Pulsar detection - Machine learning and pattern recognition exam

Alberto Baroso - s296520

2022-07-13

Abstract

The Pulsar dataset [1] is a collection of 17.898 pulsar candidates of which 1.639 are real pulsars and 16,259 are spurious examples caused by RFI/noise. A pulsar is a neutron star that emits beams of electromagnetic radiation. As pulsars rotate, their emission beam sweeps across the sky, a periodically repeated pattern of broadband radio emission can be detected as this beam crosses our line of sight. Thus pulsar search involves looking for periodic radio signals with large radio telescopes. The goal of this project is to use machine learning techniques to solve the binary classification problem of detecting pulsars from the Pulsar dataset.

Contents

1 Problem analysis

1.1 Features

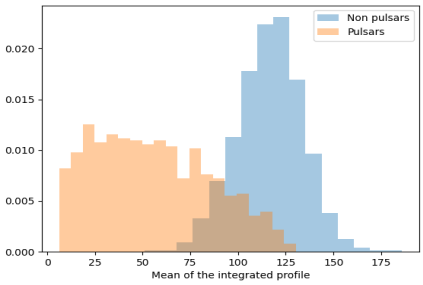
Pulsar candidates are described by a class label and 8 continuous features extracted from radio signals collected by radio telescopes.

The first four are simple statistics obtained from the integrated pulse profile (folded profile). This is an array of continuous features that describe a longitude-resolved version of the signal that has been averaged in both time and frequency (see [2] for more details). The remaining four features are similarly obtained from the DM-SNR curve (again see [2] for more details). The features are listed below:

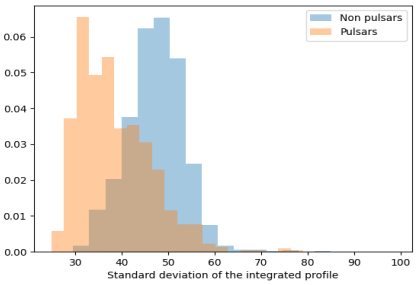
1. Mean of the integrated profile.
2. Standard deviation of the integrated profile.
3. Excess kurtosis of the integrated profile.
4. Skewness of the integrated profile.
5. Mean of the DM-SNR curve.
6. Standard deviation of the DM-SNR curve.
7. Excess kurtosis of the DM-SNR curve.
8. Skewness of the DM-SNR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

The class labels used are 0 (non pulsar) and 1 (pulsar).

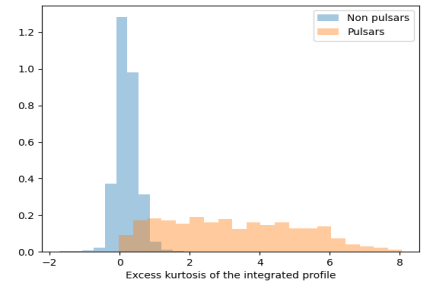
1.2 Feature distributions and ranges



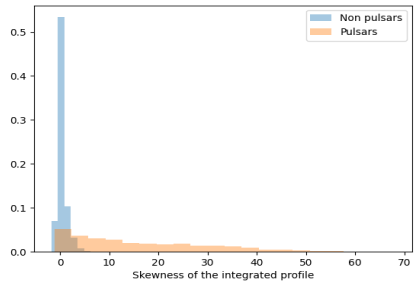
(a)



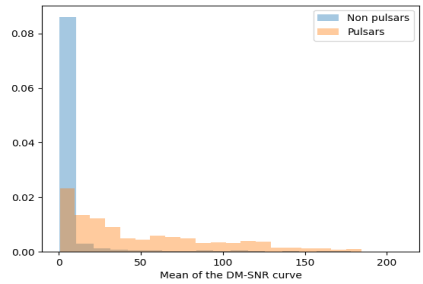
(b)



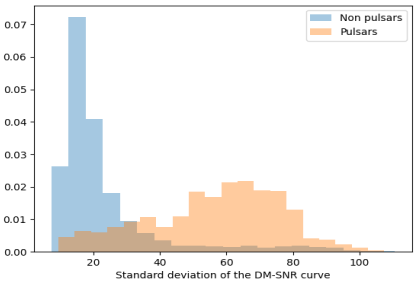
(c)



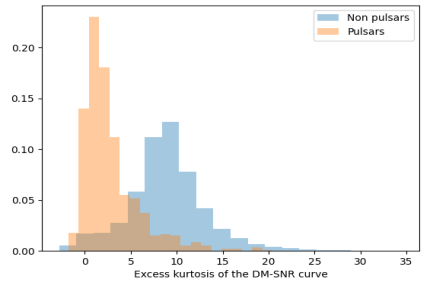
(d)



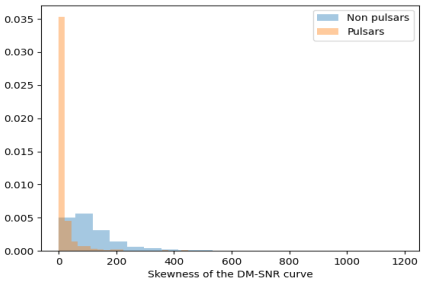
(e)



(f)



(g)



(h)

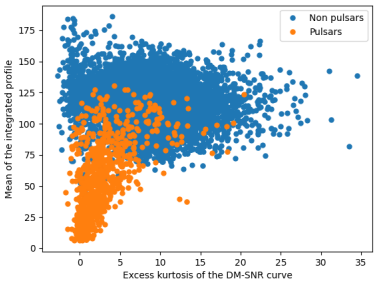
	Feature	Mean	Min	Max
a.	Mean of the integrated profile	110.86	6.18	186.02
b.	Standard deviation of the integrated profile	46.47	24.77	98.78
c.	Excess kurtosis of the integrated profile	0.49	-1.73	8.07
d.	Skewness of the integrated profile	1.84	-1.79	68.10
e.	Mean of the DM-SNR curve	12.67	0.21	209.30
f.	Standard deviation of the DM-SNR curve	26.25	7.37	110.64
g.	Excess kurtosis of the DM-SNR curve	8.33	-2.81	34.54
h.	Skewness of the DM-SNR curve	105.41	-1.98	1191.00

Table 1: Feature ranges

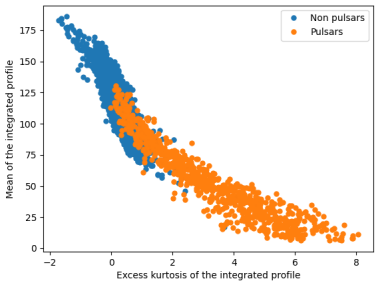
An initial analysis of the raw features of the training data shows that:

- Features have mainly irregular distributions
- Few features, the ones for non pulsars in figures (a), (b) and (g), present distributions similar to Gaussians
- Looking at the graph (b) it's possible to deduce that the "Standard Deviation of the integrated profile" could be one of the less informative features as the histograms of the two classes share one of the largest overlapping areas compared to the other graphs. But since it also has a good amount of non overlapping areas it can still provide useful information for classification purposes.
- Grphs (c) and (d) show that "Excess of kurtosis of integrated profile" and "Skewness of integrated profile" could be among the best features for sample discrimination, as they have little overlapping areas of the histograms.
- Some features, mainly the ones in graphs (b), (d) and (h), are characterized by the presence of significantly large outliers. Similar claims can be made by looking at the ranges of minimum and maximum values for the corresponding features
- We can expect classification algorithms to provide less than ideal results due to the presence of outliers.

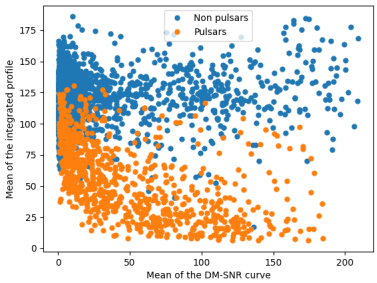
1.3 Feature pairs



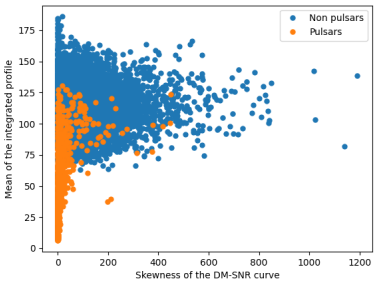
(i)



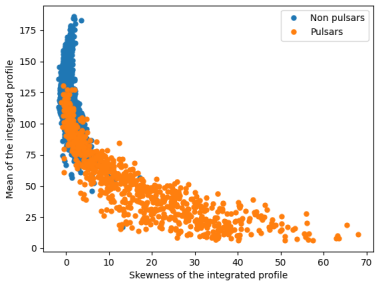
(j)



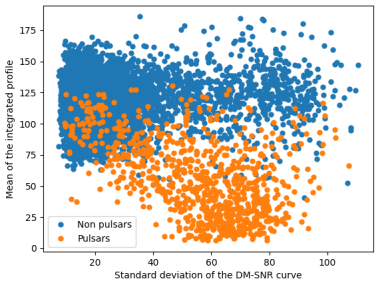
(k)



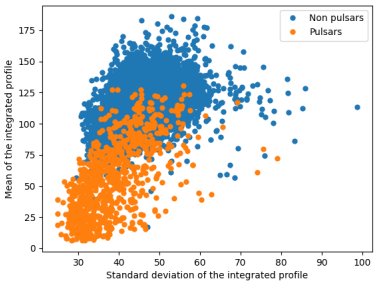
(l)



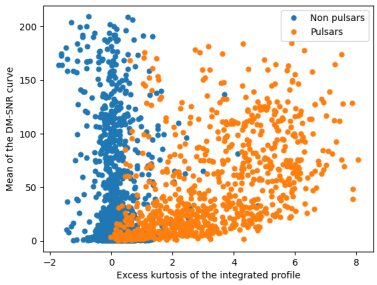
(m)



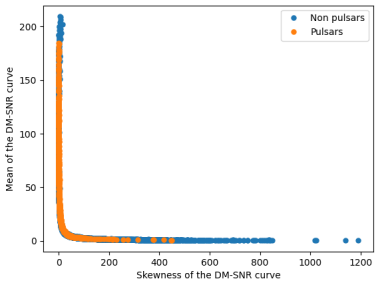
(n)



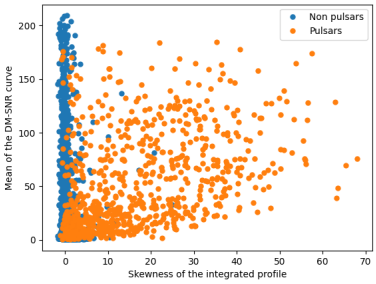
(o)



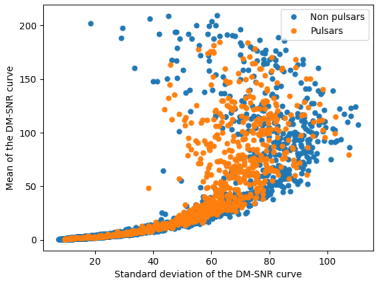
(p)



(q)

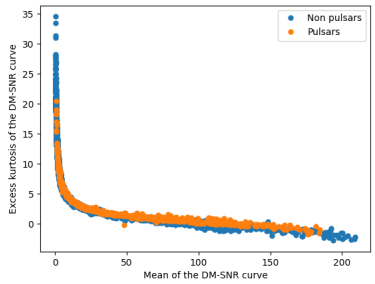


(r)

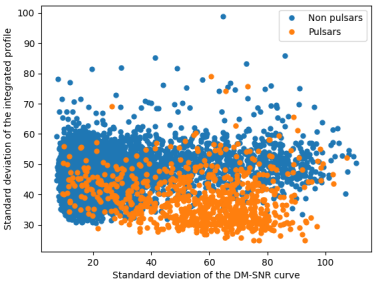


(s)

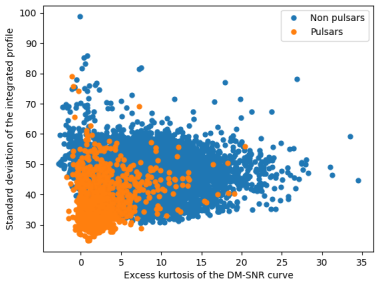
(t)



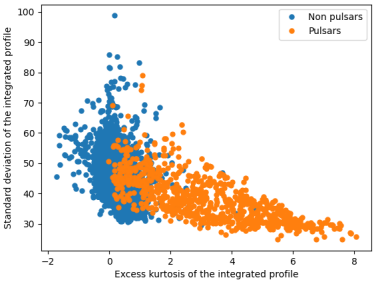
(u)



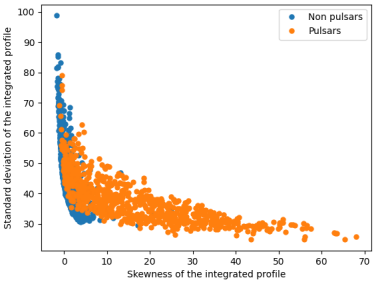
(v)



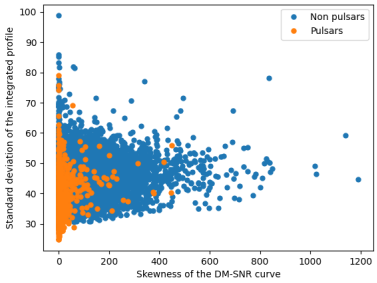
(w)



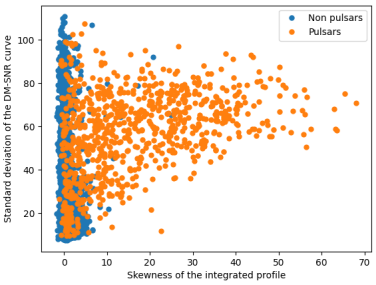
(x)



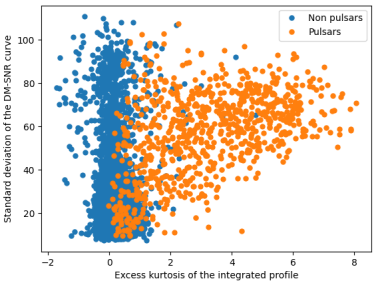
(y)



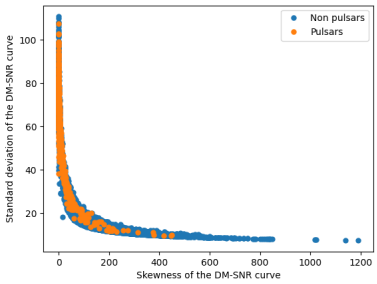
(z)



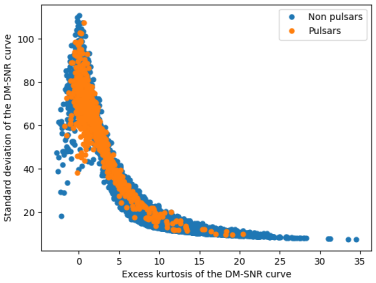
(aa)



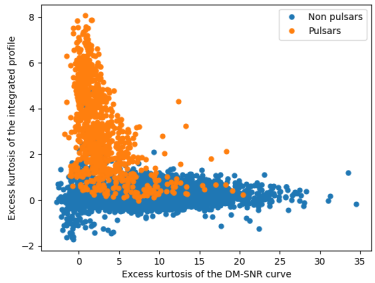
(ab)



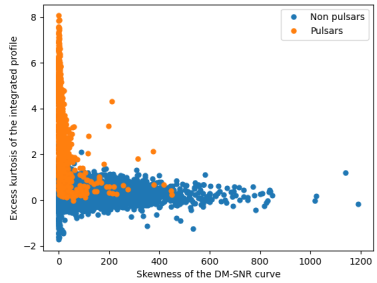
(ac)



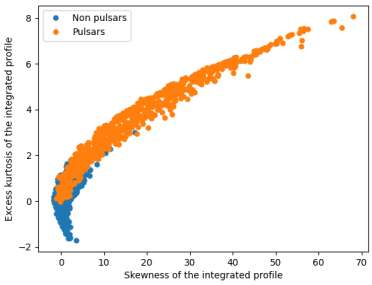
(ad)



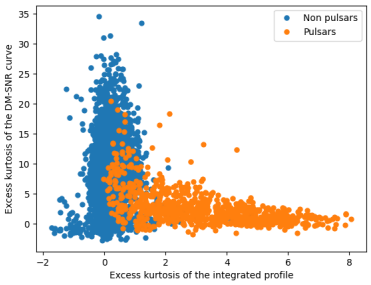
(ae)



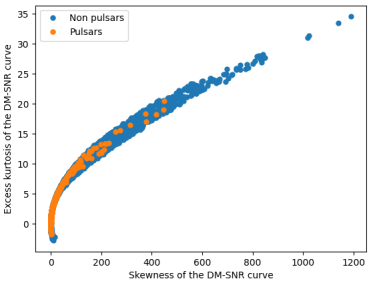
(af)



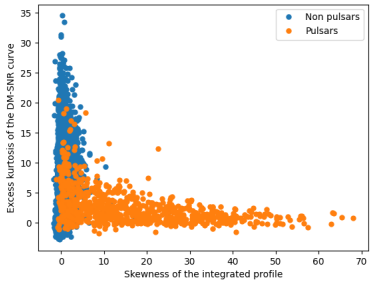
(ag)



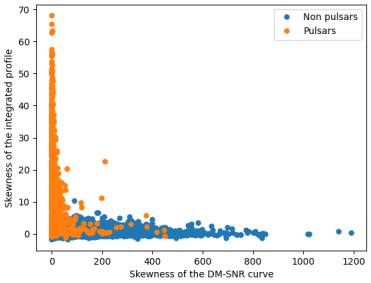
(ah)



(ai)



(aj)



(ak)

Observing the feature pairs it's possible to note that there isn't one that allows to find a clear separation of the two classes.

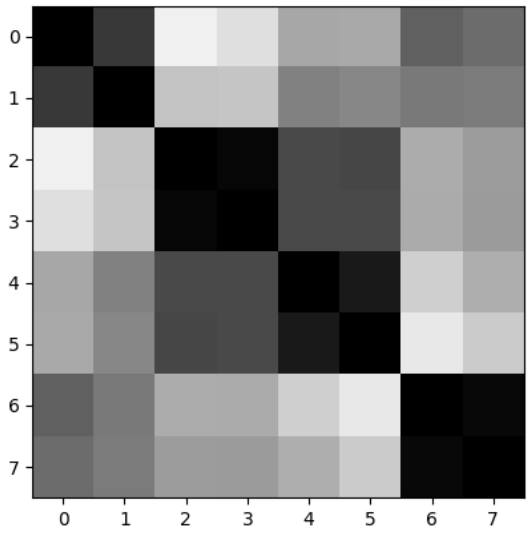
The combination of features represented in graphs (n), (p), (ab), (af) and (ah) get close to identify regions of space where samples are well separated.

The least useful pairs of features are represented in graphs (q), (s), (u) and (ac), where samples are almost entirely overlapping.

The plots also show that some pair of features share similar distributions, implying that there's a correlation between them.

1.4 Feature correlation

Computing the pearson correlation between all pairs of raw features and plotting it as a heatmap we obtain the following result:



(a) Pearson correlation heatmap RAW

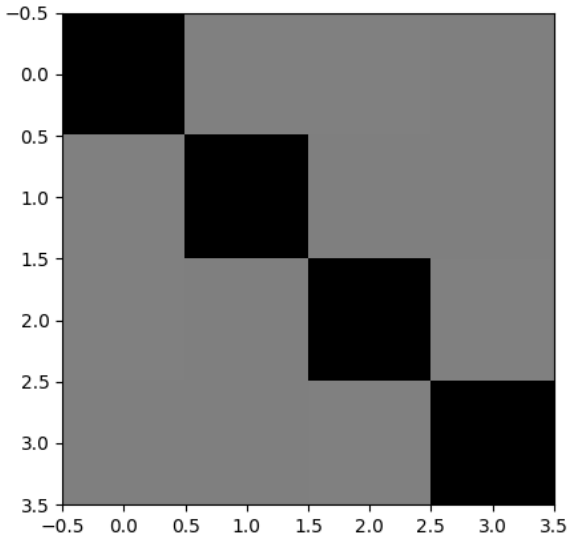
The heatmap clearly shows that the following pairs of features are strongly correlated:

1. (0) Mean of the integrated profile - (1) Standard deviation of the integrated profile.
2. (2) Excess kurtosis of the integrated profile - (3) Skewness of the integrated profile.
3. (4) Mean of the DM-SNR curve - (5) Standard deviation of the DM-SNR curve.
4. (6) Excess kurtosis of the DM-SNR curve - (7) Skewness of the DM-SNR curve.

While the following pairs of features are weakly correlated, with a pearson coefficient close to 0.4:

1. (2) Excess kurtosis of the integrated profile - (4) Mean of the DM-SNR curve
2. (2) Excess kurtosis of the integrated profile - (5) Standard deviation of the DM-SNR curve
3. (3) Skewness of the integrated profile - (4) Mean of the DM-SNR curve
4. (3) Skewness of the integrated profile - (5) Standard deviation of the DM-SNR curve

The results from the correlation analysis suggest that models might benefit from applying a dimensionality reduction technique such as PCA to reduce the feature space to 4 uncorrelated features.



(b) Pearson correlation heatmap PCA

Recomputing the pearson correlation and plotting it as a heatmap it's possible to see that the remaining 4 features are now uncorrelated.

References

[1] R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, and J. 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(1):1104–1123, 04 2016.

[2] R. J. Lyon. *Why are pulsars hard to find*. PhD thesis, University of Manchester, 2016.

Linear Discriminant Analysis (LDA) is deemed not useful as a preprocessing step for this project as LDA allows to find at most c - 1 discriminant directions, where c is the number of classes, thus we could find at most 1 direction.

We expect the MVG with naive bayes assumption model to perform worse that the others on the raw feature because we have seen from the correlation heatmap that some features are strongly correlated and by definition the naive bayes model assumes that the features are independent. One other reason to believe that this model will have poor perform is that it is usually employed when the samples are few and dimensions are a lot