# Introduction

## Description

Pulsars are kind of Neutron stars and considerably interesting for scientific research. As this exceptional kind of star produces radio emissions detectable here on Earth, machine learning tools can be used to label pulsar candidates to facilitate rapid analysis automatically. Classification systems are widely implemented, considering the candidate data sets as **binary classification problems**.

## Candidate information

Each candidate is described by eight **continuous** **variables** and a single class variable. 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. The remaining four variables are similarly obtained from the DM-SNR curve.

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

## Classification task

Given the features above, our final task is to determine whether each of the samples is a Pulsar candidate or not. So, we are faced with a **binary classification problem,** and we need to build different classifiers, analyze them, and compare their performance and cost for various applications.

# Feature and class analysis

## Class distribution

By looking at the distribution of each class, we can realize that the Dataset is **highly imbalanced**. There are <number> non-pulsars and <number> pulsars in the training set. The Dataset Might need sampling before the model-building process.

## Feature distribution

Now we turn our attention to the histogram of features. Plots for the integrated profile are presented on the left side, and the DM-SNR curve plots are on the right. The orange color represents the non-pulsars, and the blue color pulsars.

|  |  |
| --- | --- |
| Chart  Description automatically generated | Chart, histogram  Description automatically generated |
| Chart, histogram  Description automatically generated | Chart, histogram  Description automatically generated |
| Chart  Description automatically generated | Chart  Description automatically generated |
| Chart, histogram  Description automatically generated | Chart, histogram  Description automatically generated |

Apart from the mean and standard deviation of the integrated profile, all other features are highly right-skewed. Overall, the dataset is not normally distributed.

The SD of the integrated profile appears normal but has a tail on the right side, and only the mean of the Integrated Profile has a heavy tail on the left side and is left-skewed.

The mean of the DM-SNR curve, Standard deviation of the DM-SNR curve, and Skewness of the integrated profile seem to have more outliers than other features. This analysis suggests that Gaussianized features might produce better results.

## Correlation Analysis

The correlation heatmap taken from the absolute value of the Pearson correlation coefficient describes positive and negative correlations between features. (Whole dataset)

Highly positively correlated:

* The skewness of the integrated profile and Excess kurtosis of the DM-SNR curve
* The skewness of the DM-SNR curve and Excess kurtosis of the DM-SNR curve
* The Mean of the DM-SNR curve and Standard Deviation of the DM-SNR curve

Highly negatively correlated:

* The Mean of the integrated profile and Excess kurtosis of the integrated profile
* The Mean of the integrated profile and Skewness of the integrated profile
* The Excess kurtosis of the DM-SNR curve and Standard Deviation of the DM-SNR curve

Chart, histogram

Description automatically generated

We can use PCA to map data to 6 or 5 uncorrelated features to reduce the amount of computation and the number of parameters to estimate. But removing 2 or 3 features, given the limited amount of data we have, might also decrease the accuracy of our models.

# Classification

## Applications and Cross-validation

Before analyzing different classifiers, we need to select applications and the cross-validation method for our analysis. As we are not dealing with a large dataset, re-training the models doesn’t have a considerable cost. K-fold **cross-validation** would use additional data and give us more accurate results than a single split.

Our main application will be a uniform prior one:

We will also consider unbalanced applications in which the prior is biased towards one of the two classes:

At first, we measure performance in terms of normalized minimum detection costs. The cost we would pay if we made optimal decisions for the validation set. We will consider how to choose an optimal threshold in the second stage.

## MVG Classifiers

MVG Classifiers – minDCF on the validation set

|  |  |  |  |
| --- | --- | --- | --- |
|  | 5-fold | | |
|  |  |  |  |
|  | Gaussianized features - no PCA | | |
| Full-Cov | 0.10 | 0.10 | 0.10 |
| Diag-Cov | 0.10 | 0.10 | 0.10 |
| Tied Full-Cov | 0.10 | 0.10 | 0.10 |
| Tied Diag-Cov | 0.10 | 0.10 | 0.10 |
|  | Gaussianized features - PCA (m=7) | | |
| Full-Cov | 0.10 | 0.10 | 0.10 |
| Diag-Cov | 0.10 | 0.10 | 0.10 |
| Tied Full-Cov | 0.10 | 0.10 | 0.10 |
| Tied Diag-Cov | 0.10 | 0.10 | 0.10 |
|  | Gaussianized features - PCA (m=6) | | |
| Full-Cov | 0.10 | 0.10 | 0.10 |
| Diag-Cov | 0.10 | 0.10 | 0.10 |
|  | Raw features - no PCA | | |
| Full-Cov | 0.10 | 0.10 | 0.10 |

# Experimental Validation

# Conclusion