Index

[Introduction 2](#_Toc118383365)

[HTRU2 Data Set 2](#_Toc118383366)

[Aalysis of the features 3](#_Toc118383367)

[References 4](#_Toc118383368)

# Introduction

The goal of this report is to analyze the results of different machine learning algorithms applied to the HTRU2 Data Set [1].

Initially an analysis of the feature will be employed. Then will be applied gaussian, logistic regression, SVM and GMM models….

## HTRU2 Data Set

HTRU2 is a data set which describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey (South) [2].

**Pulsars** are a rare type of Neutron star that produce radio emission detectable here on Earth.

As pulsars rotate, their emission beam sweeps across the sky, and when this crosses our line of sight, produces a detectable pattern of broadband radio emission. As pulsars rotate rapidly, this pattern repeats periodically. Thus pulsar search involves looking for periodic radio signals with large radio telescopes.

Each sample in the HTRU2 Data set is a **'candidate'**, a potential signal detection known averaged over many rotations of the pulsar, as determined by the length of an observation. In the absence of additional info, each candidate could potentially describe a real pulsar. However in practice almost all detections are caused by radio frequency interference (RFI) and noise, making legitimate signals (actual pulsar) hard to find.

This HTRU2 contains a set of pulsar candidates, the actual pulsar are labeled with 1, while the others samples are labeled with 0. More in details the HTRU2 data set contains 16,259 spurious examples caused by RFI/noise, and 1,639 real pulsar examples. These examples have all been checked by human annotators. The HTRU2 is thus **highly imbalanced**: the legitimate pulsar examples are a minority positive class, and spurious examples the majority negative class.

**Classification systems** which treat the candidate data sets as **binary classification problems** can be employed to automatically label pulsar candidates to facilitate rapid analysis. Results of these classifiers are reported in the next sections of this document.

Each candidate is described by 8 continuous variables. The first four are simple statistics obtained from the integrated pulse profile (folded profile). This is an array of continuous variables that describe a longitude-resolved version of the signal that has been averaged in both time and frequency (see [3] for more details). The remaining four variables are similarly obtained from the DM-SNR curve (again see [3] for more details). Each samples has the features (where the last one is the label) summarized 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 HTRU2 contains 17,898 total examples: 1,639 positive examples (labeled with 1) and 16,259 negative examples (labeled with 0).

The dataset has been split into **Train** and Evaluation (**Test**) data.

The training set contains 8108 negative examples and 821 positive examples.

The test set contains 8151 negative examples and 818 positive examples

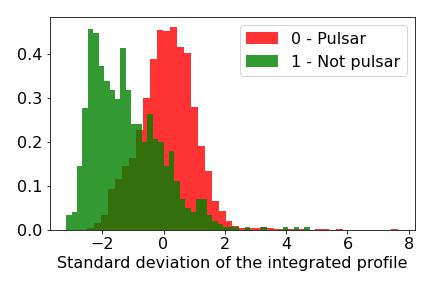
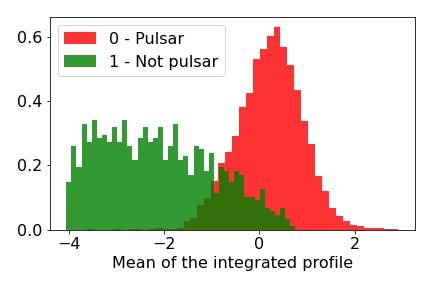
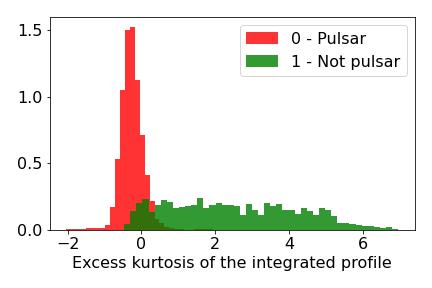
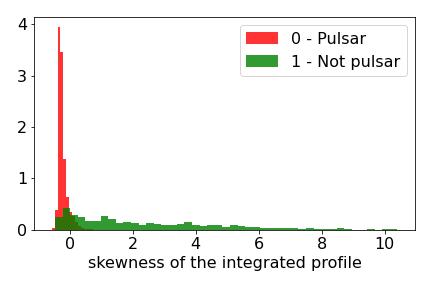
# Analysis of the features

HTRU2 contains 8 continuous variables. The mean (µ) and standard deviation (σ) of the eac features for the training dataset are:

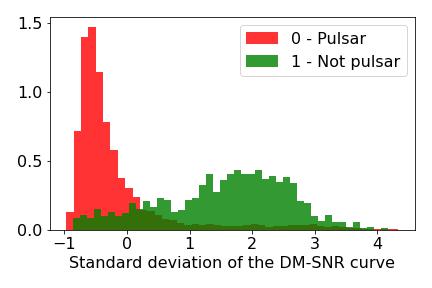
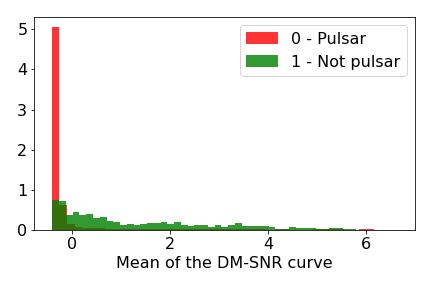
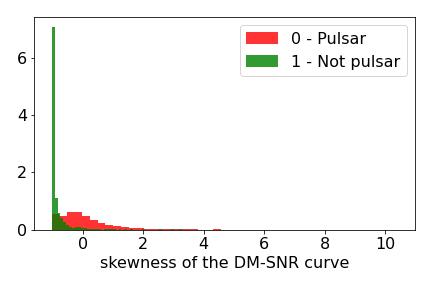
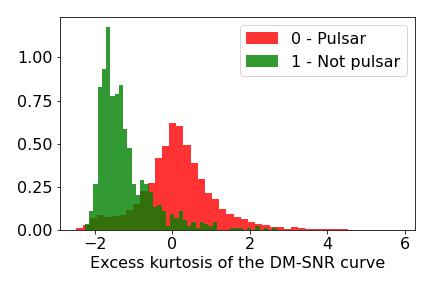
It can be observed that features of the data set have different scales, they have large differences between their ranges. So, in this case, **Z-normalization** on the data-set to bring all the features on the same scale. Using Z-normalization, we center the feature columns at mean 0 with standard deviation 1. Thus before applying any operation each sample of the training set has been transformed through the expression:

Where **x’** is the sample after the Z-score normalization, while **x** is the original sample in the original data set.

The first thing we can do is plotting the different features of different classes for the entire training set. Below are shown the histogram of the features



The 8 features result easy separable. The histograms shows that in the most of cases the features of the class 0 are concentrated around certain values. The shape of the histograms of features of class 0 is expected to be good for some classifiers we’ll implement (gaussian classifiers). On the other hand the features of the class 1 results more spread and irregular. So, it could worth to apply **Gaussianization** to the features. This is a pre-processing technique which helps to deal with the fact that some distributions are not Gaussian like. Since one of the technique we will apply is the Gaussian classifier, we will apply Gaussianization pre-processing stage which transforms our features sample as if they behave more likely to belong to a Gaussian distribution



# References

[1] https://archive.ics.uci.edu/ml/datasets/HTRU2

[2] M. J. Keith et al., 'The High Time Resolution Universe Pulsar Survey - I. System Configuration and Initial Discoveries',2010, Monthly Notices of the Royal Astronomical Society, vol. 409, pp. 619-627. DOI: 10.1111/j.1365-2966.2010.17325.x

[3] R. J. Lyon, 'Why Are Pulsars Hard To Find?', PhD Thesis, University of Manchester, 2016.