PHASE 1 : Data analysis and preparation

Prathyusha Velupula

December 1, 2023

Overleaf Link for this report is here
Link for video description is here

Contents

List of Figures

1 Introduction

The dataset for this project is chosen from here. HTRU2 is a data set which describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey (South). The data set shared here contains 16,259 spurious examples caused by RFI/noise, and 1,639 real pulsar examples. These examples have all been checked by human annotators. 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. The remaining four variables are similarly obtained from the DM-SNR curve . These are summarised below:

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

HTRU 2 Summary: 17,898 total examples ,1,639 positive examples and 16,259 negative examples.
The data is presented in two formats: CSV and ARFF (used by the WEKA data mining tool). Candidates are stored in both files in separate rows. Each row lists the variables first, and the class label is the final entry. The class labels used are 0 (negative) and 1 (positive).
Please not that the data contains no positional information or other astronomical details. It is simply feature data extracted from candidate files using the PulsarFeatureLab tool

2 Cleaning of data

The dataset downloaded is cleaned and updated using methods given here. The updated csv file is then loaded as a pandas dataframe in the jupyter notebook.

3 Processing and analysing of data

3.1 Visualization of distribution of input varibales

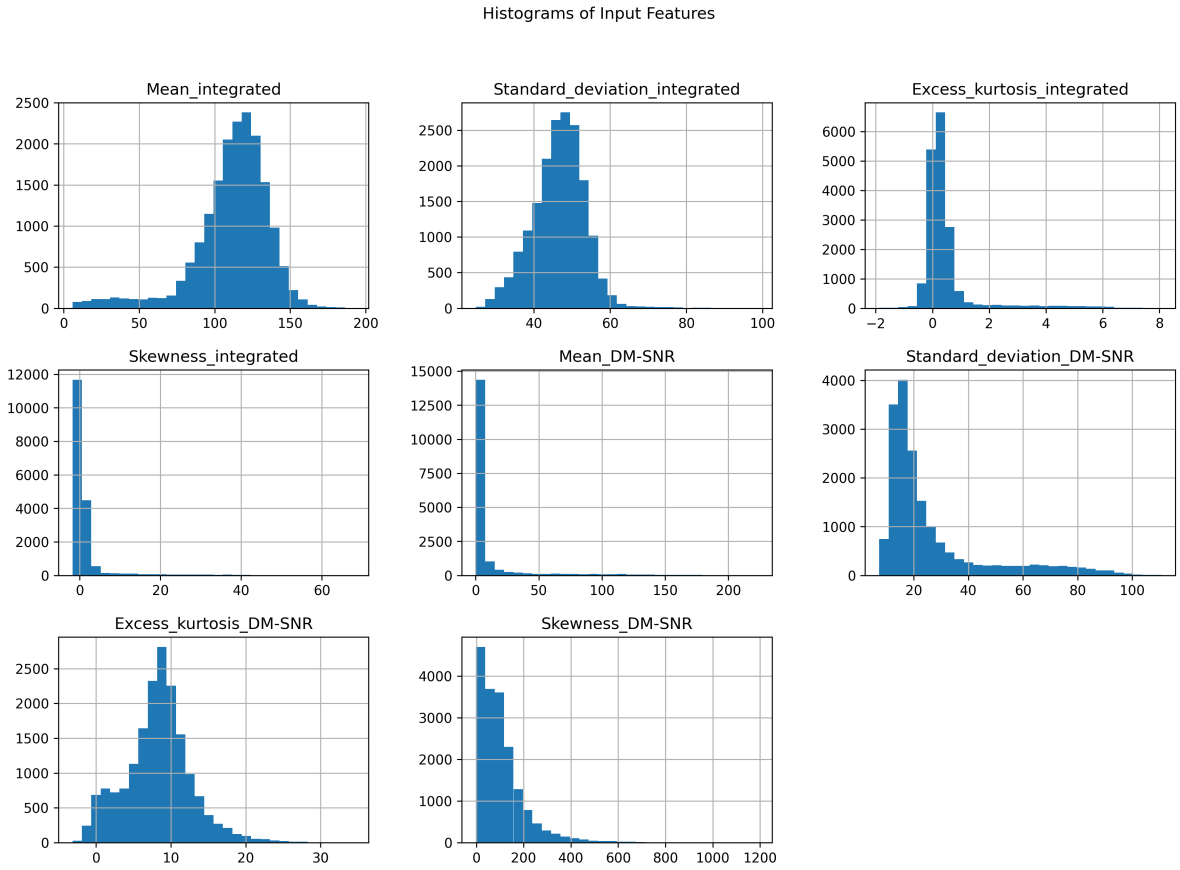

Figure 1: Histograms of Distribution of input features - before normalization

3.2 Distribution of Output label

Below figure shows how class feature is distributed (see **Figure 2**).

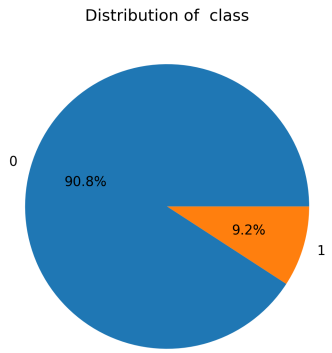

Figure 2: Pie diagram of Distribution of class label

3.3 Data normalization

NOTE: Data normalization is done before shuffling and splitting the data Data pre-processing is essential before data mining to address the non-uniform distribution of data. To achieve this, normalization techniques are used to make the optimization problem more numerically stable and improve training. Normalization helps to ensure all values lie between 0 and 1 and outliers are visible within the normalized data. There are two normalization techniques available, each with its own consequences, but either technique can be used for now.

Mean Data Normalization Formula :

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Z-Scored Normalization:

$$X_{normalized} = \frac{X - Xmean}{X_{standarddeviation}}$$

Before splitting" refers to the fact that the data-set has not yet been divided into separate training and validation sets. In machine learning, it is common to split the data-set into separate training and validation sets, where the training set is used to train the model and the validation set is used to evaluate the performance of the model.
"Before normalization" refers to the fact that the data-set has not yet been scaled to a common range. In machine learning, it is common to normalize or scale the data-set to ensure that all variables are on a similar scale. This is done to prevent variables with larger values from dominating the training process and to ensure that the model is sensitive to all variables equally.

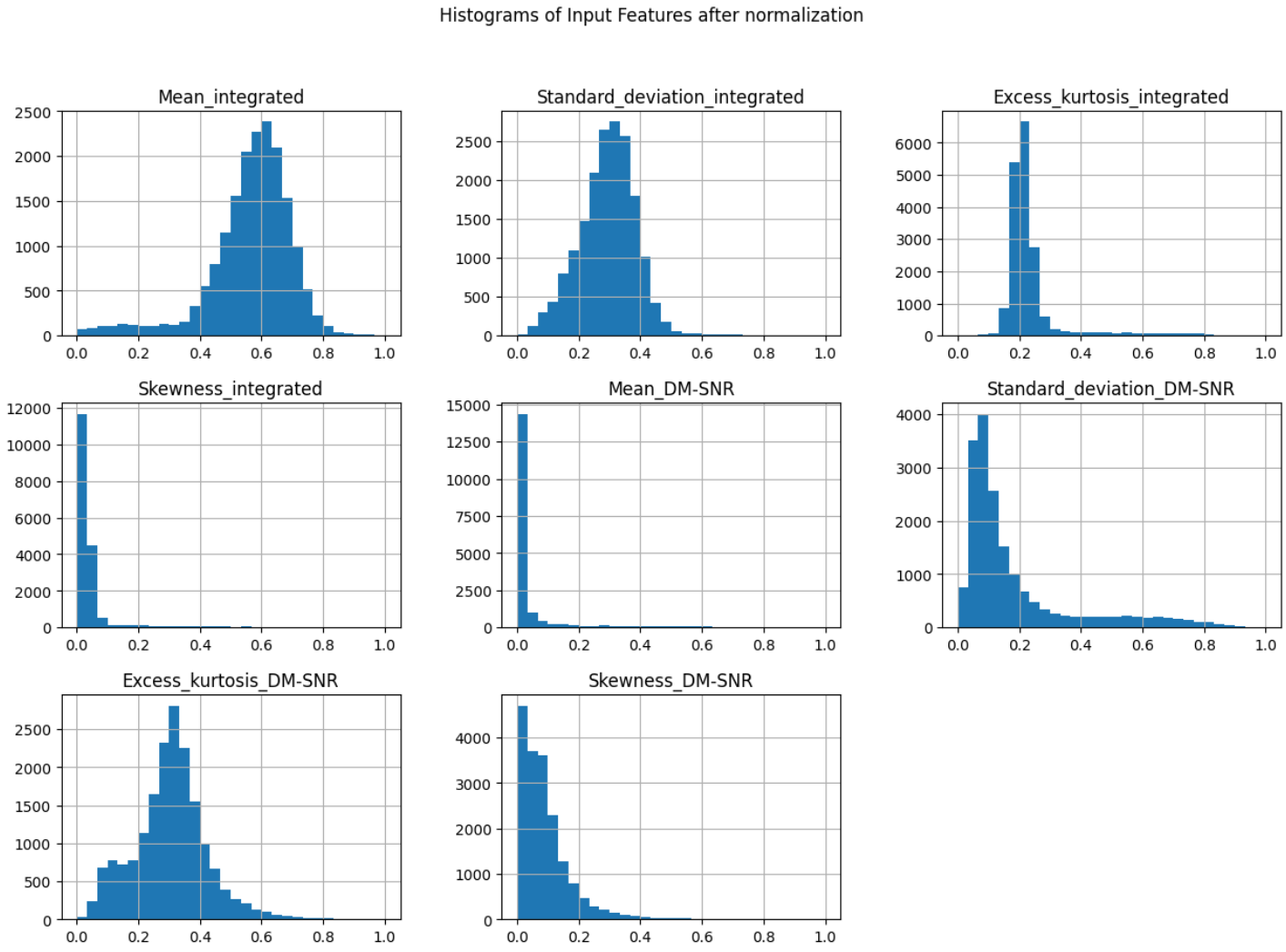

Figure 3: Histograms of Distribution of input features - after normalization

First 10 rows of data before and after normalization

Mean_integrated	Standard_deviation_integrated	Excess_kurtosis_integrated	Skewness_integrated	Mean_DM-SNR	Standard_deviation_DM-SNR	Excess_kurtosis_DM-SNR	Skewness_DM-SNR	Class
140.562500	55.683782	-0.234571	-0.699648	3.199833	19.110426	7.975532	74.242225	0
102.507812	58.882430	0.465318	-0.515088	1.677258	14.860146	10.576487	127.393580	0
103.015625	39.341649	0.323328	1.051164	3.121237	21.744669	7.735822	63.171909	0
136.750000	57.178449	-0.068415	-0.636238	3.642977	20.959280	6.896499	53.593661	0
88.726562	40.672225	0.600866	1.123492	1.178930	11.468720	14.269573	252.567306	0
93.570312	46.698114	0.531905	0.416721	1.636288	14.545074	10.621748	131.394004	0
119.484375	48.765059	0.031460	-0.112168	0.999164	9.279612	19.206230	479.756567	0
130.382812	39.844056	-0.158323	0.389540	1.220736	14.378941	13.539456	198.236457	0
107.250000	52.627078	0.452688	0.170347	2.331940	14.486853	9.001004	107.972506	0
107.257812	39.496488	0.465882	1.162877	4.079431	24.980418	7.397080	57.784738	0

Figure 4: First 10 rows of data - before normalization

Mean_integrated	Standard_deviation_integrated	Excess_kurtosis_integrated	Skewness_integrated	Mean_DM-SNR	Standard_deviation_DM-SNR	Excess_kurtosis_DM-SNR	Skewness_DM-SNR	Class
0.721342	0.417687	0.165043	0.015627	0.013382	0.113681	0.294986	0.063890	0
0.517628	0.460908	0.235415	0.018268	0.006560	0.072524	0.364015	0.108443	0
0.520346	0.196868	0.221138	0.040677	0.013030	0.139188	0.288624	0.054610	0
0.700933	0.437884	0.181750	0.016534	0.015368	0.131583	0.266348	0.046581	0
0.443854	0.214847	0.249044	0.041712	0.004327	0.039684	0.462029	0.213369	0
0.469784	0.296271	0.242110	0.031600	0.006376	0.069473	0.365216	0.111797	0
0.608507	0.324200	0.191792	0.024033	0.003522	0.018487	0.593047	0.403808	0
0.666848	0.203657	0.172710	0.031211	0.004514	0.067865	0.442652	0.167827	0
0.543014	0.376384	0.234145	0.028075	0.009493	0.068910	0.322202	0.092164	0
0.543055	0.198961	0.235472	0.042275	0.017323	0.170521	0.279634	0.050095	0

Figure 5: First 10 rows of data - after normalization