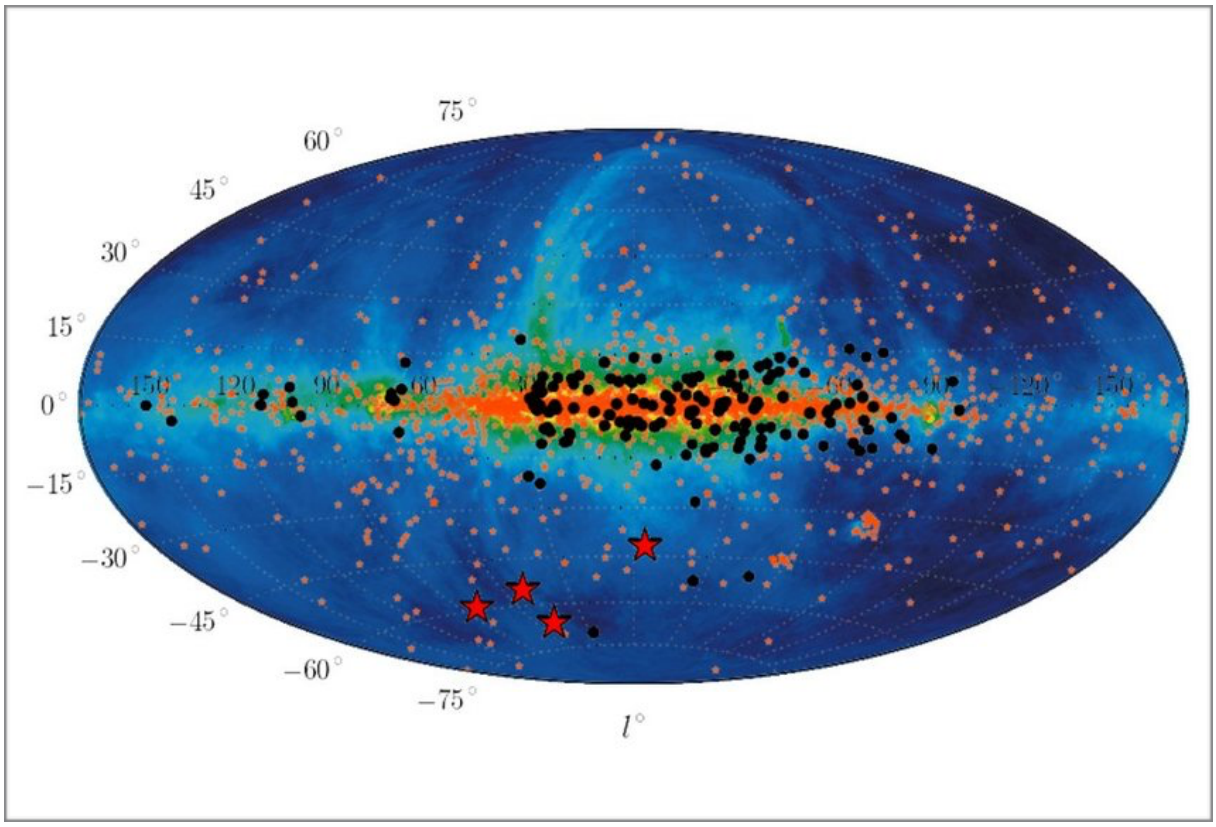
EEE5020: Machine Learning

Capstone Project Report



HTRU2 Dataset | Pulsar Detection

Akhil Punia | Ashish Sardana

14BEE0100 | 14BEE0099

under the guidance of

Prof. Monica Subashini

Winter 2017

Contents

Introduction

Topic Summary and Problem Statement

Pulsars are a rare type of Neutron star that produce radio emission detectable here on Earth. They are of considerable scientific interest as probes of space-time, the inter-stellar medium, and states of matter (see [2] for more uses).

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.

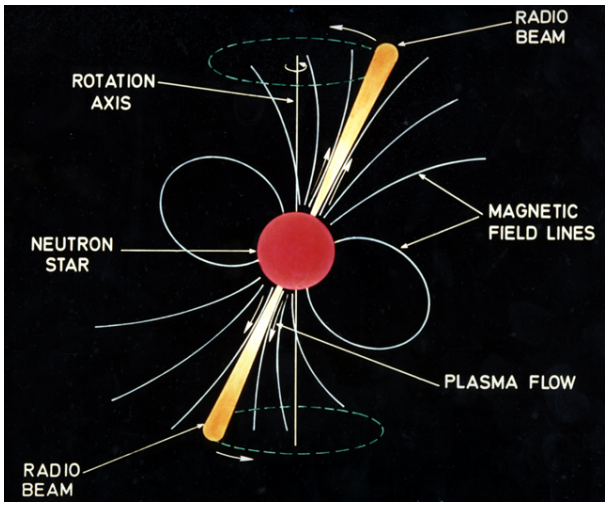


Fig-1: Pulsar Structure

Each pulsar produces a slightly different emission pattern, which varies slightly with each rotation (see [2] for an introduction to pulsar astrophysics to find out why). Thus a potential signal detection known as a 'candidate', is 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 hard to find.

Machine learning tools can now be used to automatically label pulsar candidates to facilitate rapid analysis. Classification systems in particular are being widely adopted, (see [4,5,6,7,8,9]) which treat the candidate data sets as binary classification problems. Here the legitimate pulsar examples are a minority positive class, and spurious examples the majority negative class. At present multi-class labels are unavailable, given the costs associated with data annotation.

About the Dataset

HTRU2 is a data set which describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey (South) [1].

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. 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).

Each candidate is described by 8 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 (see [3] for more details). The remaining four variables are similarly obtained from the DM-SNR curve (again see [3] for more details). These are summarised 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

HTRU 2 Summary :
17,898 total examples.
1,639 positive examples.
16,259 negative examples.

Literature Review

The High Time Resolution Universe surveys for pulsars and fast transients.

The High Time Resolution Universe survey for pulsars and transients is the first truly all-sky pulsar survey, taking place at the Parkes Radio Telescope in Australia and the Ef- felsberg Radio Telescope in Germany. Utilising multibeam receivers with custom built all-digital recorders the survey targets the fastest millisecond pulsars and radio transients on timescales of 64 μs to a few seconds. The new multibeam digital filter-bank system has a factor of eight improvement in frequency resolution over previous Parkes multibeam surveys, allowing researchers to probe further into the Galactic plane for short duration signals. To date, the total number of discoveries in the combined survey is 135 and 29 MSPs These discoveries include the first magnetar to be discovered by it's radio emission, unusual low-mass binaries, gamma-ray pulsars and pulsars suitable for pulsar timing array experiments.

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, DOI: 10.1093/mnras/stw656

Improving survey specifications are causing an exponential rise in pulsar candidate numbers and data volumes. Here, researchers study the candidate filters used to mitigate these problems during the past fifty years. They find that some existing methods such as applying constraints on the total number of candidates collected per observation, may have detrimental effects on the success of pulsar searches. Those methods immune to such effects are found to be ill-equipped to deal with the problems associated with increasing data volumes and candidate numbers, motivating the development of new approaches. Therefore researchers present a new method designed for on-line operation. It selects promising candidates using a purpose-built tree-based machine learning classifier, the Gaussian Hellinger Very Fast Decision Tree (GH-VFDT), and a new set of features for describing candidates. The features have been chosen so as to i) maximise the separation between candidates arising from noise and those of probable astrophysical origin, and ii) be as survey-independent as possible. Using these features, the new approach can process millions of candidates in seconds (~1 million every 15 seconds), with high levels of pulsar recall (90%+). This technique is therefore applicable to the large volumes of data expected to be produced by the Square Kilometre Array (SKA). Use of this approach has assisted in the discovery of 20 new pulsars in data obtained during the LOFAR Tied-Array All-Sky Survey (LOTAAS).

The High Time Resolution Universe Pulsar Survey – VI. An Artificial Neural Network and timing of 75 pulsars , Monthly Notices of the Royal Astronomical Society, vol. 427, no. 2, pp. 1052-1065, 2012

Researchers present 75 pulsars discovered in the mid-latitude portion of the High Time Resolution Universe survey, 54 of which have full timing solutions. All the pulsars have spin periods greater than 100ms, and none of those with timing solutions is in binaries. Two display particularly interesting behaviour; PSR J1054−5944 is found to be an intermittent pulsar, and PSR J1809−0119 has glitched twice since its discovery.

In the second half of the paper, researchers discuss the development and application of an artificial neural network in the data-processing pipeline for the survey. They discuss the tests that were used to generate scores and find that the neural network was able to reject over 99 per cent of the candidates produced in the data processing, and able to blindly detect 85 per cent of pulsars. They suggest that improvements to the accuracy should be possible if further care is taken when training an artificial neural network; for example, ensuring that a representative sample of the pulsar population is used during the training process, or the use of different artificial neural networks for the detection of different types of pulsars.

PEACE: pulsar evaluation algorithm for candidate extraction – a software package for post-analysis processing of pulsar survey candidates, Monthly Notices of the Royal Astronomical Society, vol. 433, no. 1, pp. 688-694, 2013.

Modern radio pulsar surveys produce a large volume of prospective candidates, the majority of which are polluted by human-created radio frequency interference or other forms of noise. Typically, large numbers of candidates need to be visually inspected in order to determine if they are real pulsars. This process can be labour intensive. In this paper, researchers introduce an algorithm called Pulsar Evaluation Algorithm for Candidate Extraction (PEACE) which improves the efficiency of identifying pulsar signals. The algorithm ranks the candidates based on a score function. Unlike popular machine-learning-based algorithms, no prior training data sets are required. This algorithm has been applied to data from several large-scale radio pulsar surveys. Using the human-based ranking results generated by students in the Arecibo Remote Command Center programme, the statistical performance of PEACE was evaluated. It was found that PEACE ranked 68percent of the student-identified pulsars within the top 0.17percent of sorted candidates, 95percent within the top 0.34percent and 100percent within the top 3.7percent. This clearly demonstrates that PEACE significantly increases the pulsar identification rate by a factor of about 50 to 1000. To date, PEACE has been directly responsible for the discovery of 47 new pulsars, 5 of which are millisecond pulsars that may be useful for pulsar timing based gravitational-wave detection projects.

SPINN: a straightforward machine learning solution to the pulsar candidate selection problem, Monthly Notices of the Royal Astronomical Society, vol. 443, no. 2, pp. 1651-1662, 2014.

Here, Researchers describe SPINN (Straightforward Pulsar Identification using Neural Networks), a high- performance machine learning solution developed to process increasingly large data outputs from pulsar surveys. SPINN has been cross-validated on candidates from the southern High Time Resolution Universe (HTRU) survey and shown to identify every known pulsar found in the survey data while maintaining a false positive rate of 0.64 per cent. Furthermore, it ranks 99 per cent of pulsars among the top 0.11 per cent of candidates, and 95 per cent among the top 0.01 per cent. In conjunction with the PEASOUP pipeline, it has already discovered four new pulsars in a re-processing of the intermediate Galactic latitude area of HTRU, three of which have spin periods shorter than 5 ms. SPINN's ability to reduce the amount of candidates to visually inspect by up to four orders of magnitude makes it a very promising tool for future large-scale pulsar surveys. In an effort to provide a common testing ground for pulsar candidate selection tools and stimulate interest in their development, we also make publicly available the set of candidates on which SPINN was cross-validated.

Selection of radio pulsar candidates using Artificial Neural Networks, Monthly Notices of the Royal Astronomical Society, vol. 407, no. 4, pp. 2443-2450, 2010

Radio pulsar surveys are producing many more pulsar candidates than can be inspected by human experts in a practical length of time. Here researchers present a technique to automatically identify credible pulsar candidates from pulsar surveys using an artificial neural network. The technique has been applied to candidates from a recent re-analysis of the Parkes multi-beam pulsar survey resulting in the discovery of a previously unidentified pulsar.

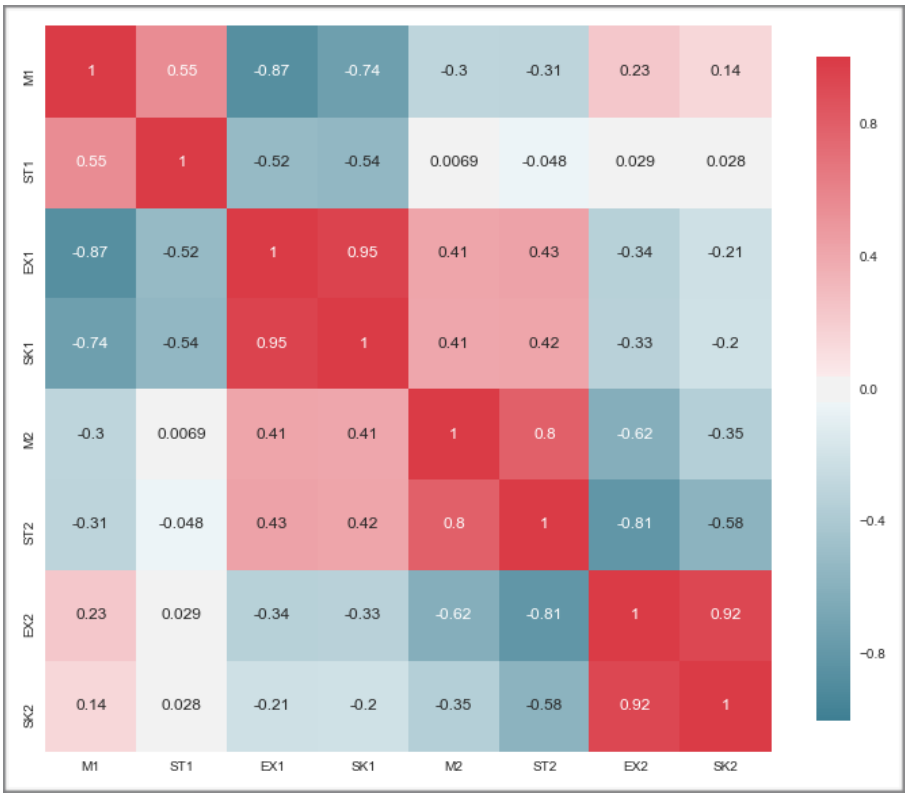
Methodology

The goal of this work is to develop an automatic supervised and unsupervised classification method to achieve faster and more detailed selection and classification of stars into Pulsars or non-Pulsars. The method was devised to supersede the experts' criteria and retrieve in addition to pulsars and non-pulsars, potentially interesting characteristics of stars that remained unexplored because of the use of predetermined templates.

Part I. Visual inspection of candidates:

The dataset consists of features rather than raw data. The pre-processed data gives an added advantage: the computation power required to extract features is not required.

Finding correlation between features gives an insight about the relationship between them. The goal of a correlation analysis is to see whether two measurement variables co vary, and to quantify the strength of the relationship between the variables.



After observing the correlation between the features, it is important to do feature selection. These methods aid in the mission to create an accurate predictive model. They help us by choosing features that will give you as good or better accuracy whilst requiring less data.

Part II. Clustering - An Unsupervised Learning Task

Cluster analysis is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

Clustering generates natural clusters and is not dependent on any driving objective function. Hence such a cluster can be used to analyse the stars based on their similar properties.

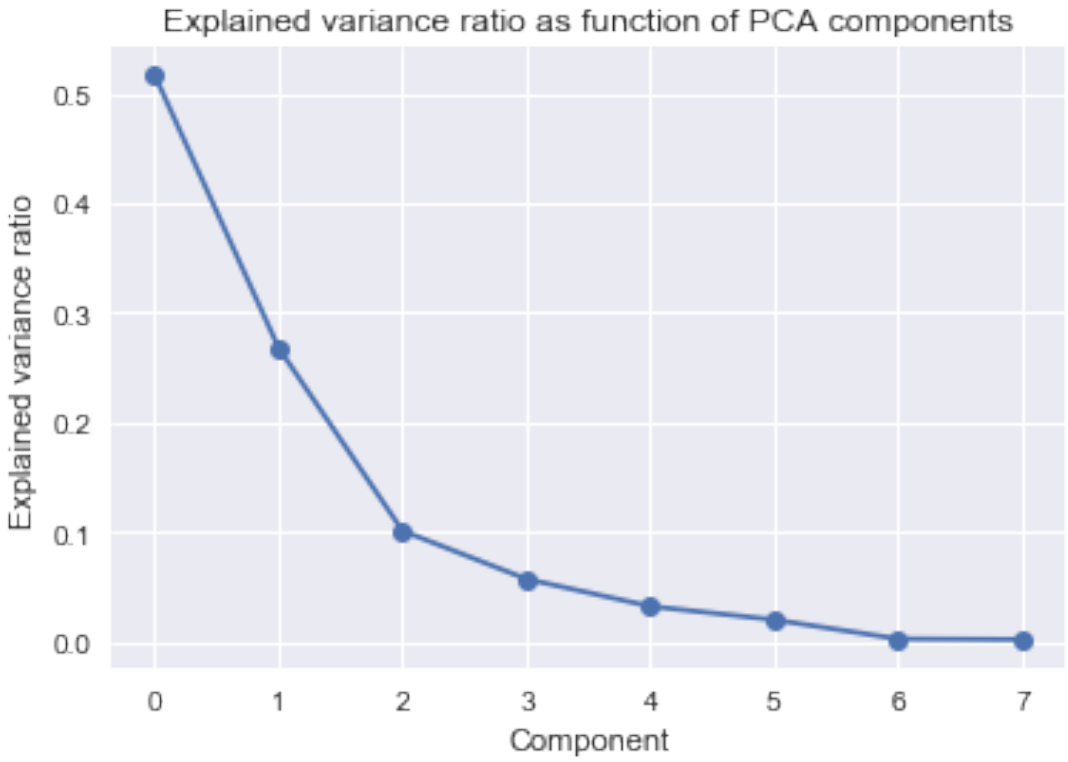
PCA - Principal Component Analysis:

It is a precursor step to any analysis that we may subject to our dataset. In such a high-dimensional space, Euclidean distances tend to become inflated and meaningless. This can severely impact our algorithms performance. Such a situation demands more data to train our model and this problem is called the 'Curse of Dimensionality.'

The PCA algorithm solves this problem by finding out the features that explain the maximum variance. So, instead of training our models over 8 features we will be training them over 'some' features that explain the maximum variance.

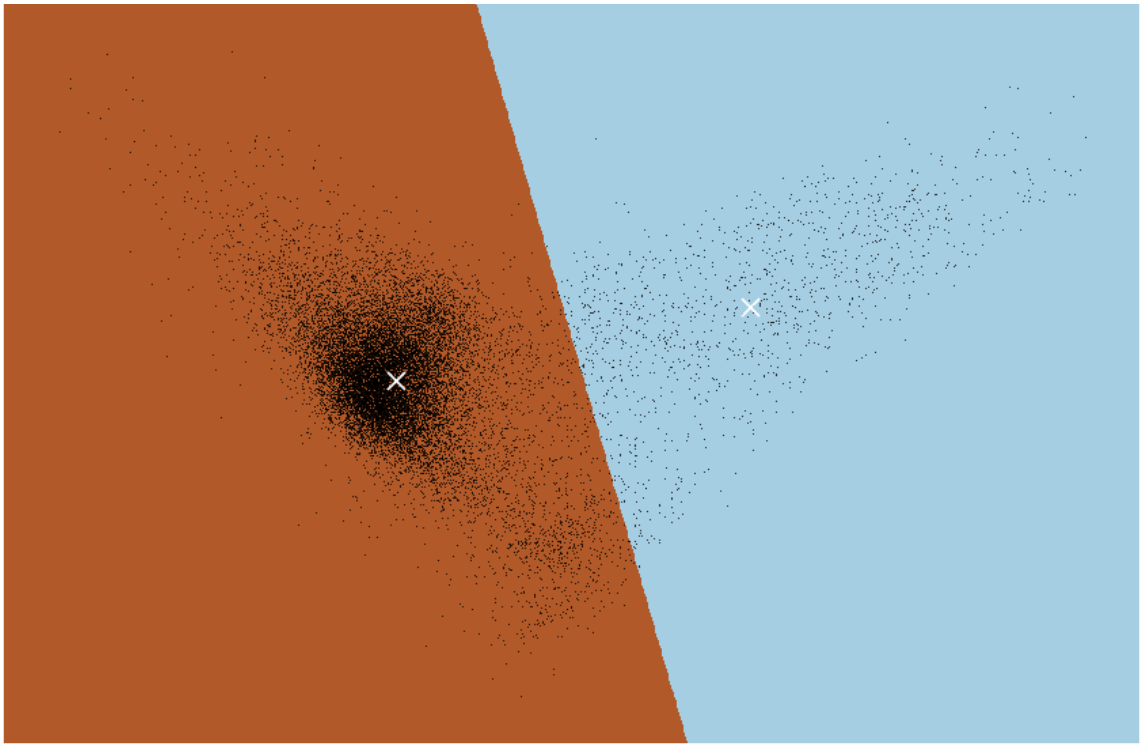
After observing the variance of each feature, we can conclude that 2 out of 8 features are responsible for 80% of the variance.

Selecting those 2 features for our prediction model will save computation as well as help us to visualise the data.



K-Means Clustering Analysis:

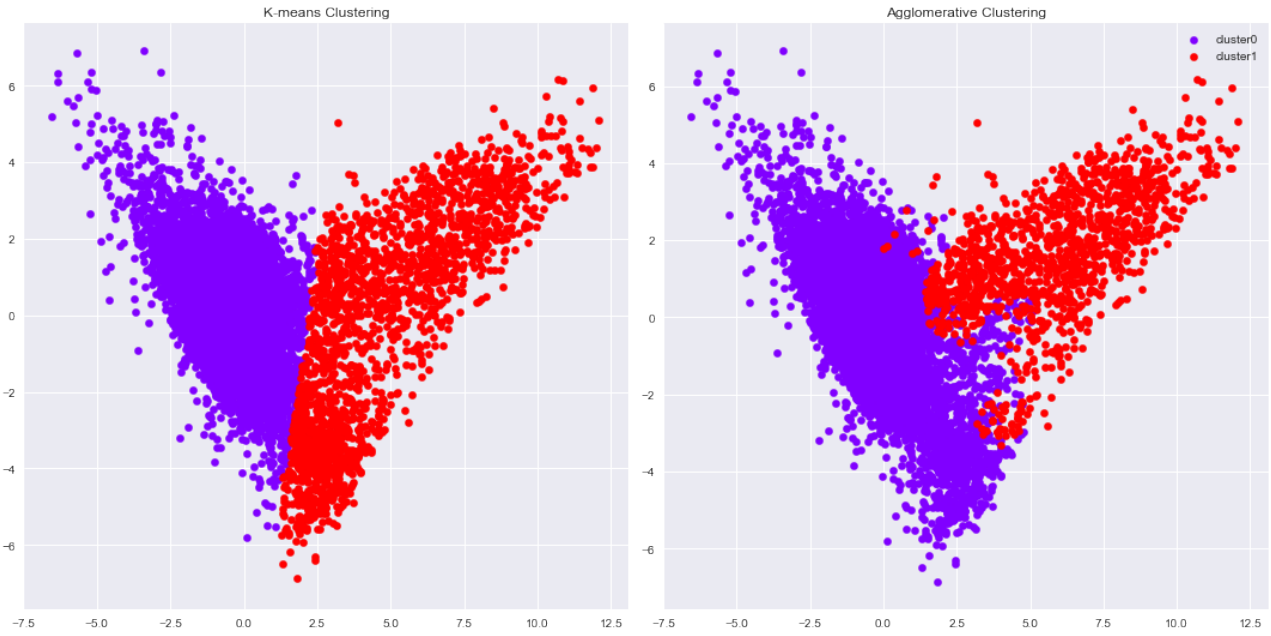
k-means clustering aims to partition n observations into k clusters (here, 2) in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.



KMeans vs. Agglomerative Clustering:

Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters. Some of its advantages over K-Means clustering are:

- It can produce an ordering of the objects, which may be informative for data display.
- Smaller clusters are generated, which may be helpful for discovery.



Part III. Classification - A Supervised Learning Task

Decision Tree:

Decision Tree Classifier is a simple and widely used classification technique. It applies a straightforward idea to solve the classification problem. Decision Tree Classifier poses a series of carefully crafted questions about the attributes of the test record. Each time time it receive an answer, a follow-up question is asked until a conclusion about the class label of the record is reached.

Random Forest Classifier:

Random forests or random decision forests[1][2] are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set'

Naive Bayes Classifier:

Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. The technique is easiest to understand when described using binary or categorical input values.

It is called *naive Bayes* or *idiot Bayes* because the calculation of the probabilities for each hypothesis are simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value P(d1, d2, d3|h), they are assumed to be conditionally independent given the target value

K nearest neighbour classifier:

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms.

The neighbours are taken from a set of objects for which the class (for *k*-NN classification) or the object property value (for *k*-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

Support Vector Machine classifier:

A SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

Part IV. Leveraging weak learners via Adaptive Boosting

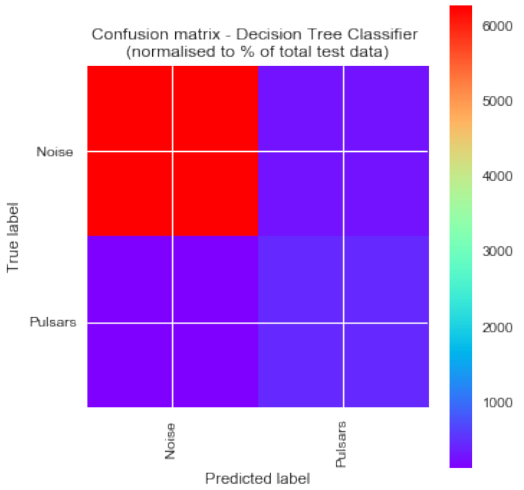
AdaBoost is a popular and effective leveraging procedure for improving the hypotheses generated by weak learning algorithms. AdaBoost and many other leveraging algorithms can be viewed as performing a constrained gradient descent over a potential function. At each iteration the distribution over the sample given to the weak learner is proportional to the direction of steepest descent.

The resulting algorithms have bounds that are incomparable to AdaBoost's. The analysis suggests that our algorithm is likely to perform better than AdaBoost on noisy data and with weak learners returning low confidence hypotheses. Modest experiments confirm that our algorithm can perform better than AdaBoost in these situations.

Observations & Results

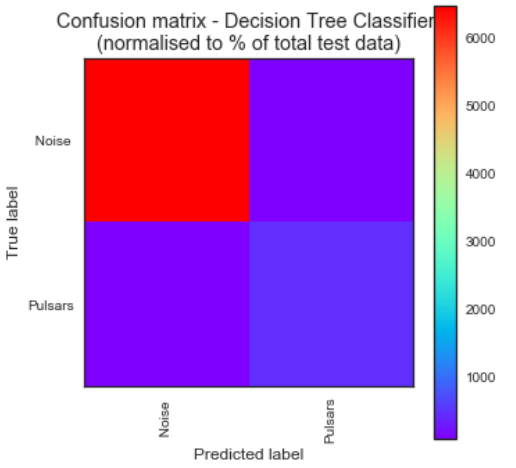
1. K-Means Clustering Performance:

	precision	recall	f1-score	support
0	0.98	0.95	0.97	6566
1	0.60	0.77	0.68	594
avg / total	0.95	0.94	0.94	7160



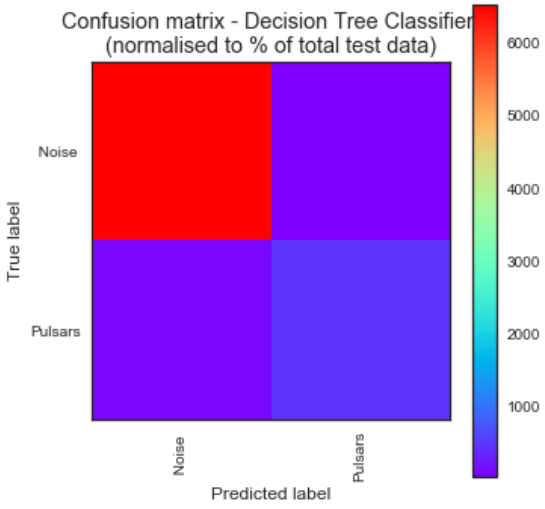
2. Decision Tree Classifier Performance:

	precision	recall	f1-score	support
0	0.98	0.99	0.98	6566
1	0.83	0.80	0.82	594
avg / total	0.97	0.97	0.97	7160



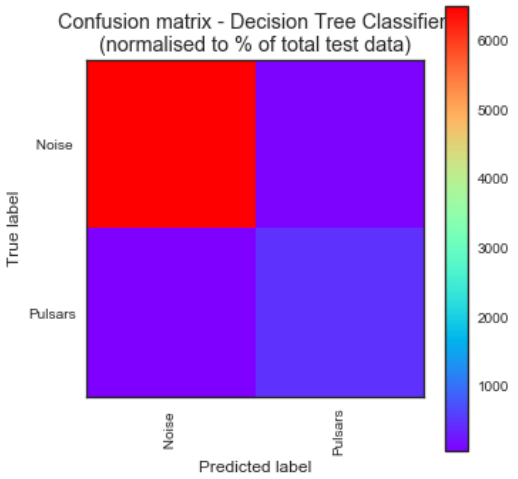
3. Random Forest Classifier Performance:

	precision	recall	f1-score	support
0	0.98	0.99	0.99	6566
1	0.93	0.82	0.87	594
avg / total	0.98	0.98	0.98	7160



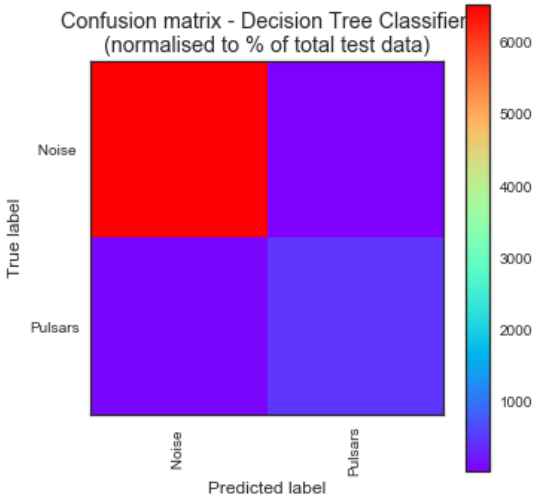
4. Naive Bayes Classifier Performance:

	precision	recall	f1-score	support
0	0.99	0.96	0.97	6566
1	0.65	0.86	0.74	594
avg / total	0.96	0.95	0.95	7160



5. KNN Classifier Performance:

	precision	recall	f1-score	support
0	0.98	0.99	0.99	6566
1	0.92	0.81	0.86	594
avg / total	0.98	0.98	0.98	7160



6. Support Vector Machine Classifier Performance:

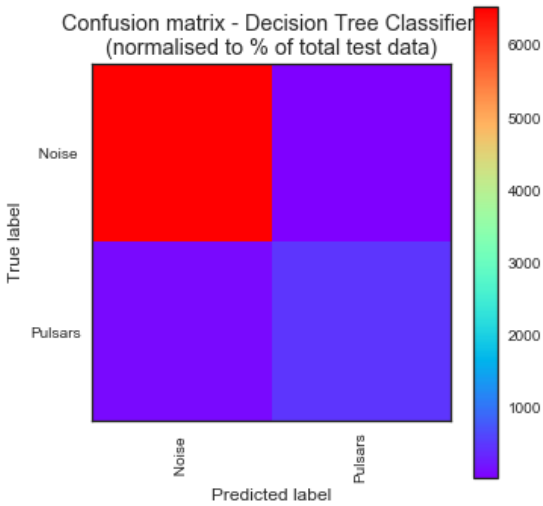
Best score: 0.980

Best parameters set:

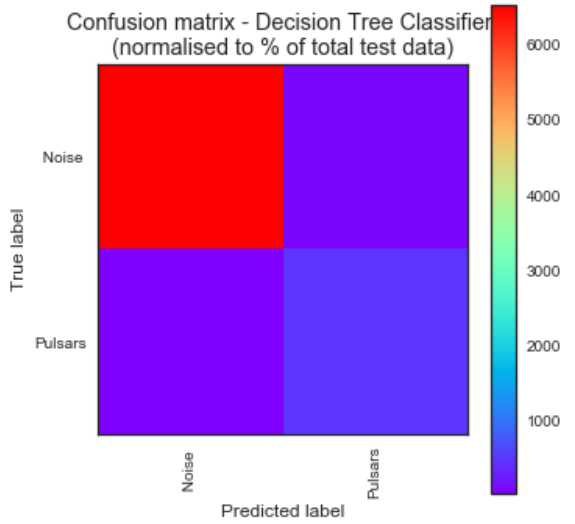
C: 10

kernel: 'rbf'

	precision	recall	f1-score	support
0	0.98	0.99	0.99	6566
1	0.93	0.80	0.86	594
avg / total	0.98	0.98	0.98	7160



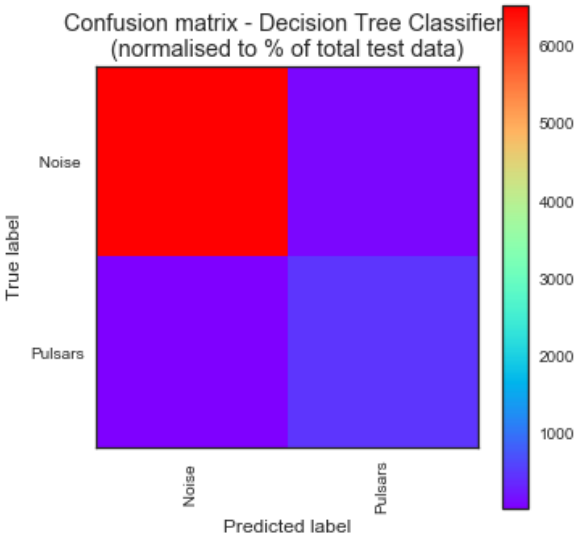
7. Adaptive Boosting of Decision Tree Classifier:



Decision tree train/test accuracies 0.976/0.977

	precision	recall	f1-score	support
0	0.99	0.98	0.99	6609
1	0.82	0.89	0.85	551
avg / total	0.98	0.98	0.98	7160

8. AdaBoost Classifier Performance:



AdaBoost train/test accuracies 0.979/0.980

	precision	recall	f1-score	support
0	0.99	0.98	0.99	6640
1	0.82	0.93	0.87	520
avg / total	0.98	0.98	0.98	7160

Conclusion

After analysing the performance of different supervised algorithms, the results have been obtained with a Random Forest Classifier.

Random Forest Classifier: It classifies 489/594 samples in Test Case correctly going ~82% Accuracy and only misclassifies 116/6648 R/F Noise Samples as legitimate sources. Thus, with probability of False Detection around 1.74%.

References

1] 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

[2] D. R. Lorimer and M. Kramer, 'Handbook of Pulsar Astronomy', Cambridge University Press, 2005.

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

[4] R. J. Lyon et al., '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, DOI: 10.1093/mnras/stw656

[5] R. P. Eatough et al., 'Selection of radio pulsar candidates using artificial neural networks', Monthly Notices of the Royal Astronomical Society, vol. 407, no. 4, pp. 2443-2450, 2010.

[6] S. D. Bates et al., 'The high time resolution universe pulsar survey vi. an artificial neural network and timing of 75 pulsars', Monthly Notices of the Royal Astronomical Society, vol. 427, no. 2, pp. 1052-1065, 2012.

[7] D. Thornton, 'The High Time Resolution Radio Sky', PhD thesis, University of Manchester, Jodrell Bank Centre for Astrophysics School of Physics and Astronomy, 2013.

[8] K. J. Lee et al., 'PEACE: pulsar evaluation algorithm for candidate extraction a software package for post-analysis processing of pulsar survey candidates', Monthly Notices of the Royal Astronomical Society, vol. 433, no. 1, pp. 688-694, 2013.

[9] V. Morello et al., 'SPINN: a straightforward machine learning solution to the pulsar candidate selection problem', Monthly Notices of the Royal Astronomical Society, vol. 443, no. 2, pp. 1651-1662, 2014.

[10] R. J. Lyon, 'PulsarFeatureLab', 2015, [Web Link].

[11] Sci-Kit Learn [Web Link]