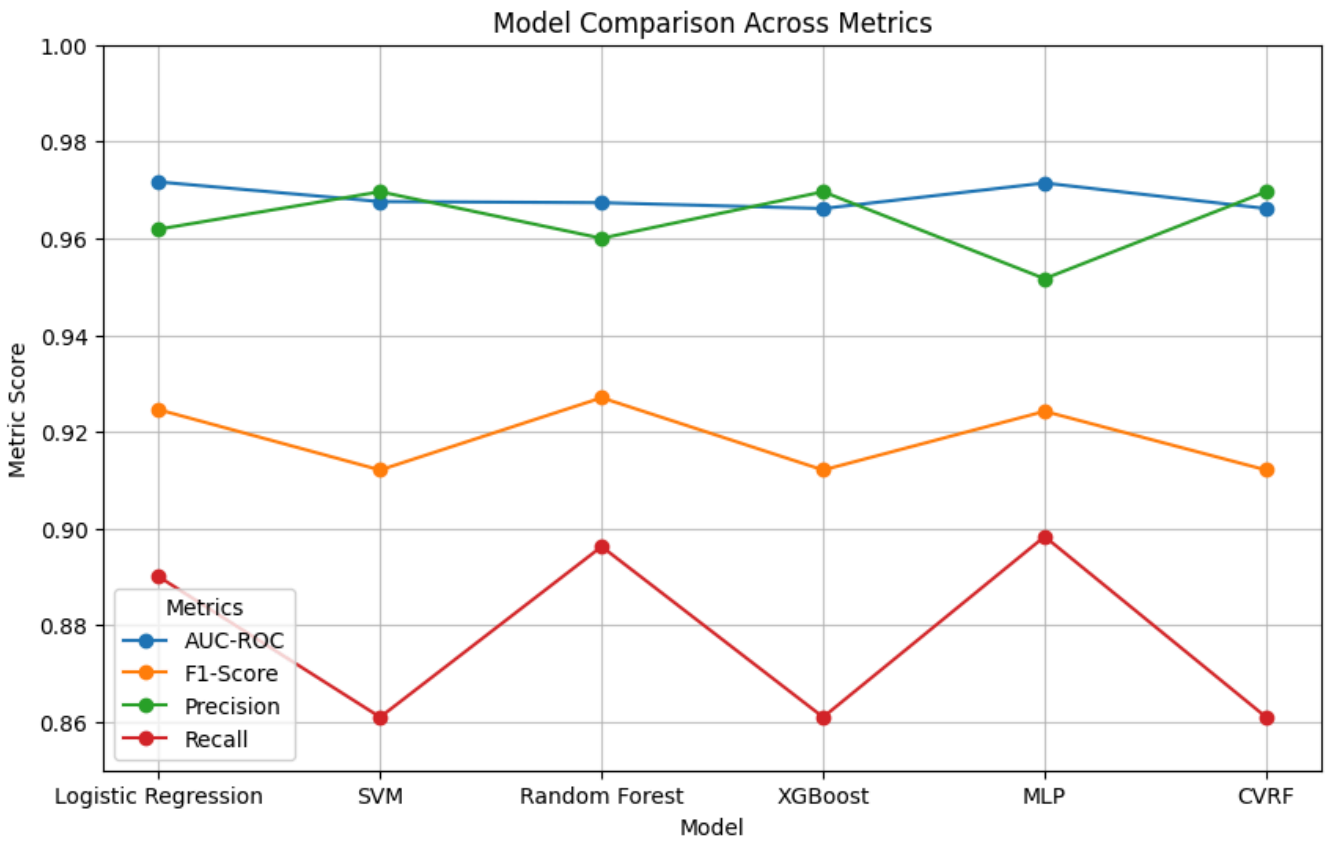
Model Selection:

I started by testing simpler models and progressively moved to more complex ones to determine which performed best:

- **Logistic Regression:**
  Baseline model for binary classification. It assumes a linear decision boundary and works well when the classes are linearly separable.

- **Support Vector Machines (SVM):**
  Performs well in high-dimensional spaces and is suitable for finding the optimal hyperplane that maximizes the margin between classes. I used the radial basis function (RBF) kernel to capture non-linear relationships.
- **Random Forest:** Ensemble learning method used for classification and regression tasks by making multiple decision trees with bagging and random feature selection.
- **XGBoost:** Gradient boosting with building decision trees sequentially, where each tree corrects the errors of its predecessor.
- **Neural Networks (MLP):**
  Multi-Layer Perceptron (MLP) neural network model, which consists of multiple layers of neurons capable of learning complex non-linear relationships.

Model	AUC-ROC	F1-Score	Precision	Recall
Logistic Regression	0.971657	0.924569	0.961883	0.890041
Support Vector Machines (SVM)	0.967590	0.912088	0.969626	0.860996
Random Forest	0.967361	0.927039	0.960000	0.896266
XGBoost	0.966152	0.912088	0.969626	0.860996
Multi-Layer Perceptron (MLP)	0.971430	0.924226	0.951648	0.898340
CVRF	0.966152	0.912088	0.969626	0.860996



Model Comparison:

Based on the evaluation metrics, Logistic Regression and MLP achieved the highest AUC-ROC and F1-Score, making them strong contenders for this task. However, Random Forest emerged as the most balanced model, with the highest F1-Score (0.9270) and solid precision and recall values, indicating its ability to correctly identify pulsars while maintaining low false positive rates.

XGBoost and SVM, though capable models, performed slightly worse than Random Forest in terms of AUC-ROC and recall.

Thus, Random Forest is the recommended model, while Logistic Regression and MLP can serve as viable alternatives if specific metrics (like AUC-ROC or recall) are prioritized.

- **Random Forest (default)**: Performed well because it naturally handled the data's complexity without overfitting
- **RandomizedSearchCV-Optimized Random Forest (CVRF)**: The hyperparameter tuning led to overfitting, making it too specific to the training data, which hurt its performance
- **XGBoost**: XGBoost is sensitive to hyperparameters and may have overfitted as well

Random Forest Model: Comprehensive Performance Analysis

Overall Performance Metrics

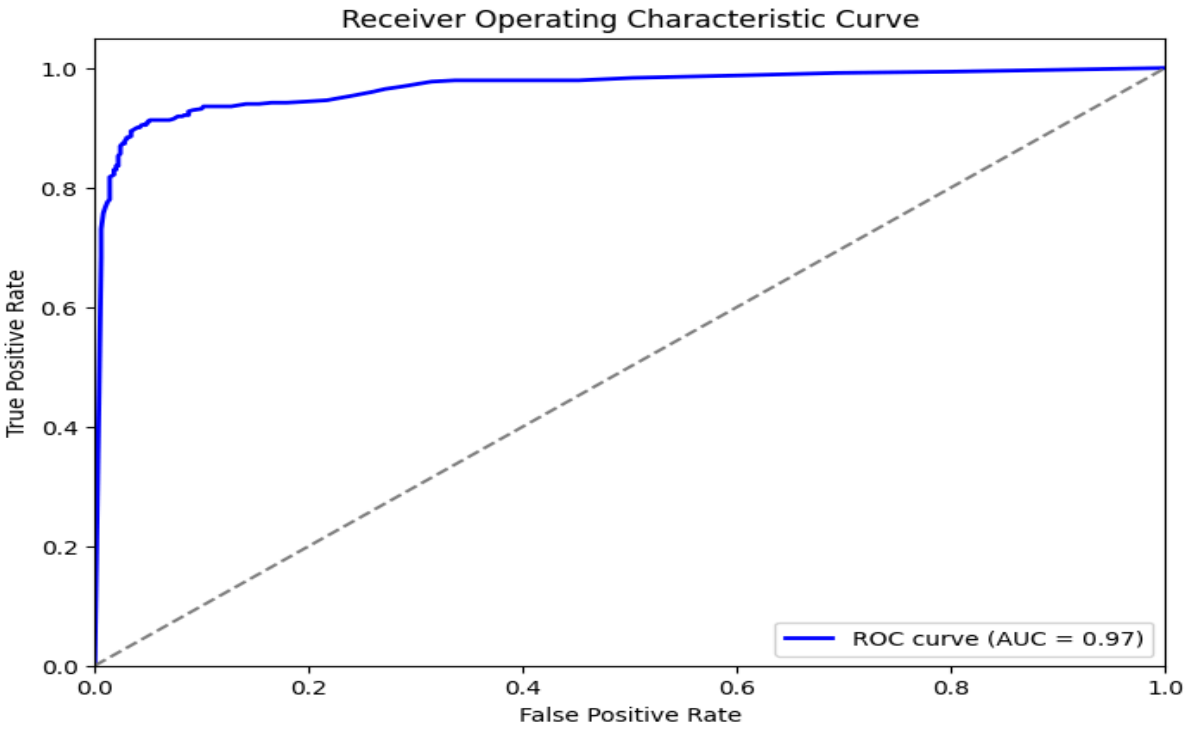
- **Accuracy**: The model achieves 93% accuracy, indicating that the model is effective in distinguishing between pulsars and spurious signals.

- **Precision**:

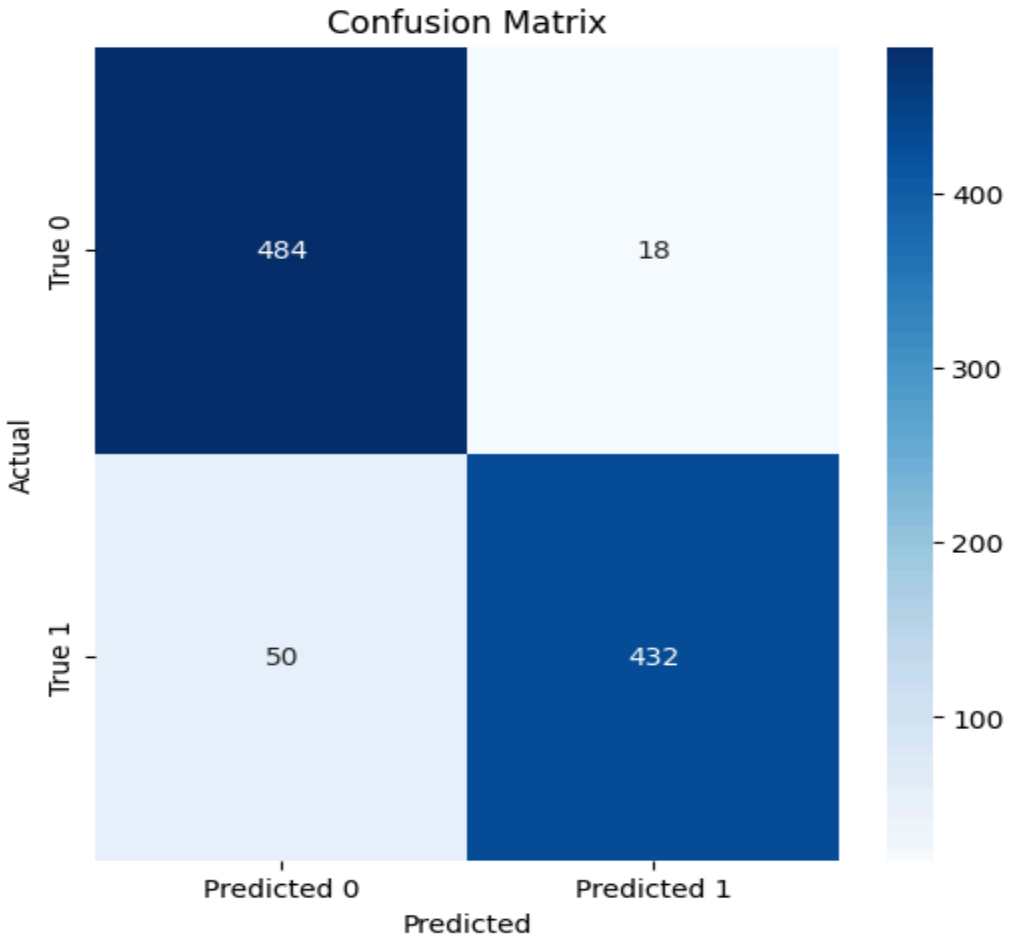
    - **Class 0 (Spurious Signals)**: Precision of **0.91** indicates that when the model predicts a spurious signal, it is correct 91% of the time. This is a strong result as it minimizes the chances of incorrectly classifying non-pulsar signals as pulsars.
    - **Class 1 (Real Pulsars)**: Precision of **0.96** is excellent, showing that the model is highly reliable in its identification of real pulsars, ensuring minimal misclassification of pulsars as spurious signals.
- **Recall**:

    - **Class 0 (Spurious Signals)**: The model achieves **96% recall**, meaning it successfully identifies 96% of all spurious signals. This is crucial for preventing false positives, which are especially detrimental when dealing with large datasets like HTRU2 where noise contamination can be high.
    - **Class 1 (Real Pulsars)**: The **90% recall** for pulsars is strong, but not perfect. There are 18 false negatives (i.e., pulsars misclassified as spurious), which implies that the model is missing some pulsar candidates. Given the dataset's imbalanced nature, this is a reasonable trade-off for ensuring minimal false positives.
- **F1-Score**:

    - **Macro Average F1-Score** of **0.93** confirms balanced performance across both classes, meaning the model is not overly biased toward either class.
    - **Weighted Average F1-Score** of **0.93** aligns with the overall model performance and indicates that both classes are treated equally despite the original class imbalance.


Receiver Operating Characteristic Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (AUC = 0.97)

Confusion Matrix Breakdown



The confusion matrix offers detailed insights into the model's behavior:

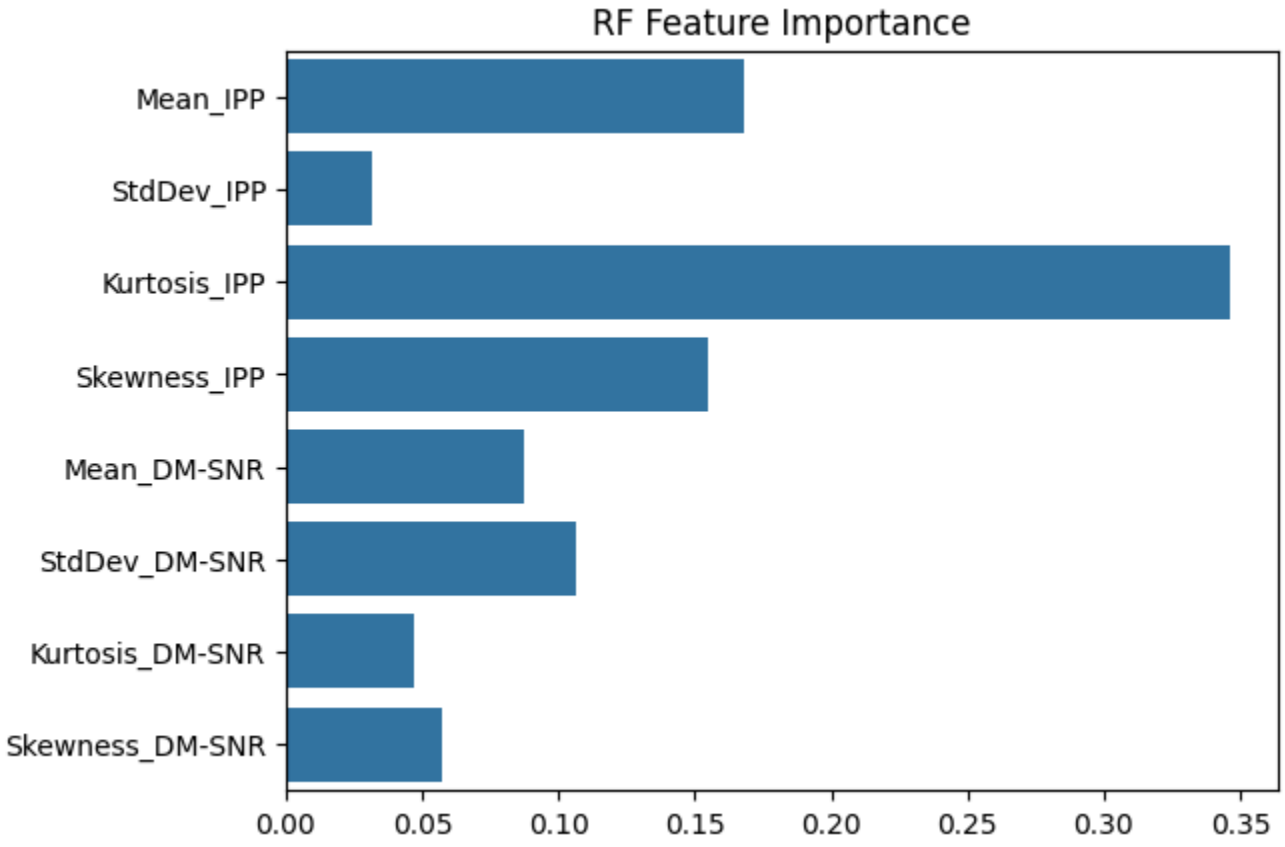
- **True Positives (TP)**: **432 pulsars** correctly identified.
- **True Negatives (TN)**: **484 spurious signals** correctly identified.
- **False Positives (FP)**: **50 spurious signals** incorrectly labeled as pulsars.
- **False Negatives (FN)**: **18 pulsars** incorrectly labeled as spurious signals.

The relatively low number of false positives (50) suggests the model is good at distinguishing pulsar candidates from noise.

- **Class 1 (Real Pulsars)**: The 18 false negatives indicate a minor deficiency in recall, which could be improved with further model tuning (GridSearchCV) or more exploration of complex models like neural networks if the recall for pulsars is crucial.

- **Class 0 (Spurious Signals)**: The 96% recall for spurious signals is excellent, confirming that the model effectively filters out noise and ensures that the majority class is well-recognized.

Feature Importance



The Random Forest model's ability to provide feature importance is beneficial for interpretability. The model ranks the features based on their contribution to the classification task:

- Kurtosis_IPP (0.35) is the most important feature. This suggests that the peakedness of the integrated pulse profile (IPP) is a key distinguishing factor between pulsars and noise. Kurtosis measures the outliers or extreme values in a distribution, which may be more pronounced in pulsar signals.
- Mean_IPP (0.17) and Skewness_IPP (0.16) follow as important features. These characteristics describe the central tendency and asymmetry of the pulse profile, both crucial for detecting pulsar patterns.
- Other DM-SNR features like StdDev_DM-SNR and Skewness_DM-SNR are less influential, highlighting the relatively stronger role of the IPP-derived features in distinguishing the classes.

This feature importance analysis can guide future feature engineering or model refinements, focusing on optimizing the most influential features and understanding how pulsar properties manifest in the dataset. The pulsars clearly have unique distinguishable qualities that indicate it is possible to identify a set of characteristics or phenomena which makes it obvious the signals are different. Research could adjust to these findings to take new approaches to measuring pulsar features.

Conclusion

The Random Forest classifier emerged as the best-performing model for distinguishing real pulsars from spurious signals in the HTRU2 dataset. It achieved a strong AUC-ROC of 0.967 and a balanced F1-Score of 0.93, indicating that it can effectively detect pulsars while minimizing false positives and false negatives. This model is particularly notable for its precision of 0.96, meaning that it is highly reliable in predicting real pulsars, and its recall of 0.90, which ensures that the majority of pulsars are captured by the model.

Through feature importance analysis, I observed that the Kurtosis_IPP and Mean_IPP features played the most significant roles in classification, highlighting that the shape (kurtosis) and average value (mean) of the integrated pulse profiles are crucial for distinguishing between pulsar and spurious signals. This finding not only confirms the effectiveness of these features for identifying pulsars but also provides a deeper understanding of the underlying statistical characteristics of pulsar signals, further informing the modeling process.

The decision to undersample the majority class (spurious signals) helped mitigate the impact of class imbalance. This strategy reduced bias and improved the model's sensitivity to the minority class (real pulsars), though it came at the cost of reducing the total sample size and diversity of the training data. Despite this, the approach improved model fairness and ensured that the model could effectively detect pulsars without favoring the dominant class.

Practical Implications:

The success of the Random Forest classifier in this task demonstrates its practical utility for automating the classification of pulsar candidates in large-scale radio surveys like the HTRU. By automating the detection process, the model significantly reduces the need for manual verification, thereby accelerating the identification of real pulsars and improving the efficiency of surveys. In addition, its strong precision and recall scores mean that the model can confidently identify pulsars while minimizing false identifications which can be critical in astrophysics due to high costs of research and small margin for error in calculations.

The results of the model have immediate applications in the HTRU and similar surveys, where large volumes of candidate pulsar signals need to be sifted through. With automation in place, astronomers can allocate resources more effectively, focusing on analyzing confirmed pulsar candidates for theoretical exploration such as creating space maps and keeping time in space.

The feature importance analysis reveals insightful patterns in the data, emphasizing that Kurtosis_IPP and Mean_IPP are the most influential features for identifying pulsars. This suggests that pulsar signals, which are periodic and exhibit distinct patterns in their pulse profiles, tend to have unique statistical characteristics in terms of their kurtosis (sharpness or flatness of the distribution) and mean values. Understanding this is monumental for both further model refinement and for astrophysical interpretation of pulsar properties. This insight also suggests that models can be further optimized by adding features that capture more detailed characteristics of pulsar signals, such as higher-order statistical moments or frequency-domain features. Moments could inform about the impact of kurtosis

and look into frequency to even further distinguish between the two types of signals due to the overall importance of mean in distinguishing the classes.

The model's reliance on specific features like Kurtosis_IPP and Mean_IPP indicates that a better understanding of these statistical metrics in pulsar signals could enhance data collection methods, possibly guiding the design of research equipment itself to focus on these features associated with these attributes over others for more efficient detection, such as increasing the interval rate of measurement due to potential spiking pulse profiles.

The model also has its limitations:

1. Generalizability to Other Datasets: The model was trained and validated on the HTRU2 dataset, which represents a specific understanding of pulsar signals. While the model performs well on this dataset, its generalizability to other surveys or datasets with different types of noise or signal characteristics is uncertain. The model may require retraining or fine-tuning if applied to other datasets, especially if the nature of the pulsar signals differs (e.g., in different frequency bands or with different observational conditions).

2. Impact of Data Preprocessing: The use of undersampling was effective in addressing the class imbalance, but it also limited the training data. While this led to a more balanced model, it reduced the overall data volume and might have resulted in some loss of valuable signal diversity. More sophisticated techniques, such as SMOTE (Synthetic Minority Over-sampling Technique) or weighted loss functions, could be explored to balance the dataset without reducing sample size.

In conclusion, the Random Forest model has shown great promise in automating pulsar classification and can be applied effectively to large-scale surveys. Its balance between precision and recall, combined with feature importance insights, makes it an excellent tool for pulsar research. However, as with any model, it is important to recognize its limitations and continue exploring ways to refine and improve it through better data, advanced techniques, and deeper scientific understanding.