Pulsar Star Classificiation

Hanan Salim

Contents

1 Introducation **1**

2 EDA and Feature Engineering **3**

3 Logistic Regression **8**

4 Perceptron **8**

5 Decision Tree **8**

6 Bootstrap **8**

7 Conclusion **8**

8 References **12**

1 Introducation

On a dark, clear night, the sky is a canvas dotted with countless stars, shining brightly due to their intense heat. In the 1960s, astronomers began studying these stars through different wavelengths of light and made a surprising discovery: some stars were blinking. While initially, this phenomenon sparked speculation about extraterrestrial life, scientists soon realized that these were unique stars known as pulsars.

Pulsars are formed from the remnants of a massive star that has undergone a supernova explosion. During this cataclysmic event, the star's core collapses under its immense gravity, forming an incredibly dense neutron star. This newly formed neutron star spins rapidly, retaining much of the original star's angular momentum. A pulsar is a special type of neutron star that emits radiation.

Pulsars possess incredibly strong magnetic fields, which propel particles outward along their magnetic poles. These accelerated particles generate powerful beams of light, causing the pulsating effect observed by astronomers. As the pulsar rotates, these beams sweep across the sky, making them appear to blink.

Pulsars are the lighthouses of the universe and have become important celestial objects to study. Unfortunately, they emit very weak signals which can be lost in background noise when making observations. Pulse emissions from each individual rotation (single pulses), are highly variable from pulse to pulse. To overcome this, we can average multiple single pulses into an integrated pulse profile which is more consistent across different observations and stable in time.

But, an integrated pulse profile is not enough to know if the signal is actually from a pulsar. It could be mixed in with other information or be from a completely different source. This is where the DM-SNR curve

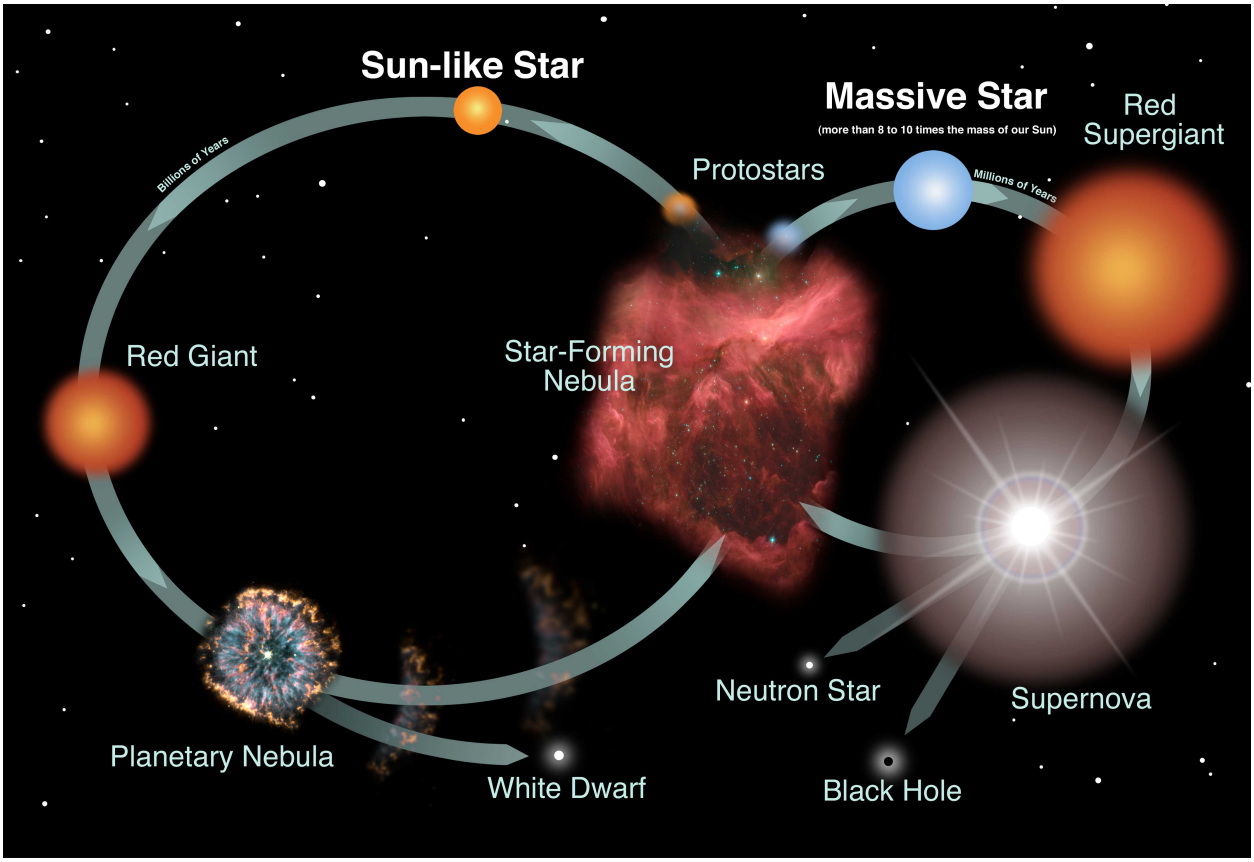


Figure 1: Life Cycle of a Star

comes into play. The DM-SNR curve has two components, the dispersion measure (DM) and signal-to-noise ratio (SNR). SNR measures the strength of a pulsar's signal relative to background noise.The dispersion measure, on the other hand, measures the spread of the pulse. When a pulsar emits a pulse of radiation, it contains a range of frequencies. As this pulse travels through space, the different frequencies travel at slightly different speeds. This causes the pulse to spread out, or disperse, over time. We capture this through the dispersion measure.

The DM is plotted on the x-axis and the SNR on the y-axis which give us the the DM-SNR curve. Notice that since both of these are curves, we can view them as distributions and thus describe them completely by calculating there mean, standard deviation, skewness, and kurtosis. This exactly what our data does. It describes each obsersation based on 8 features which are summarized below.

Our feature are:

```
1. ‘mean_ip‘ : Mean of the integrated profile \
2. ‘sd_ip‘ : Standard deviation of the integrated profile \
3. ‘kurtosis_ip‘ : Excess kurtosis of the integrated profile \
4. ‘skewness_ip‘ : Skewness of the integrated profile \
5. ‘mean_ds‘ : Mean of the DM-SNR curve \
6. ‘sd_ds‘ : Standard deviation of the DM-SNR curve \
7. ‘kurtosis_ds‘ : Excess kurtosis of the DM-SNR curve \
8. ‘skewness_ds‘ : Skewness of the DM-SNR curve
```

Along with our features we have our response variable under the `results` column. The data set contains 16,259 observations caused by noise, and 1,639 real pulsar observations, for a total of 17898 observations.

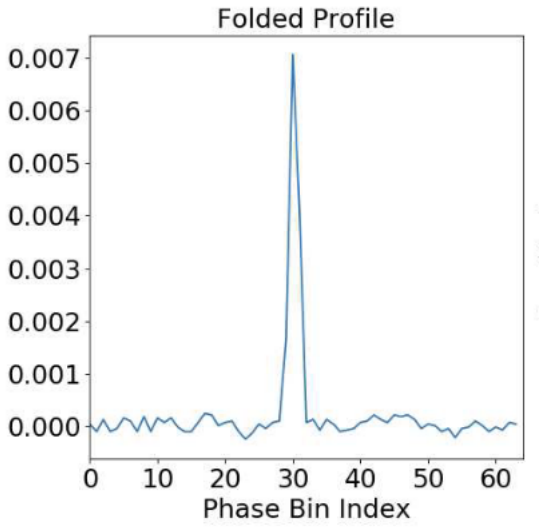


Figure 2: Integrated Pulsar Profile

2 EDA and Feature Engineering

```
   mean_ip            sd_ip         kurtosis_ip       skewness_ip
Min.   :  5.812   Min.   :24.77   Min.   :-1.8760   Min.   :-1.7919
1st Qu.:100.930   1st Qu.:42.38   1st Qu.: 0.0271   1st Qu.:-0.1886
Median :115.078   Median :46.95   Median : 0.2232   Median : 0.1987
Mean   :111.080   Mean   :46.55   Mean   : 0.4779   Mean   : 1.7703
```

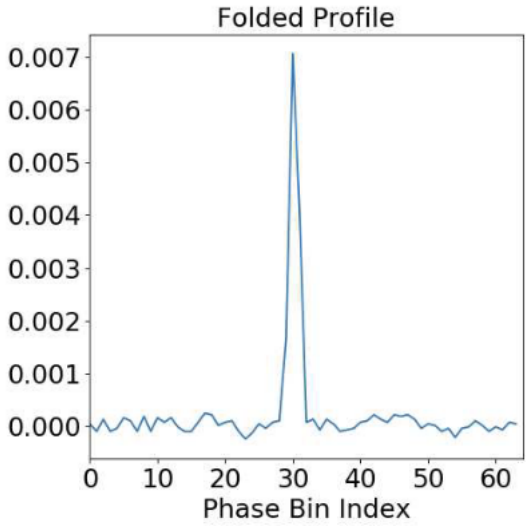


Figure 3: DM-SNR Curve

```
3rd Qu.:127.086   3rd Qu.:51.02   3rd Qu.: 0.4733   3rd Qu.: 0.9278
Max.   :192.617   Max.   :98.78   Max.   : 8.0695   Max.   :68.1016
   mean_ds            sd_ds         kurtosis_ds       skewness_ds
Min.   :  0.2132   Min.   :  7.37   Min.   :-3.139   Min.   :  -1.977
1st Qu.:  1.9231   1st Qu.: 14.44   1st Qu.: 5.782   1st Qu.:  34.961
Median :  2.8018   Median : 18.46   Median : 8.434   Median :  83.065
Mean   : 12.6144   Mean   : 26.33   Mean   : 8.304   Mean   : 104.858
3rd Qu.:  5.4643   3rd Qu.: 28.43   3rd Qu.:10.703   3rd Qu.: 139.309
Max.   :223.3921   Max.   :110.64   Max.   :34.540   Max.   :1191.001
    result
Min.   :0.00000
1st Qu.:0.00000
Median :0.00000
Mean   :0.09157
3rd Qu.:0.00000
Max.   :1.00000
```

We see from the summary statement, that our data does not have any missing data. Furthermore, since our features are statistics describing two distributions, we do not intend to combine or categorize them. But, we would like to assess their distributions. In figure 4, we see that the mean and standard deviation for the integrated pulse profile are slightly skewed, whereas the kurtosis and skewness are strongly skewed. Similarly, in figure 5 we see that all four features associated with the DM-SNR curve are skewed. We may want to log transform our data before downstream analysis.

Figure 6, shows the distribution for each statistic comparing pulsar candidates vs non-pulsar candidates. All features show a strong separation between our response variable categories. From figures 4 and 5, we can see that our features have different scales and are heavily skewed. To fix the scales, we can apply min-max normalization. To fix the skew, we would typically apply a log transformation but since our data has negative values and zeros, we instead apply a negative log transformation shown below.

$$T(x) = \text{sign}(x) \cdot \log(|x| + 1)$$

In figure 7, we can see that applying $T(x)$ followed by normalization helps reduce the skew while still retaining separation.

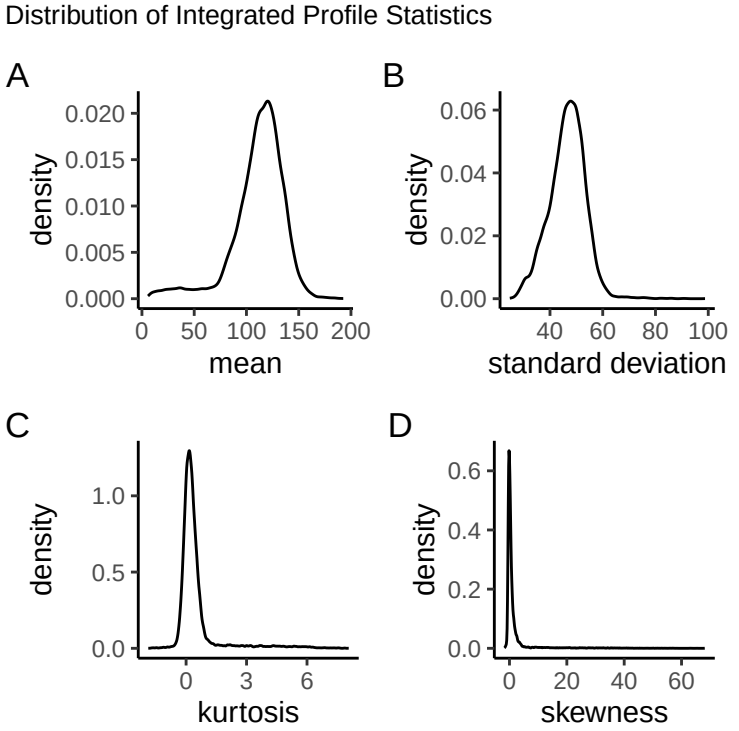


Figure 4: Distribution plots of integrated pulse profile statistics

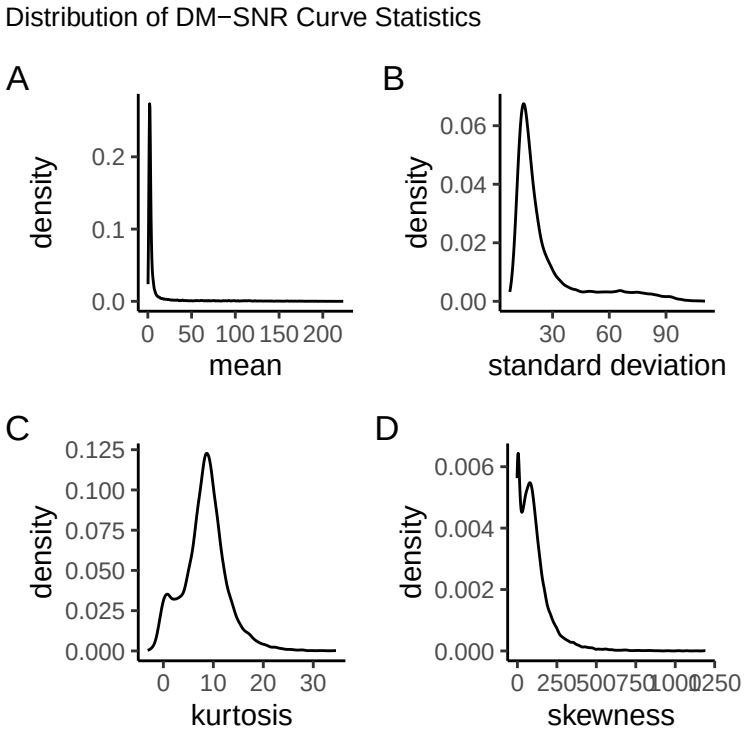


Figure 5: Distribution plots of DM-SNR curve statistics

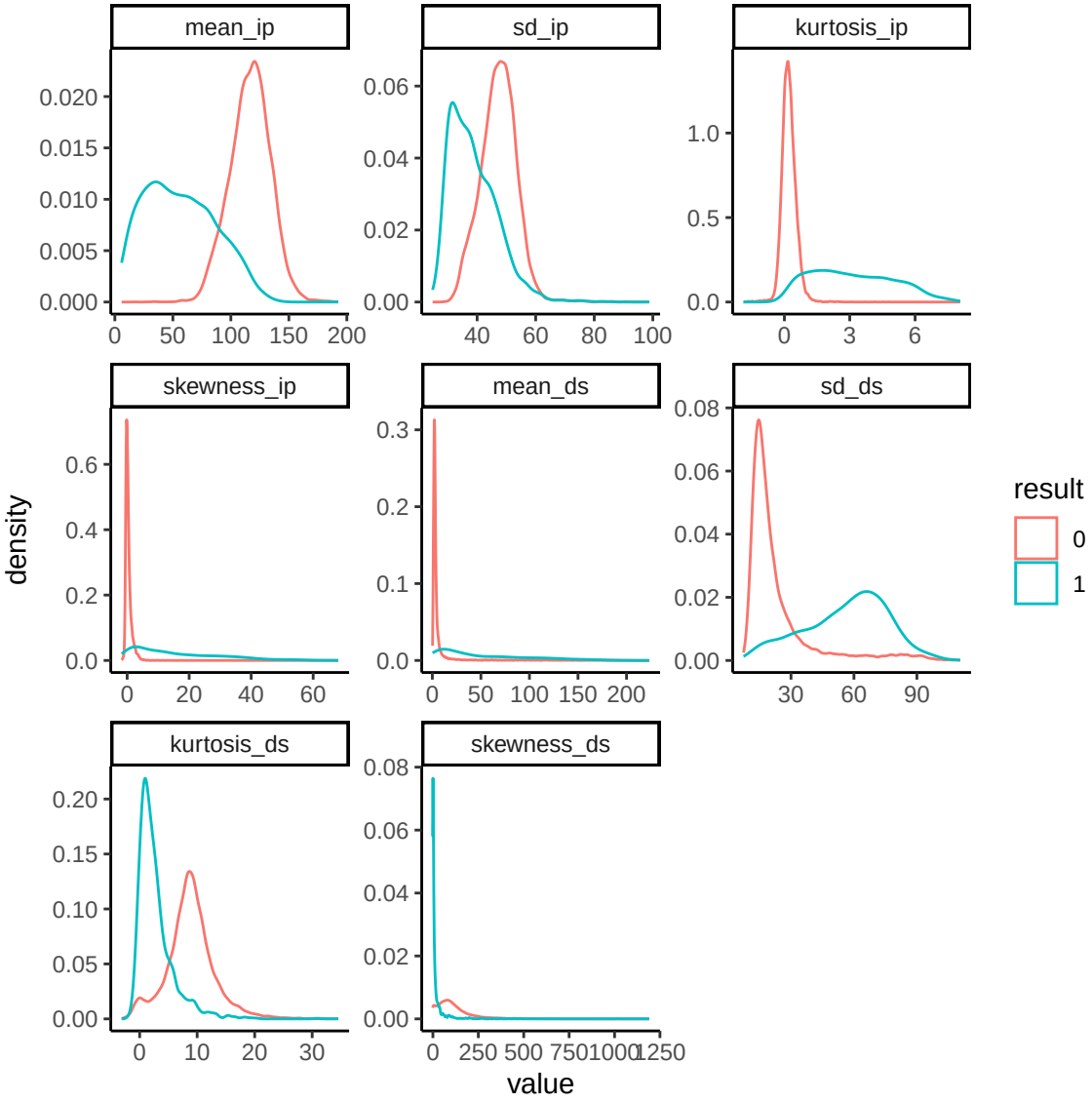


Figure 6: Density plots comparing pulsar candidates vs non-pulsars candidates

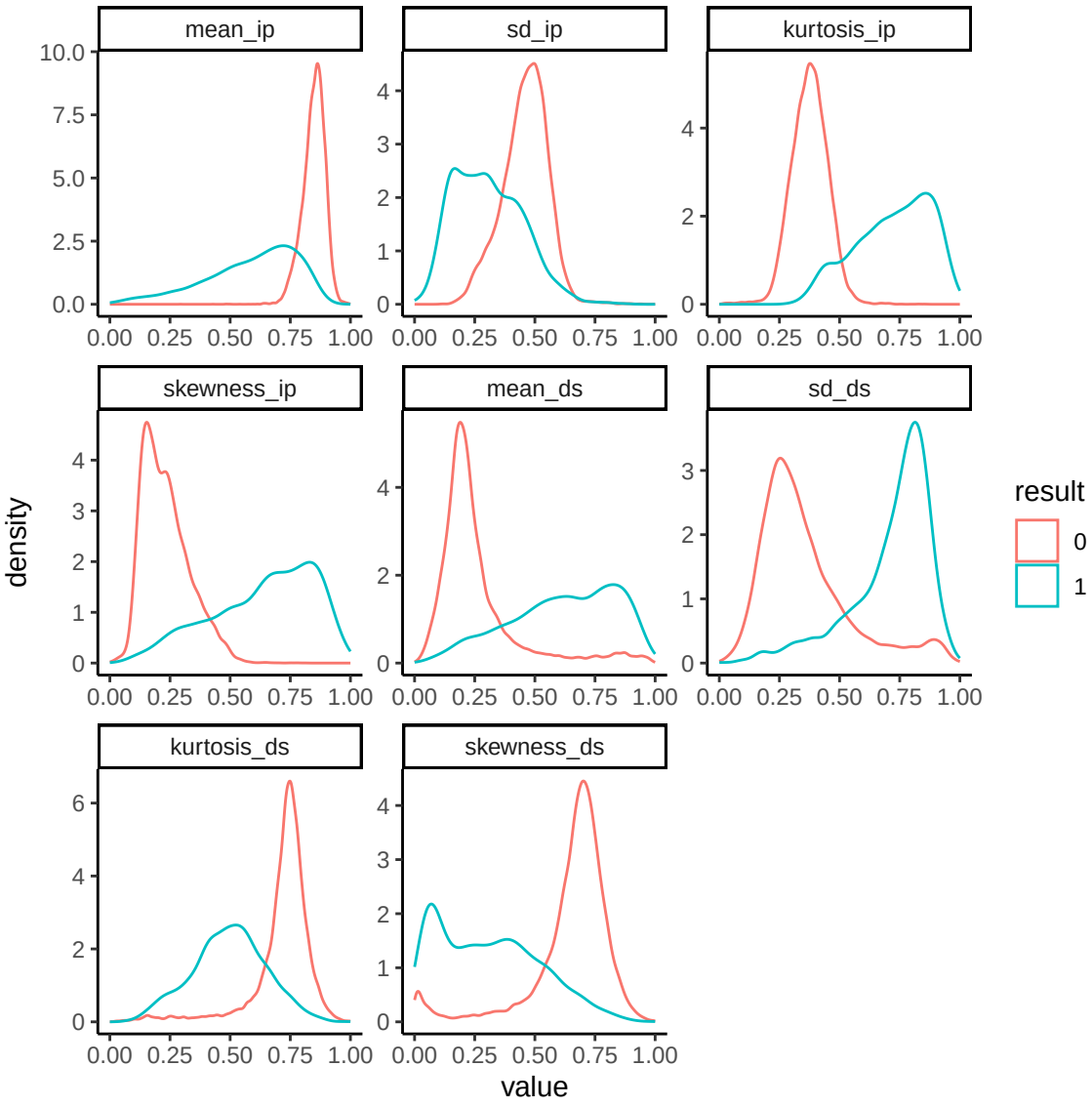


Figure 7: Normalized density plots comparing pulsar candidates vs non-pulsars candidates

3 Logistic Regression

The first model we build is a logistic regression model which uses all 8 features. From the output, we can see that the model indexes heavily on the kurtosis of the integrated pulse profiles.

Table 1: Logistic Model Coefficent

	x
(Intercept)	-39.8155233
mean_ip	8.8879814
sd_ip	4.0174345
kurtosis_ip	26.9537405
skewness_ip	3.0309738
mean_ds	0.8160058
sd_ds	14.5160718
kurtosis_ds	2.1412300
skewness_ds	10.7681070

4 Perceptron

The next model we build is a perception using a logistic activation function. Again, we see that the kurtosis of the integrated pulse profiles has the largest weight.

5 Decision Tree

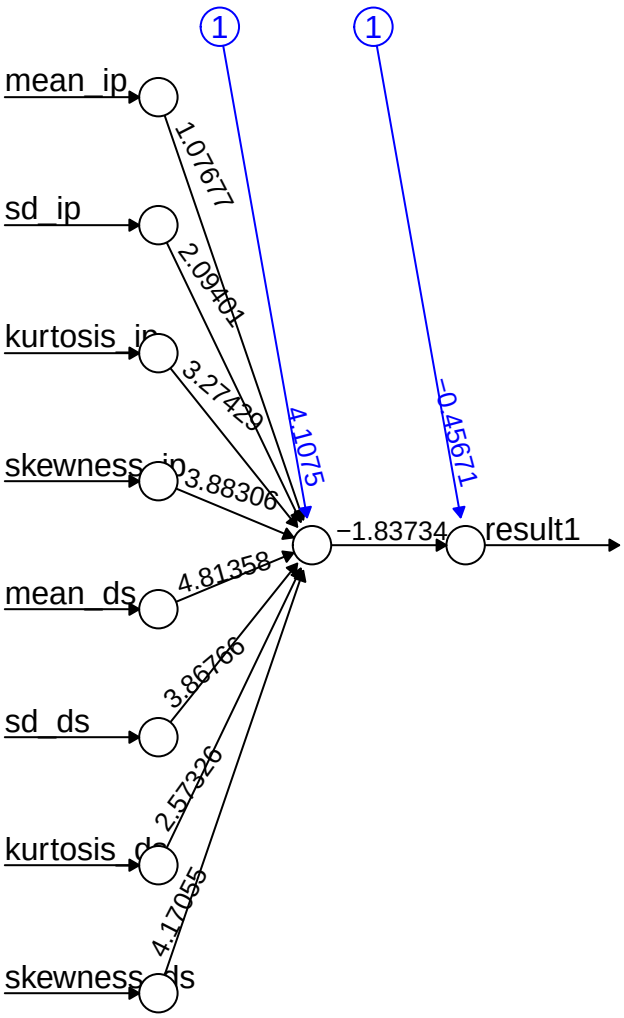
To build our third set of models, we will use decision trees. If we build a simple model where false negatives and false positives are equally penalized, we will get only 1 split based on kurtosis of the integrated pulse profile. But, we do not want an equal penalty because in our case, false negatives are worse than false positive. This is because machine learning models are only 1 step in the verification process for detecting pulsars. We would like to use our model to remove as many negative observations while keeping a small subset which can than we verified by a person. Thus, we build to decision tree models, one where false negatives are penalized 10x more than false positive and another with a 50x penalty.

6 Bootstrap

Finally, we build a bagging model which creates 200 trees and assigns a 10x penalty to false negatives.

7 Conclusion

Below, we plot the ROC curves for all 5 our models. Notice that all 5 perform fairly well with our perceptron model having a slight edge. There is a small drop off in performance with our decision tree models but this is expected due to our penalty settings. In fact, whatever cutoff we use, we would like to minimize the the false negatives.



Error: 595.897252 Steps: 48

Figure 8: Neural Network Model

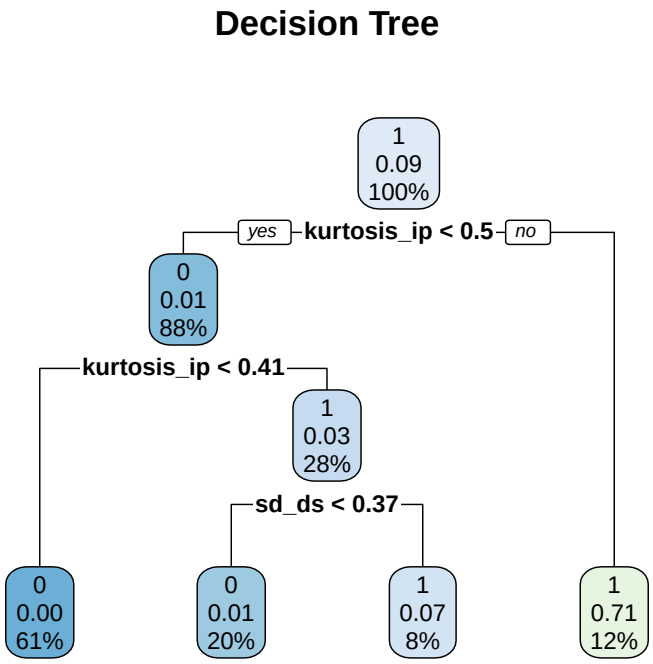


Figure 9: Decision Tree where false negatives are penalized 50 to 1 compared to false positives

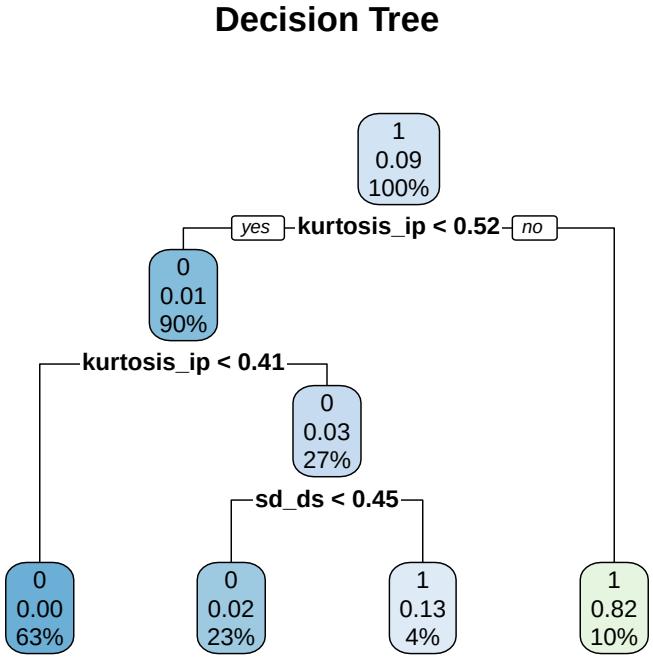


Figure 10: Decision Tree where false negatives are penalized 10 to 1 compared to false positives

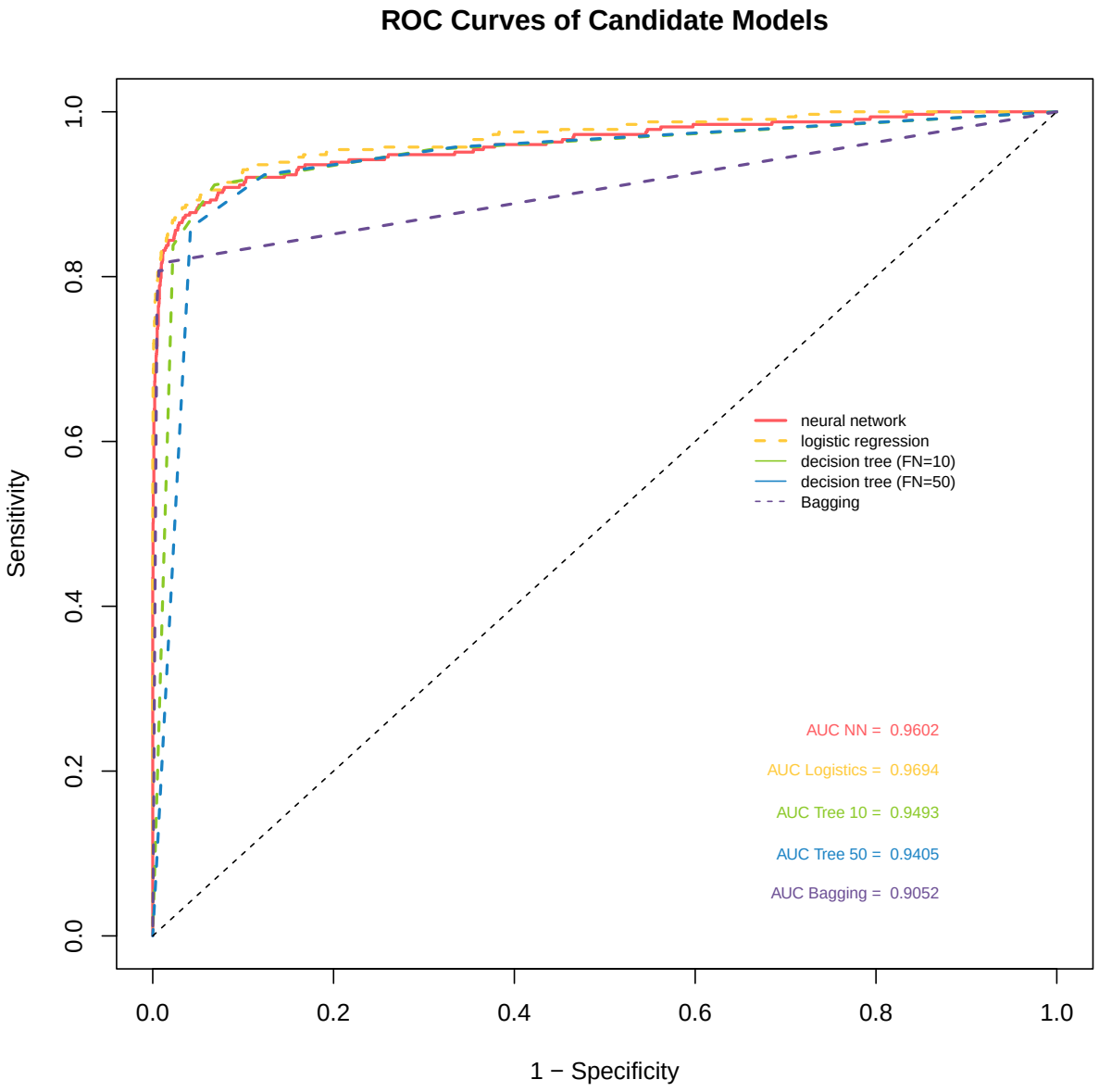


Figure 11: ROC curves of our 5 models

8 References

1. Lyon, R. J., Stappers, B. W., Cooper, S., Brooke, J. M., & Knowles, J. D. (2016). 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(1), 1104-1123.
2. Wang, Y., Pan, Z., Zheng, J., Qian, L., & Li, M. (2019). A hybrid ensemble method for pulsar candidate classification. Astrophysics and Space Science, 364, 1-13.