Neural Networks for Classifying Pulsar Candidates

Nadison Kannan

Department of Geology, Portland State University

G410: Machine Learning for the Natural Sciences

Professor David Percy

13 June 2025

Abstract

Pulsars are rotating neutron stars that emit beams of radio frequency light that can be observed on Earth. The analysis of these signals offers scientists ways of exploring states of matter in hyper dense environments and a way to analyze gravitational waves. Classifying pulsars is done manually because sometimes signals can be noise and not actual pulsars. Machine learning is being explored to efficiently classify pulsars based on their pulsar profile.

The goal of this project is to explore the use of a neural network machine learning model for classifying candidate pulsar signals using supervised learning techniques. This paper aims to answer the following question: can neural networks accurately classify a pulsar based on its candidate profile? The experiment is carried out using Orange which is an open-source machine learning and data visualization software.

The experiment outlined in this paper trains three models on the HTRU2 data set from the UC Irvine Machine Learning Repository. The models trained are a neural network, a random forest, and naive bayes. Multiple models are trained to compare the accuracy and computational efficiency of the neural network against simpler and faster models.

Models are trained using cross-validation and evaluated with a series of metrics including precision tests and classification accuracy. The neural network and the random forest models performed equally well and the naive bayes performing slightly worse but still with high accuracy. Because the neural network's accuracy was comparable to the random forests and because the neural network is less computationally efficient than the random forest, it is concluded that there is no real advantage to using a neural network for classifying pulsars.

Introduction

Pulsars

Pulsars are super dense and highly magnetized rotating neutron stars, created by massive stars that stop nuclear fusion and collapse on themselves, resulting in a supernova (Socorro). They channel particles along their magnetic poles, accelerating them to relativistic speeds and thus emitting beams of electromagnetic radiation out from these poles. These beams are not aligned with the axis that the pulsar rotates about and are swept around as the pulsar rotates (Lea). These can be detected as they pass through Earth, appearing as a pulse of radio waves to detection instruments.

The understanding and discovery of pulsars are significant for the purposes of physics and astronomy. The light emitted from pulsars carry important information about their nature. Because of the immense pressure that the pulsar is subjected to, it can give scientists insights into otherwise unobservable states of matter (Lea). Additionally, pulsars are used to directly test the general theory of relativity which theorizes that there should exist gravitational waves cause by the motion of massive celestial bodies in the universe (NRAO). These gravitational waves cause ripples in the fabric of space-time which will alter the steady frequency of the observed rotation.

Motivation

Because of their important application in physics and astronomy, it is important to be able to correctly identify them so they can be used in research. Each pulsar has a unique emission pattern with slight variations in each rotation (Lyon). As the light signal from the pulsar travels through the interstellar medium, they experience a unique dispersion that is characteristic of a pulsar (R.J. Lyon). These signals are recorded in a collection of plots and summary statistics which comprises what is called a pulsar ‘candidate,’ which characterizes the potential pulsar (R.J. Lyon). However, most candidates turn out to be radio frequency interference (RFI) and noise, so each candidate must be inspected manually or by an automated method (Lyon). Due to advances in telescope technology and detection capabilities, the number of candidates detected has increased significantly to the point where manual inspection has become impractical, referred to as the ‘candidate selection problem’ (R.J. Lyon).

Automated methods address the candidate selection problem by allowing large numbers of candidates to be quickly evaluated (R.J. Lyon). Machine learning one of the automated methods that is being used for evaluating pulsar candidates efficiently classify candidates as a pulsar or noise. However, machine learning models built for pulsar candidate evaluation are often specific to a particular pulsar survey search pipeline, which makes these models unsuitable as a general solution to the candidate selection problem (R.J. Lyon).

Objective

The current research uses a specific machine learning model called the “Gaussian Hellinger Very Fast Decision Tree” (GH-VFDT) (R.J. Lyon). This model is a decision tree algorithm and was designed specifically for the data set obtained with the researcher’s particular pulsar survey search pipeline. The goal of this project is to evaluate the efficacy of a neural network machine learning model for accurately classifying candidate pulsar signals from non-pulsar signals using supervised learning techniques. Thus, the research question driving this project is as follows: can neural networks accurately classify a pulsar based on its candidate profile? Neural networks are powerful models that might serve as the basis for custom models used on other pulsar survey search pipelines.

Methods

Neural Networks

Neural networks are machine learning models that consist of interconnected nodes, called neurons, that process data non-linearly (GeeksforGeeks). Neural networks are modeled to mimic the functions of the human brain; thus, these models are effective at learning patterns without pre-defined rules. Learning in a neural network works in a three-stage process; 1) input data is given to the neural network, 2) the network generates an output based on the current parameters, 3) the network adjusts weights and biases iteratively to improve the performance of the model.

Neural networks can be used for supervised learning techniques for the prediction of classes based on labeled data. The neural network is trained on the labeled input-output pairs and will generate outputs (GeeksforGeeks). These outputs are compared with the desired outputs, and a prediction error is created. This prediction error is then used as the model enters subsequent iterations as it adjusts its weights and biases to minimize the error until it converges on the lowest error it can achieve.

Neural networks have several advantages that make them appealing for supervised learning applications. They can be effective in identifying relationships between the input-output data due to the non-linear activation functions in the neurons as well as their adaptability with pattern recognition through iterative learning. They can also be computationally efficient by parallelizing tasks due to the nature of their design (GeeksforGeeks). Thus, neural networks can be both computationally efficient and powerful learning algorithms, which are advantageous considerations when choosing machine learning algorithms.

Data Sources

The HTRU2 data set was downloaded from the UC Irvine Machine Learning Repository which is a "collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms" (UC Irvine MLR). The following citation is copied from the data set readme file:

> *"This data was obtained with the support of grant EP/I028099/1 for the University of Manchester Centre for Doctoral Training in Computer Science, from the UK Engineering and Physical Sciences Research Council (EPSRC). The raw observational data was collected by the High Time Resolution Universe Collaboration using the Parkes Observatory, funded by the Commonwealth of Australia and managed by the CSIRO."*

The data set has eight features that describe each pulsar candidate collected during the HTRU survey. Each of the eight features are continuous variables and the candidate is represented by a single class variable that denotes whether the data point is a pulsar or not a pulsar. The list of features is as follows:

1. Mean of integrated profile.
2. Standard deviation of the integrated profile.
3. Excess kurtosis of the integrated profile.
4. Skewness of the integrated profile.
5. Mean of the DM-SNR curve.
6. Standard deviation of the DM-SNR curve.
7. Excess kurtosis of the DM-SNR curve.
8. Skewness of the DM-SNR curve.
9. Class

These candidate features were chosen because they are hypothesized to maximize the separation between noise and non-noise candidates (R.J. Lyon). The first four features all come from the integrated pulsar profile, which is a collection of observations from the pulsar that characterize it, like a fingerprint. The next four variables are from the DM-SNR curve, which is a curve that represents the relationship between the dispersion measure (DM) and the signal to noise ratio (SNR) (Holewik J). In total, the data set contains 16,259 samples caused by RFI/noise, and 1,639 real pulsars that were manually confirmed (Lyon, R).

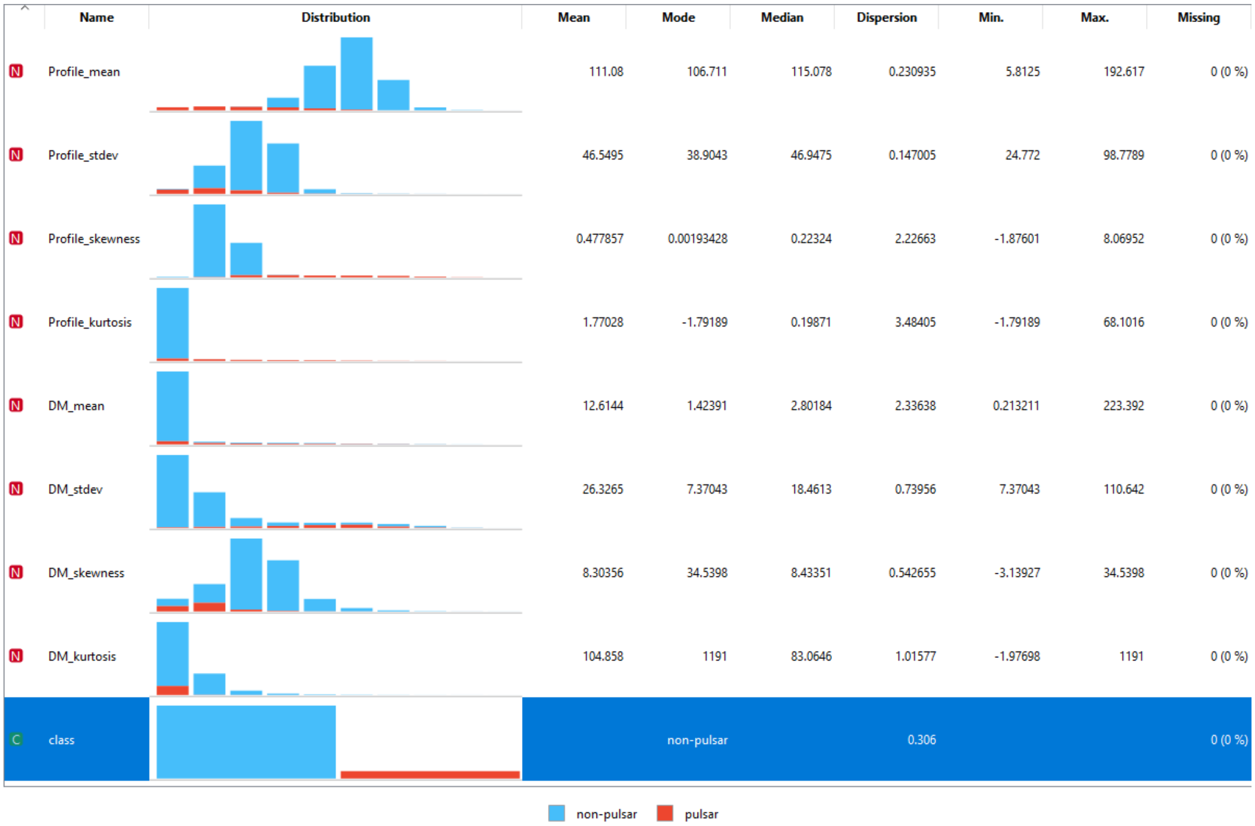
Exploratory Data Analysis

Before beginning the train the model, I looked at the data to get a general idea of distribution and other important summary statistics. First, however, after uploading the data set to the Orange workspace, I noticed that the feature variable names were numbers and not the names of what the represented. I used the 'Edit Domain' widget to rename the variables accordingly so that working with the data for the remainder of the project would be easier.

With the data columns renamed, I started the exploratory data analysis by looking at the features statistics. Looking at the distributions and summary statistics in Figure 1, there is a significant difference between the number of pulsars (red) and non-pulsars (blue) in the data. This is important to note because training on model with such a large difference might deprive the model of important instances of learning from pulsar data since it is so sparse. This might lead to an imbalanced model, but I will proceed with training to see how the model performs despite this.

Figure 1

Feature statistics of the HTRU2 pulsar candidates data set.



	Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
N	Profile_mean		111.08	106.711	115.078	0.230935	5.8125	192.617	0 (0 %)
N	Profile_stdev		46.5495	38.9043	46.9475	0.147005	24.772	98.7789	0 (0 %)
N	Profile_skewness		0.477857	0.00193428	0.22324	2.22663	-1.87601	8.06952	0 (0 %)
N	Profile_kurtosis		1.77028	-1.79189	0.19871	3.48405	-1.79189	68.1016	0 (0 %)
N	DM_mean		12.6144	1.42391	2.80184	2.33638	0.213211	223.392	0 (0 %)
N	DM_stdev		26.3265	7.37043	18.4613	0.73956	7.37043	110.642	0 (0 %)
N	DM_skewness		8.30356	34.5398	8.43351	0.542655	-3.13927	34.5398	0 (0 %)
N	DM_kurtosis		104.858	1191	83.0646	1.01577	-1.97698	1191	0 (0 %)
C	class			non-pulsar		0.306			0 (0 %)

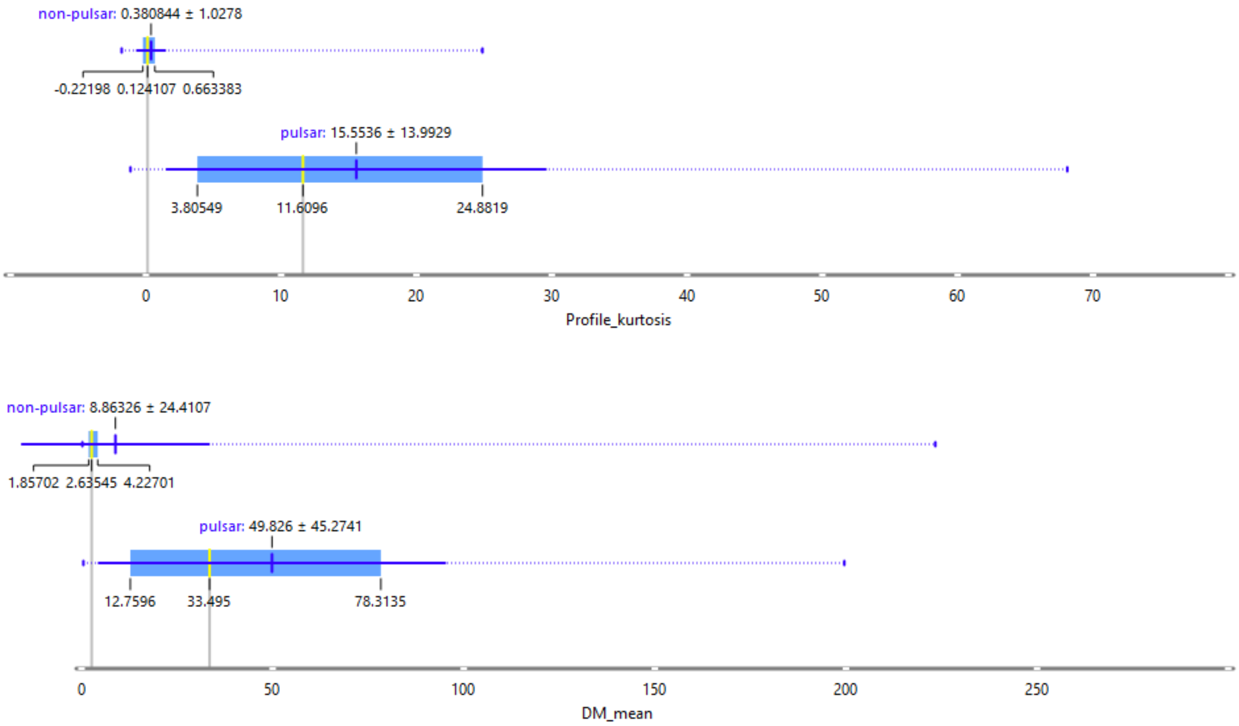
🟦 non-pulsar 🟥 pulsar

Additionally, most of the variables have a low dispersion value, which suggests that most of the data is close to the median and consistent. However, the profile kurtosis and the DM mean variables appear to have a much larger spread from the median which causes a significant positive skew of the data. Since neural network models are generally more robust, leaving these positively skewed distributions as they are may still yield an accurate model. However, it is unclear whether the pulsar candidates or the non-pulsar candidates are skewed, for this it may be useful to look at another visualization.

To look at the profile kurtosis and the DM mean distributions more closely, I will visualize them on box plots to see the spreads of the pulsars vs the non-pulsars. Looking at Figure 2, there is a clear distinction between the spreads of the pulsars and the non-pulsars. The non-pulsars for both profile kurtosis and DM mean have a small spread with a median close to zero. The non-pulsars for both features still exhibit a positive skew due to some large outliers. Looking at the pulsar data points for both features, these data points have a larger spread which may have contributed to the shape of the distributions for these features that was seen in Figure 1. The pulsars also exhibit a slight positive skew which indicates that the data is being pulled right by some large values.

Figure 2

Box plots of the pulsar candidate features profile kurtosis (top) and DM mean (bottom). The median is shown as the solid grey line connecting the box to the axis.

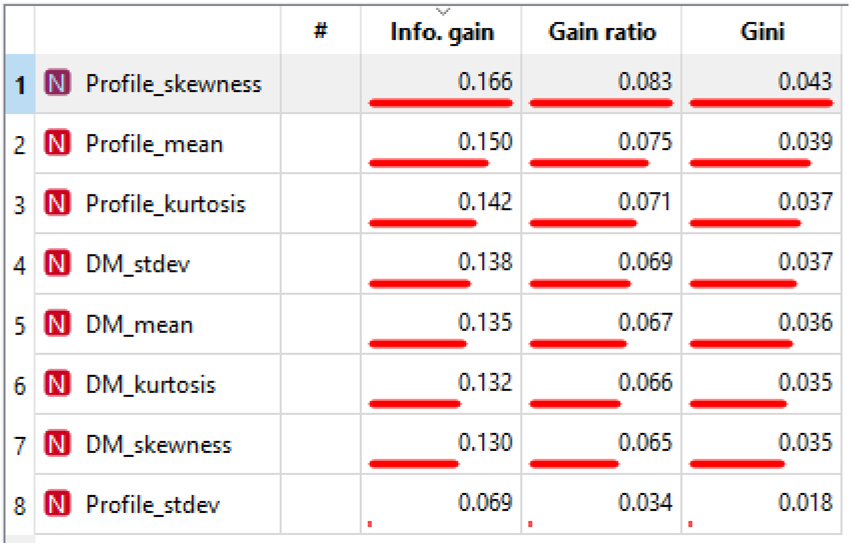


Interestingly, the pulsars and non-pulsars for profile kurtosis and DM mean have largely different distributions. This distinction in distribution might turn out to be beneficial for training a machine learning model since these data points within these features are so distinct in their individual characteristics.

To conclude the exploratory data analysis, I will use the ‘Rank’ widget in Orange to see the correlations with the discrete target variable. The correlations will be evaluated with three different scoring methods; information gain, the expected amount of information; information gain ratio, a ratio of the information gain and the attribute's intrinsic information; and Gini decrease, the inequality among values of a frequency distribution (Demsar J). From Figure 3, the highest correlated feature by all scoring methods is the profile skewness, however, all the features except for profile standard deviation have a high correlation with the target variable. This suggests that all the variables in the features matrix will be good choices for training a model.

Figure 3

Features ranked by their correlation with the discrete target variable and sorted in descending order for various scoring methods.



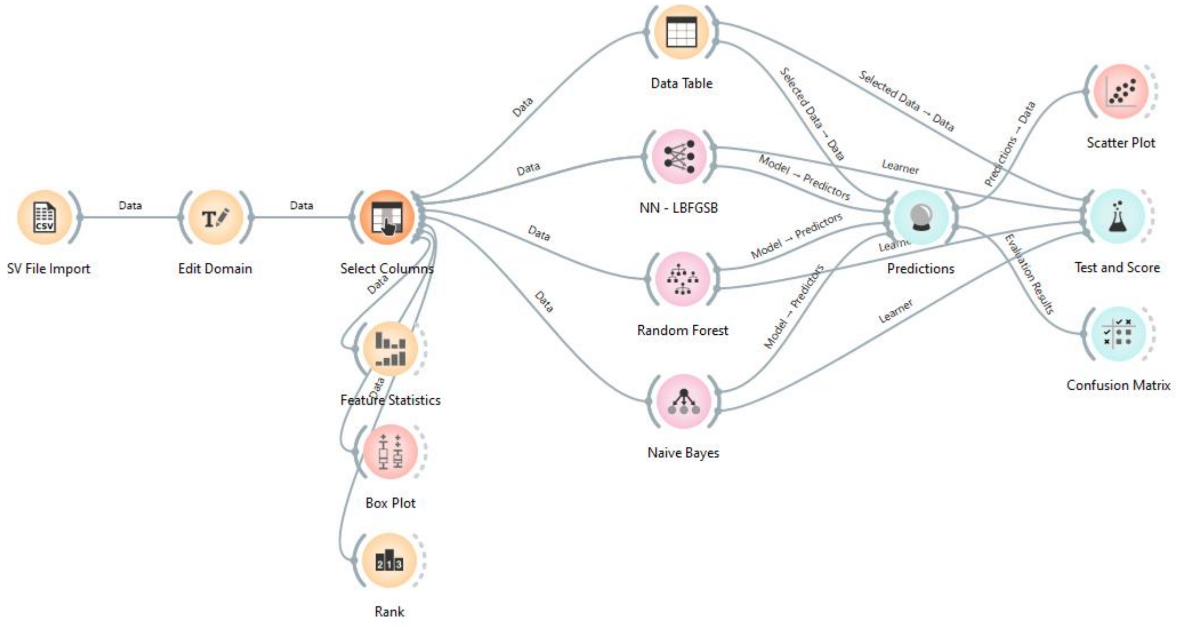
		#	Info. gain	Gain ratio	Gini
1	N Profile_skewness		0.166	0.083	0.043
2	N Profile_mean		0.150	0.075	0.039
3	N Profile_kurtosis		0.142	0.071	0.037
4	N DM_stdev		0.138	0.069	0.037
5	N DM_mean		0.135	0.067	0.036
6	N DM_kurtosis		0.132	0.066	0.035
7	N DM_skewness		0.130	0.065	0.035
8	N Profile_stdev		0.069	0.034	0.018

Training the Models

With a better understanding of how the features relate to the target variable, I begin training a neural network, random forest and naive bayes model to gauge the neural network’s performance for classifying pulsars based on pulsar profiles. The three models will be trained on all eight of the features with the neural network using 3 hidden layers, a logistic activation function, a regularization of 0.0001 and a maximum of 200 iterations. The random forest will use 5 trees with 2 attributes considered at each split. All the models will be trained using cross-validation with five folds and evaluated with accuracy metrics and a confusion matrix. The full workflow in orange is shown below in Figure 4. The comparison models were chosen based on their difference in learning. The random forest is an ensemble method that uses decision trees which is the same type of model as the GH-VFDT model that was used in the research paper that the data is from. The Naive Bayes model is probabilistic and might address areas of overlap better.

Figure 4

Workflow of the entire training process designed and carried out in Orange.



Results

With the models trained, I can now evaluate their performance in classifying pulsars and non-pulsars and compare them to one another. The tabulated results of the cross-validation testing are included below in Figure 5. From the table, all three models performed equally well with scores for most tests above 0.9. The neural network and the random forest performed equally well and the naive bayes model only slightly underperformed the other two.

Figure 5

Output of the ‘Test and Score’ widget in Orange that shows the performance of the neural network, random forest, and naive bayes models for various evaluation metrics.

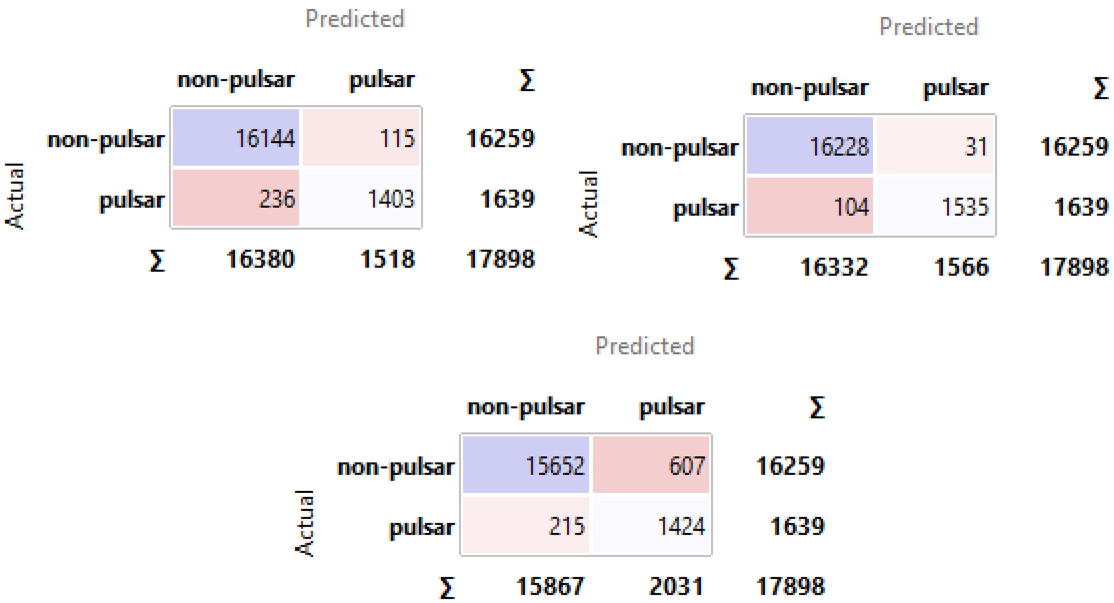
Test and Score						
Model	AUC	CA	F1	Prec	Recall	MCC
Neural Network	0.978	0.979	0.979	0.979	0.979	0.871
Random Forest	0.953	0.979	0.979	0.979	0.979	0.870
Naive Bayes	0.961	0.954	0.956	0.960	0.954	0.757

Most notably, all the models performed well for important classification tests like the precision (Prec) test, which measures the proportion of true positives among instances classified as positive (Demsar J). This means that all three models are predicting pulsars with high accuracy. Another metric is the classification accuracy (CA) test which measures the proportion of correctly classified instances (Demsar J). The neural network and the random forest both scored 0.979 for this test, which suggest that both models have a high proportion of correctly classified instances. The naive bayes model also performed similar well for this metric with a score of 0.954.

Another way to evaluate classification accuracy is to plot the outputs of these models in a confusion matrix. Figure 6 below shows the confusion matrices for each of the three models. The confusion matrices reveal an interesting insight between the neural network and random forest matrix. Despite the neural network generally performing better than the random forest, the confusion matrix for the random forest appears to have classified more instances correctly. This is unexpected based on the outputs in Figure 5 since both models have similar scores for most metrics. This might be explained as the types of points that were classified correctly and incorrectly. It is possible that the neural network classifies points that are in areas of overlap better than the decision tree does and thus scored better on some metrics because of this.

Figure 6

Confusion matrices for the neural network (top left), random forest (top right), and the naive bayes (bottom) models.



		Predicted		
		non-pulsar	pulsar	Σ
Actual	non-pulsar	16144	115	16259
	pulsar	236	1403	1639
	Σ	16380	1518	17898

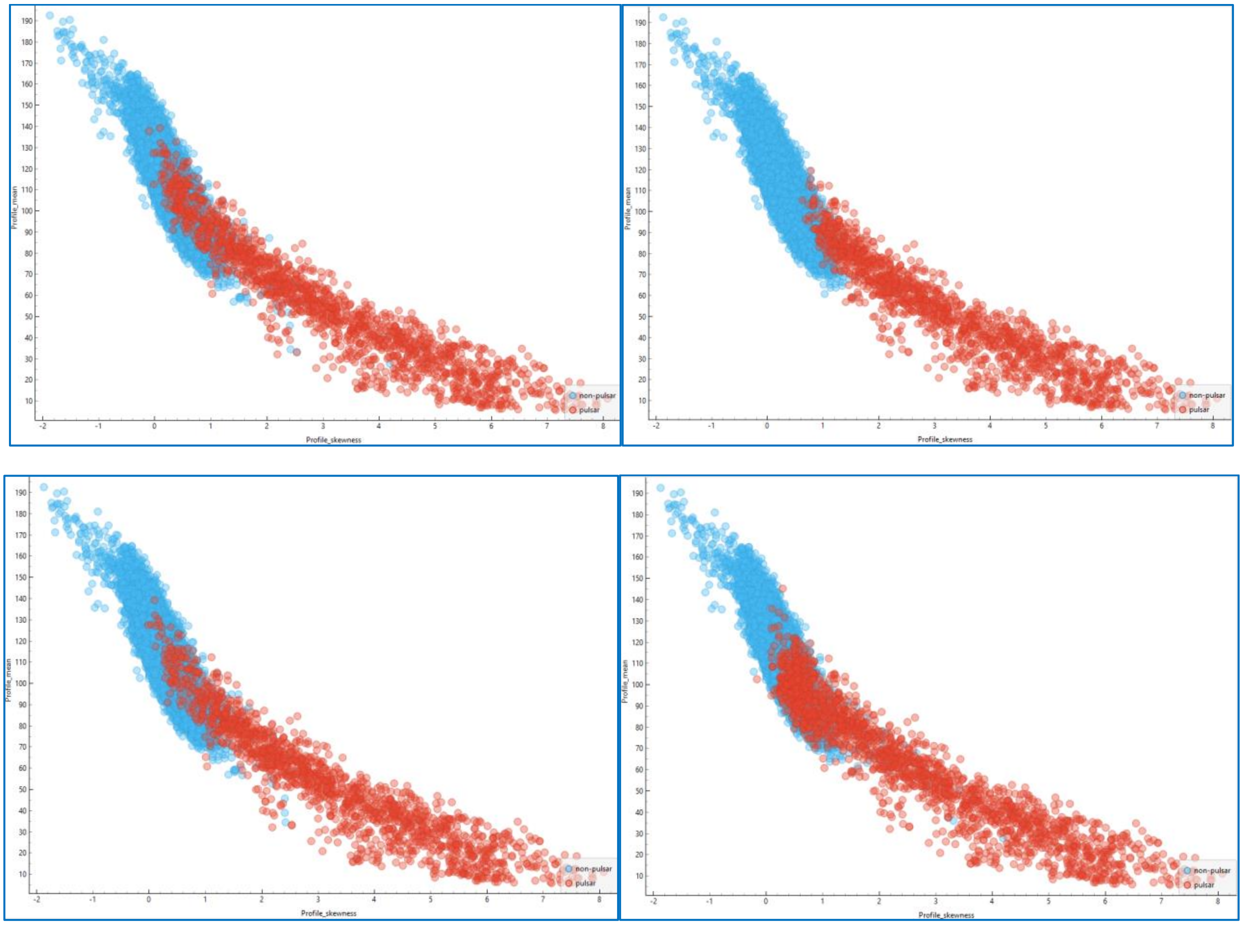
		Predicted		
		non-pulsar	pulsar	Σ
Actual	non-pulsar	16228	31	16259
	pulsar	104	1535	1639
	Σ	16332	1566	17898

		Predicted		
		non-pulsar	pulsar	Σ
Actual	non-pulsar	15652	607	16259
	pulsar	215	1424	1639
	Σ	15867	2031	17898

To visualize the performance of the models in their ability to classify points in areas of significant overlap, I will plot scatter plots that show the predictions of the three models compared to a scatter plot of the actual labeled data. Figure 7 shows the scatter plots for the actual labeled data and the predictions of each model. The axes were chosen due to their high correlation with the target variable and the visualization of areas of overlap.

Figure 7

Scatter plots of the actual labeled data compared to scatter plots of the model predictions. (Top left) Actual data, (top right) neural network predictions, (bottom left) random forest predictions, (bottom right) naive bayes predictions.



Interestingly, these scatter plots give another interesting insight into the models that the test and score table in Figure 5 does not show. Originally, I thought the discrepancy in the confusion matrix was due to the neural network's ability to classify points that were in areas of significant overlap. However, comparing the scatter plots of the model predictions with the scatter plot of the actual labeled data, it appears as though the neural network classifies the least number of points in the overlapping area. This could be partially obscured since there are so many points, but it appears that both the naive bayes and the random forest can classify pulsars that are far into the non-pulsar group, and vice versa. It is important to keep in mind that this is one way to visualize the data, and the neural network might show better results in a scatter plot using different variables.

Conclusion

Based on the results of the model scores, confusion matrices, and scatter plots, I can conclude that neural networks, offer no real advantage over decision trees or naive bayes models. Even though the accuracy of the neural network is notable, it is more computationally demanding than the random forest and naive bayes models. As parameters are scaled up in the neural network, the training becomes even more demanding, whereas the random forest model has much more headroom for scaling up model complexity making it ideal for a larger data set of pulsar candidates.

Overall, neural networks might offer an alternative to random forest or ensemble models because of their ability to learn complex patterns in non-linear and multidimensional data. The neural network performs well with pulsar candidate data, and if computation time is not a concern, it might be a viable alternative. Additionally, if there are more features that must be analyzed, the neural network may be better suited to capture complex interactions between seemingly disparate features.

Future experiments may need address the significant overlap in the pulsars and non-pulsars. Finding models that can address this area of overlap or project the features into a space where they are linearly separable may provide more accurate classification in otherwise uncertain areas. Exploring and experimenting with various models is an important part of not only finding efficient ways of classifying data but is also an important step in learning about the data and finding relationships between features that may not have been immediately obvious.

References

Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python, Journal of Machine Learning Research 14(Aug): 2349−2353.

GeeksforGeeks. What is a neural network?. https://www.geeksforgeeks.org/neural-networks-a-beginners-guide/

Holewik J, Schaefer G, Korovin I. Imbalanced Ensemble Learning for Enhanced Pulsar Identification. Advances in Swarm Intelligence. 2020 Jun 22;12145:515–24. doi: 10.1007/978-3-030-53956-6_47. PMCID: PMC7354782.

Lea, R. (2016, April 22). What are pulsars?. Space. https://www.space.com/32661-pulsars.html

Lyon, R. (2015). HTRU2 [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5DK6R.

R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, J. D. Knowles, Fifty Years of Pulsar Candidate Selection: From simple filters to a new principled real-time classification approach MNRAS, 2016.

R. J. Lyon, HTRU2, DOI: 10.6084/m9.figshare.3080389.v1.

Socorro, NRAO. (2012, February 20). Pulsars: The universe’s gift to physics. Astronomy.com. https://www.astronomy.com/science/pulsars-the-universes-gift-to-physics/