**MACHINE LEARNING APPROACH TO DETECT PULSARS**

*Presented By*

**Bhavna Dora**

**Roll no- 12102V203004**

*Supervised By*

**Dr. Jagadish Kumar**

****

Post Graduate Department of Physics, Utkal University

Bhubaneswar

**CERTIFICATE**

This is to certify that the project entitled ***“Machine learning approach to detect pulsars”*** is submitted by Bhavna Dora, P.G. Department of Physics, Utkal University under my supervision and I consider it worthy of consideration for M.Sc. semester examination in the Department of Physics.

Date: Dr. Jagadish Kumar

**DECLARATION**

I certify that the work contained in this project is a review work and has been done by myself under the general supervision of my advisor Dr. Jagadish Kumar. The work has not been submitted to any other institute for any degree. I have followed the guidelines provided by the institute in writing the project.

Bhavna Dora

**ACKNOWLEDGEMENT**

Foremost, I would like to express my sincere gratitude to my thesis advisor, Dr. Jagadish Kumar, for granting me the opportunity to work under his guidance. He has been a true inspiration for me. He was there when I needed help and was always available with new ideas and suggestions, whenever I had any difficulty with my study. Without his help this thesis would have never come to life.

I would like to thank my parents for their continuous support when I needed it the most. They were always there for me right to the end. Lastly, I would also like to thank my friends who had helped and motivated me immensely which would ensure this project work being propelled to the finish in time. Thank you!

Bhavna Dora

**CONTENTS**

1. INTRODUCTION
   1. Introduction to pulsars5
   2. What are Pulsars?5
   3. Emission of Radio waves- Light house model…………………………………………………………………7
   4. Why to look for Pulsars?9
2. TRANSITION TO ML
   1. Pulsar search.11
   2. Candidate selection………………………………………………………………………………………………………11
   3. Transition to Machine learning…………………………………………………………………………………..12
      1. Decision Tree…………………………………………………………………………………………….12
      2. Logistic Regression…………………………………………………………………………………….13
      3. Random Forest14
3. PULSAR DETECTION- ML CLASSIFICATION
   1. Definition of the problem…………………………………………………………………………………………….16
   2. Data extraction & visualization ……………………………………………………………………………………17
   3. Pre-processing of the data…………………………………………………………………………………………..23
   4. Choice of ML models and metrics………………………………………………………………………………..23
   5. Training and evaluating the models…………………………………………………………………………….24
   6. Conclusion on the best model……………………………………………………………………………………..32

**Future applications**………………………………………………………………………………………………………………….34

**References**……………………………………………………………………………………………………………………………….35

**CHAPTER I**

* 1. **INTRODUCTION TO PULSARS**

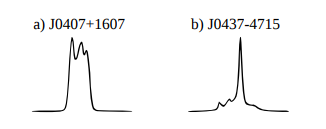
In 1931, a radio engineer Karl Jansky built an antenna to receive radio waves while working at Bell laboratories. Jansky noticed that there was a particular radio signal which was periodic in nature. Its maximum intensity was observed in every 23 hours and 56 minutes. After some time he was able to determine its source which was a region in the Sagittarius constellation at the centre of our galaxy. So, Jansky stumbled upon an astronomical radio source and made the first of many discoveries in the research field which was later termed as Radio Astronomy.

In 1967, Jocelyn Bell and Antony Hewish detected pulsed radio emission while studying the scintillation of compact radio sources using their telescope- Interplanetary Scintillation Array. The signal was periodic and repeated every 1.337 seconds. The detected signal was referred to as ‘Little Green Men One’ (LGM-1), because it was considered to be of extra terrestrial origin. It was thought of as a communication attempt by some extra terrestrial intelligent beings with us. However, it was later identified as a radio wave emitting star called Pulsar. It led to the development of a new sub field of study called Pulsar Astronomy.

* 1. **WHAT ARE PULSARS?**

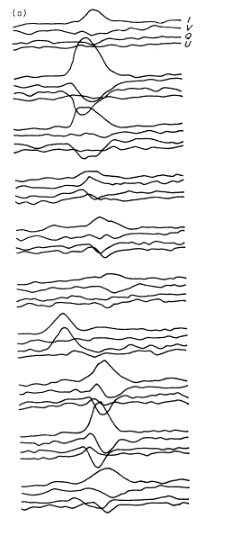
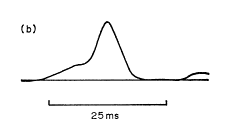
Pulsars are rapidly rotating neutron stars emitting periodic radio waves detectable here on Earth. Pulsars are extremely dense stellar remnants and posses typical mass of 1.2 times solar mass condensed into a sphere of approximately 20 km in diameter [1]. The fastest known rotational period of pulsar is 1.396 milliseconds, which corresponds to ∼ 716 rotations per second [2] while the slowest observed rotational period is 8.51 seconds. Pulsars also posses immense gravitational and magnetic fields. The gravity at the surface of pulsars is 108 – 1014 Gauss [3] in contrast to Earth’s magnetic field strength which is 0.25-0.65 Gauss.

Each pulsar produces a unique pattern of pulse emission known as its pulse profile. A pulse profile is similar to a fingerprint. We can identify the pulsars based on their pulse profiles alone.



*Figure 1.1: pulse profiles of two pulsars [4]*

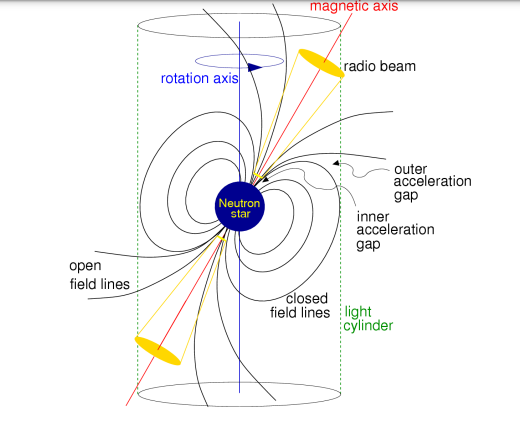
However, while the pulsar rotational periods are extremely consistent, their profiles can deviate from one-period to the next. This makes them hard to detect. However these pulse profiles do become stable, when averaged over many thousands of rotations giving rise to integrated pulse profile [4].

*Figure 1.2: Single and integrated pulse profile of PSR 0329+54 [5]*

* 1. **EMISSION OF RADIO WAVES- LIGHT HOUSE MODEL**

Radio waves emitted by pulsars originate from their magnetosphere. Magnetosphere is defined as the space surrounding a pulsar in which charged particles are influenced by a co-rotating magnetic field (the magnetic field that rotates with the pulsar at same velocity) which has both open and closed field lines.

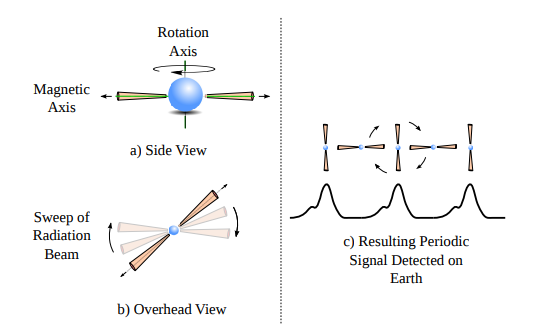


*Figure 1.3: Light house model of a radio Pulsar [1]*

The precise process that occurs within the magnetosphere that leads to the emission of radio waves is not entirely understood. It is hypothesized that the strong magnetic field causes a charge separation between the pulsar surface and magnetosphere. This leads to a voltage gap which causes the extraction of charged particles from the surface. These particles are accelerated along the co-rotating magnetic field lines which endows them with high energy. So, they emit radiations in the form of high energy photons. These photons being subjected to the strong magnetic field of pulsars undergo pair production- electron and positron which are also accelerated along the co-rotating magnetic field lines. So, they emit radiations like radio waves, gamma waves etc. [4]

**LIGHT HOUSE MODEL-**

The pulsar’s magnetic axis is inclined with respect to its rotational axis. So, each time a pulsar rotates, the radiation beam produced near the magnetic poles is swept at an angle across the sky. If the beam crosses the line of sight of an observer on Earth, the radio waves emitted are detectable. This pattern repeats periodically with each rotation of pulsar. This is called as light house model of emission. [4]



*Figure 1.4: A pulsar with magnetic axis orthogonal to rotational axis and its radio signal detected on Earth [6]*

* 1. **WHY TO LOOK FOR PULSARS?**

Pulsars are of significant scientific interest as they provide a means through which we can study stellar evolution, nature of gravitation, composition of the Inter Stellar Medium etc. Pulsars are valuable laboratories to investigate fundamental physics.

1. PROBES OF THE INTER STELLAR MEDIUM (ISM)-

As radio waves emitted by pulsars travel towards Earth, it passes through the ISM. The ISM affects the radio signal in many ways. The radio waves interact with the free electrons present in the ISM and are dispersed. The low frequency parts of the signal reach later than the high frequency parts. This is the dispersive effect and causes the smearing of pulsar signals in time. This allows us to study the composition if the inter stellar medium closely. Thus pulsars are regarded as probes of the ISM.

1. TEST OF GRAVITY-

Hulse-Taylor binary system consists of a closely orbiting pulsar and neutron star. Observation of this system revealed that their orbits were decaying over time. The rate of decay was the same as predicted by Einstein’s general theory of relativity, which accounts for energy loss via the production of gravitational wave. [6]

1. GRAVITATIONAL WAVE DETECTION-

Millisecond Pulsars (MSPs) which have rotation period in milliseconds can be used to detect gravitational waves [7]. By timing the arrival of pulses produced by numerous MSPs, the presence of gravitational waves can be detected as a disturbance in the regularity of pulse arrival times on Earth. This is because a passing gravitational wave distorts the space time metric around the Earth, thus imprinting a characteristic signature in the pulse arrival times.

1. INDEPENDENT TIME STANDARD-

Measuring the time of arrival of pulses gives rise to an extremely effective time keeping system as accurate as atomic clocks [1]. It is useful for space craft navigation and time keeping here on Earth.

**CHAPTER II**

* 1. **PULSAR SEARCH**

Pulsars are searched in two modes- Targeted and Blind searches. Targeted searches are focused on those locations where pulsars are more likely to be found i.e. dense stellar regions like Galactic clusters. Blind searches observe multiple regions of the sky. They have discovered most of the pulsars which are known till date. However, detecting pulsars during blind searches is very difficult because their emitted radio signals are extremely weak on reaching the Earth due to scintillation and scattering.

Most of the signals detected are the man made Radio Frequency Interference (RFI) and noise with a very small fraction of true legitimate pulsars. So, we need a very large radio telescope which is sensitive to weak signals and a signal search pipeline to isolate the true pulsar signals by filtering out the RFI and noise.

* 1. **CANDIDATE SELECTION**

Pulsar candidates are radio signals that exhibit pulsar like characteristics making them worthy of further analysis. The candidates are describable in terms of characteristics like Signal to Noise ratio, Dispersion measure, Signal period, Pulse width etc. [1] In earlier times, candidate selection were carried out manually by visually inspecting every candidate. This was very time consuming and cognitively taxing. A lot of telescope hours went in vain. Moreover, during the past 50 years the number of candidates has increased steadily. So, numerous automated methods were developed using computers to generate plots to visually inspect the candidates. But these techniques produced results with low level filtering by using two characteristics- S/N ratio and pulse profile [8].

But the candidate numbers kept on increasing due to the advanced technical capabilities of telescopes and associated computational infrastructures. So, it became unfeasible for human inspection of filtered candidates too (which were also of ‘considerable size’).

* 1. **TRANSITION TO MACHINE LEARNING**

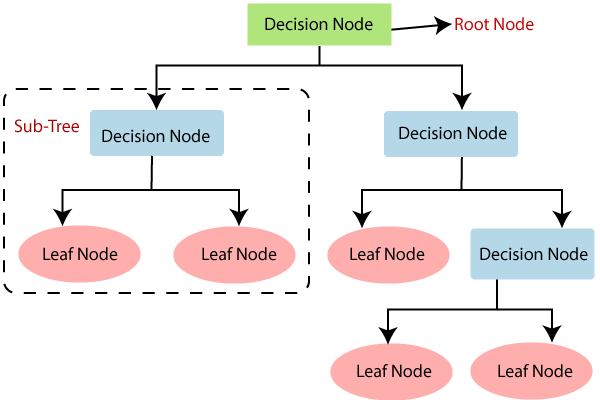
The candidate selection problem is going to worsen in the coming years because the radio telescopes under development will produce large volumes of observational data. It will become impossible to store them for a later on off-line reprocessing because of the financial cost of required storage media. [6] So, the candidate selection techniques must change. It must transition from being an off-line process to a real-time process operating on the observational data as it arrives. This calls for using machine learning algorithms to detect pulsar candidates from a large volume of data with utmost accuracy.

Detecting the radio signal received as from a legitimate pulsar or from RFI and noise is dealt in machine learning as a binary classification problem which gives output as binary value- 0 indicates not a pulsar and 1 indicates pulsar. We use the dataset HTRU2 collected during the High Time Resolution Universe Survey with 17898 pulsar candidates out of which only 1639 are legitimate pulsars [9]. We develop three machine learning classification models- Decision Tree, Logistic Regression and Random Forest to implement on the dataset and compare their accuracies in detecting pulsars.

* + 1. **DECISION TREE**

Decision trees are supervised learning method used for classification and regression [10]. The model learns decision rules inferred from the data features and predicts the value of target variable. It is a piece wise constant approximation. Decision tree classifier is a machine learning algorithm used for classification of dataset.

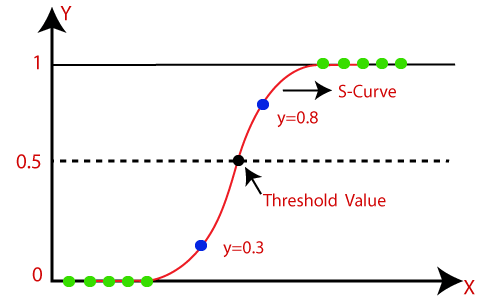
Decision tree is like a flow chart. The attributes/ features are represented be decision nodes. Each branch is a decision rule. Output is represented by leaf node. The best attribute is selected using the Attribute Selection Measure (ASM). This is made the top most decision node called root node. Then the dataset is broken to smaller subsets based on the attribute value. The process is repeated recursively to build the tree. It stops till there are no attributes left and no instances left.



*Figure 2.1: Built of decision tree [11]*

* + 1. **LOGISTIC REGRESSION**

Logistic regression is a supervised learning method. As the name suggests it is used for regression problems i.e. the output predicted assumes continuous values. But it can be used for classification problems. Logistic regression models the data using a sigmoid function (S shaped) given as

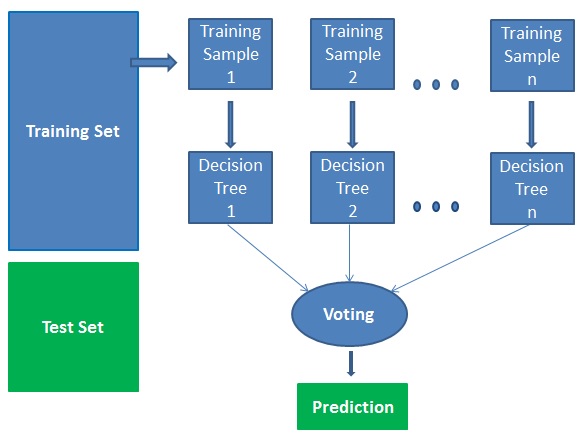
****

*Figure 2.2: Sigmoid curve [11]*

The function returns a value between 0 and 1, which is nothing but the probability. We have to choose a threshold value which is important for the regression model to behave as a classification model. Let us choose the threshold as 0.5. Then if the value returned by the sigmoid function is 0.8 then it is classified as one class (say class A) and if the value returned is 0.3 then it is classified into another class (say class B). In this way logistic regression becomes a classification model.

* + 1. **RANDOM FOREST**

Random forest is a supervised learning method. It is used for both classification and regression. It is based on the concept of ensemble learning- it contains many decision trees on various subsets of the given dataset and takes their average to predict the outcome with high accuracy [10].

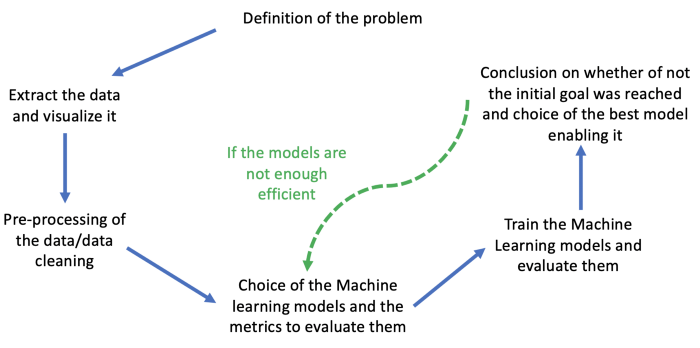


*Figure 2.3: Working of Random Forest Classifier [11]*

The training set is broken down into many smaller subsets (let n subsets). It is fed to n no. of decision trees. These decision trees are trained. When the test set is given to these decision trees, each of them predicts their output. The output with the majority of votes is given as the final output.

**CHAPTER III**

The pulsar detection problem using the machine learning classification algorithms is divided into a six step process as illustrated below.



**3.1. DEFINITION OF THE PROBLEM**

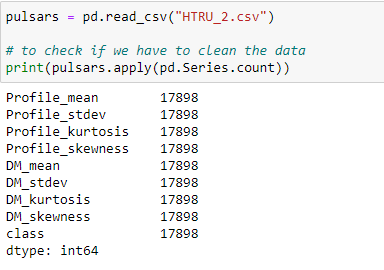
The goal of the project is to detect pulsars from the recorded dataset with utmost accuracy. We can see clearly that the result of our models will be binary: 1 if the star is a pulsar star and 0 otherwise. Thus, we implement classification models like Decision Tree, Logistic Regression and Random Forest to deal with this problem. We use the dataset HTRU2 which describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey. It can be accessed here [12].

# 3.2. EXTRACT THE DATA AND VISUALIZE IT

We use various libraries of python- Numpy for all the numerical functions, Pandas for data cleaning and statistical analysis, Seaborn and Matplotlib for data visualization and plots, Sklearn for classification models etc. we also use the metrics module which contains many functions like confusion matrix, classification report, accuracy, f1 score, roc curve, auc etc.



HTRU\_2 dataset was downloaded in the form of a CSV (comma separated values) file. Using the read\_csv function we extract the data. We check if we have to clean the data by counting the no. of data points present in the data set.



We see that there are 17898 data points. There are a total of 8 features and one target class and each of them have values corresponding to the 17898 data points. So, there are no missing values. Thus, we don’t have to clean the data set. The 8 features that describe the pulsars are- Mean, Standard deviation, Excess Kurtosis, Skewness of the integrated profile, Mean, Standard deviation, Excess Kurtosis, Skewness of the DM-SNR curve. [8]

Integrated profile- Pulsars emit radio waves and the pattern of emission varies slightly with each rotation of the Pulsar even though the time period remains same. So, these profiles are averaged thousands of times to give a stable profile called integrated profile.

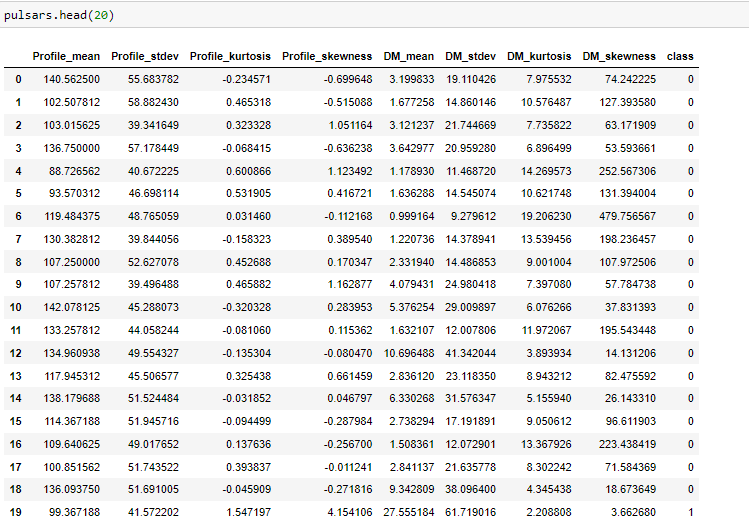
DM-SNR curve- As pulsar signals travel through the interstellar medium (ISM) to reach Earth, they interact with free electrons and are dispersed. The low frequency parts of the signal reach later than the high frequency parts. This is the dispersive effect and causes smearing of pulsar signals in time. So, the strength of the pulsar signals reduces. Signal to noise ratio (SNR) is a measure of strength of pulsar signals. S/N should be very high for the signal to be distinguishable and separable from noise. Higher S/N indicates presence of legitimate pulsars.

The smearing of pulsar signals in time is proportional to a quantity called Dispersion Measure (DM). It is the integrated column density of free electrons between the observer and pulsar. The time delay between the low and high frequency components is [1]-

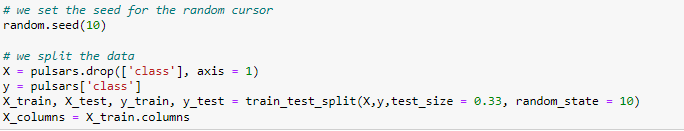
So, the time delay is proportional to DM. Using accurate value of DM we can remove the time delay, then SNR will be high and closer to the pulse profile. DM-SNR curve is a plot between DM and SNR.

Among the features, Mean is simply the average. Standard deviation indicates how much spread out the profile is. Kurtosis is a measure of size of the tail of the distribution. Normal distribution has a kurtosis of 3. So, Excess Kurtosis = Kurtosis – 3. Positive excess kurtosis indicates the distribution has fat tails. Negative excess kurtosis indicates the distribution has light tails as compared to the normal distribution. Skewness tells about the asymmetry of the distribution. If the tail of the distribution extends towards right or left, then the distribution is called right skewed or left skewed respectively.

Let us have a look at our data set and print the first 20 data points.

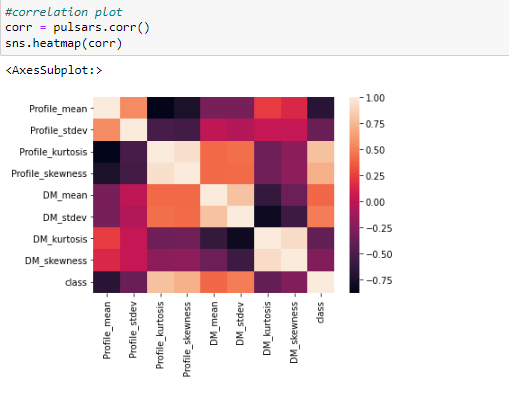


The next step is to **split the data** into two sets- train set which comprises 67% of data that will be used to train the machine learning algorithms and test set which comprises 33% of data that will be used to test the performance of each of them. We use the train\_test\_split function to do this. This function splits the data and assigns them randomly to the train and test set. So, in order to reproduce the same results later on, we used random.seed function and give it any number to initialize it. The seed function saves the random state and thus reproduces the same result on multiple executions of the code.



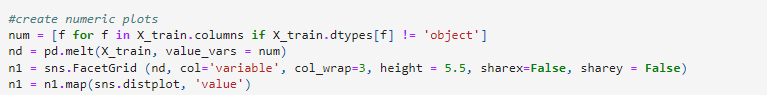
Now, let’s generate the correlation plot. Correlation plot is used to analyze the correlation between different pairs of features. Correlation coefficient ranges from a scale of -1 to +1. Values near to +1 indicate strong positive relation between the pair of features and values near to -1 indicate strong negative relation between the features. Values near to 0 indicate no relation between the features. The correlation coefficient is a measure of strength of linear relationship between the pairs of features.

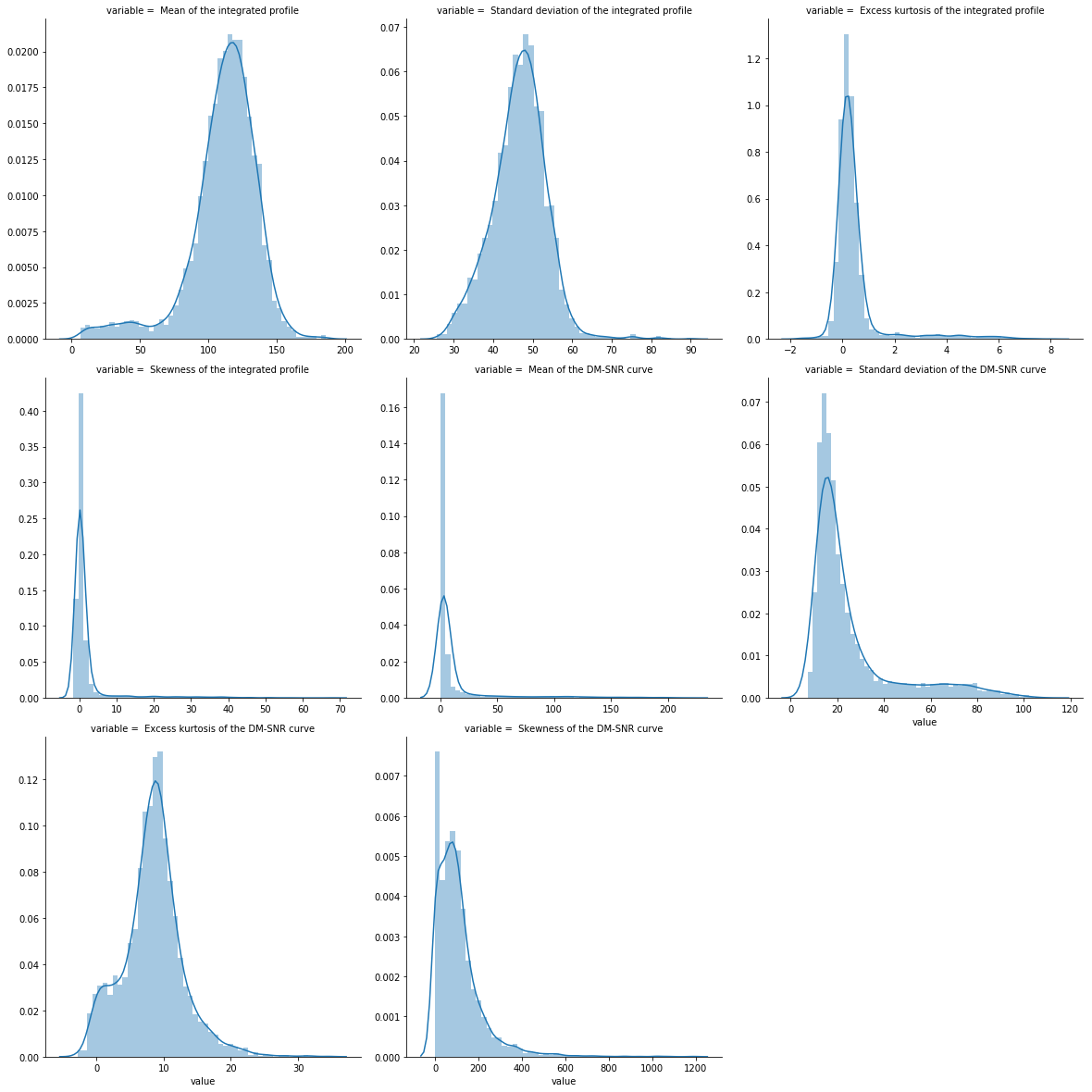
We use the corr( ) function to compute the correlation coefficients. Then we plot it using seaborn library’s heatmap function.



We observe the link of all the features with the target class. We notice that Excess Kurtosis of the integrated profile is the most correlated feature with the target class. This indicates that it is an important feature that describes pulsars.

Let’s now see the distribution of the 8 features. We create numeric plots i.e. histogram. Histogram is the distribution of numeric feature/variable where the values of the variable are split into bins and each bin is represented by a bar. We use distplot function to fit a curved line to the histogram.





We see that most of the features have a Gaussian distribution with a little amount of skewness. We also observe that the scales of the features vary widely. So, we need to standardize them.

**3.3. PRE-PROCESSING OF THE DATA**

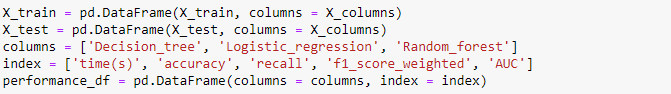
We normalize our data using the standard scaler function. It standardizes the data by removing the mean and scaling to unit variance.

This is important because if the distribution of features is not centered at 0 (i.e. mean=0) and have unit variance, then some features with high variance may dominate other features and then the model won’t be able to learn from all features equally and won’t be able to predict right results.

snip 5.PNG

**3.4. CHOICE OF THE MACHINE LEARNING MODELS AND THE METRICS TO EVALUATE THEM**

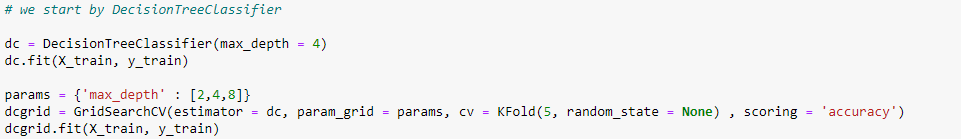
We will use 3 machine learning models- Decision Tree, Logistic Regression and Random Forest for this classification problem. We will also evaluate the performance of these using metrics like time, accuracy, recall, f1 score, AUC etc. So, let’s create a data frame with these metrics for each model.



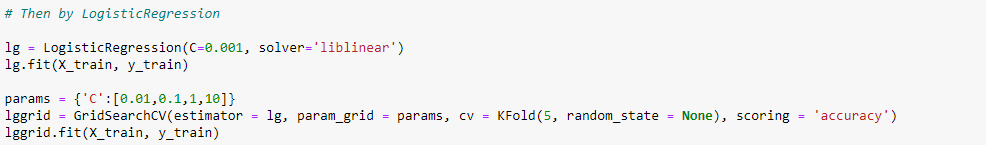
**3.5.** **TRAIN THE MODELS AND EVALUATE THEM**

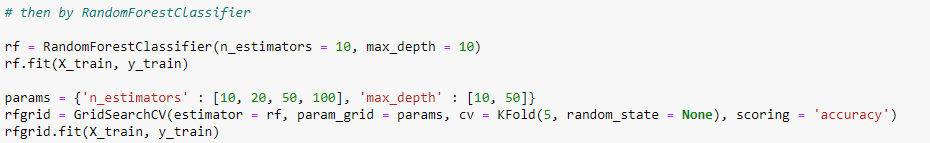
**Training of models-** We start by creating object of Decision Tree Classifier. Then we fit the model to the training set. This is called as training of model. Now we define the hyper parameters. We use a hyper parameter- max\_depth to tune our model. Max\_depth indicates how deep the tree can be i.e. how many splits the tree can have. Generally a max\_depth of 4 is used for better results.

The training model is split into K=5 smaller sets. The model is trained using K-1 sets and the remaining set is used to validate the model and see its accuracy. This is called K-fold cross validation. Then we use GridSearchCV which runs through all the parameters fed to the parameter grid and produces the best combination of parameters based on the scoring metric of our choice- accuracy, f1, recall etc. The model is again trained using fit function.

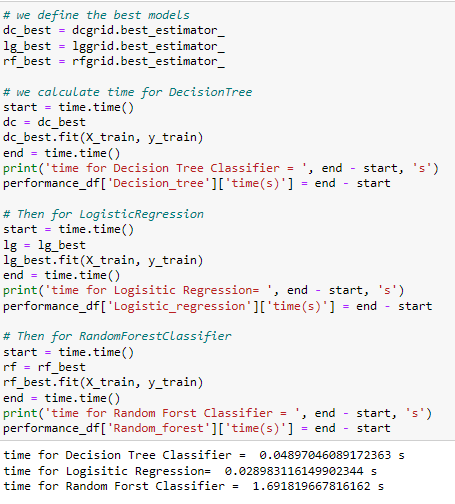


We apply the same methodology for Logistic Regression and Random Forest Classifier with different hyper parameters. The hyper parameter in case of Logistic Regression is C and the solver used is liblinear. The parameter C signifies the strength of regularization. Smaller the C, stronger is the regularization. It takes a positive float value. The hyper parameters in case of Random Forest is n\_estimators – which are the number of Decision Trees built before taking the average / majority votes to predict the final output, and max\_depth which represents the number of splits in a decision tree.

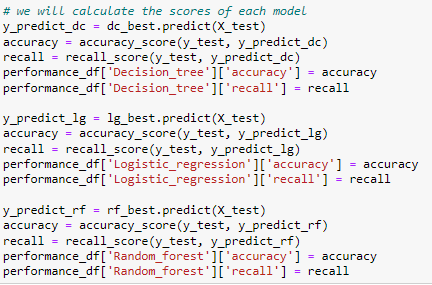




**Time taken to train-** Now let us calculate the time taken to train each model. After using the GridSearchCV we got the best combination of hyper parameters. Then we define the best models using the best\_estimator\_attribute (it returns the estimator with highest score). Now we use the time function. We measure the start time, then we fit the best model to the training set. Then we measure the end time. We subtract these two measured times to get the time taken to train the model.



**Evaluation on the basis of metrics-** Now we evaluate the best models based on other metrics like Recall, accuracy. We predict the values of target class for each data point in the test set using the predict function. Then we calculate the accuracy score and recall score.



**Generating confusion matrix and classification report-** Now let us generate the confusion matrix for Decision Tree Classifier. We plot it using seaborn library’s heatmap function. Confusion matrix is a 2×2 matrix. It has 4 segments-

TP (True Positive) - they are actually positive and are also predicted as positive.

TN (True Negative) – they are actually negative and are also predicted as negative.

FP (False Positive) – they are actually negative but are falsely predicted as positive- type I error.

FN (False Negative) – they are actually positive but are falsely predicted as negative- type II error

From the confusion matrix we can calculate-

Accuracy- it is the ratio of correct predictions by the model to the total no. of predictions =

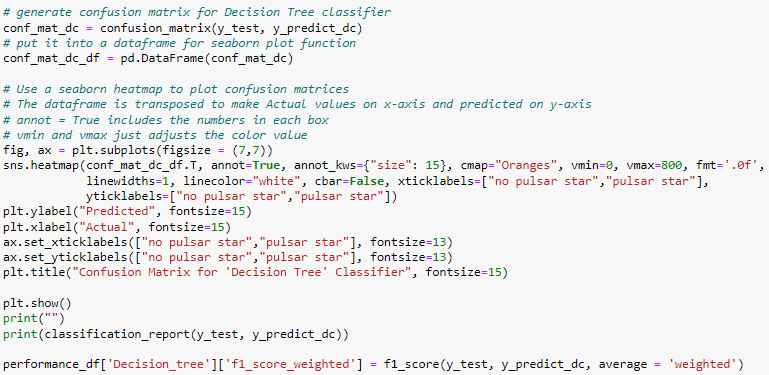
Recall/ True Positive Rate- ratio of true positives to the actual no. of positives =

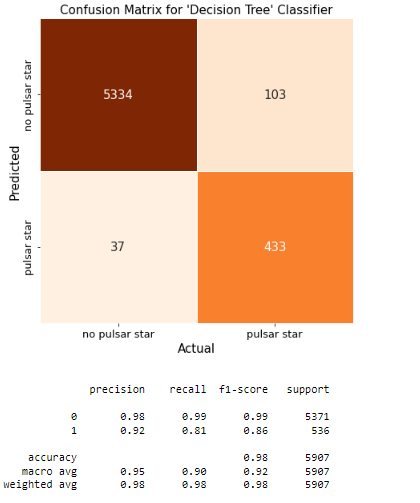
Precision- ratio of true positives to the predicted no. of positives =

False Positive Rate- ratio of false positives to the actual no. of negatives =

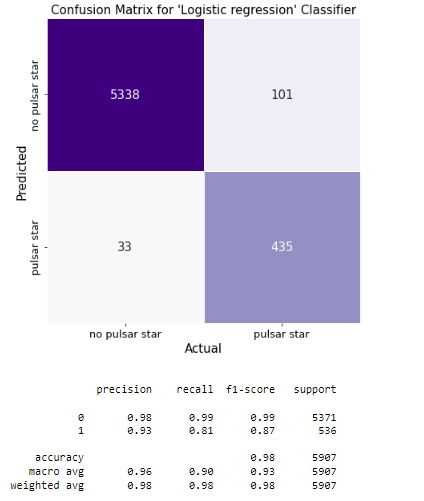
We also generate the classification report. Classification report is a report of performance of the model. It displays precision, recall, f1 score, support. F1 score is the weighted harmonic mean of precision and recall. The closer the value of f1 score is to 1, better is the performance of the model.

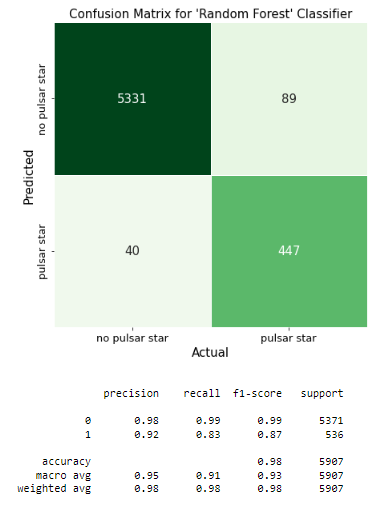
Support is the no. of true occurrences in the dataset = TP + FN





Similarly we generate the confusion matrix and classification report for Logistic Regression and Random Forest classifier.

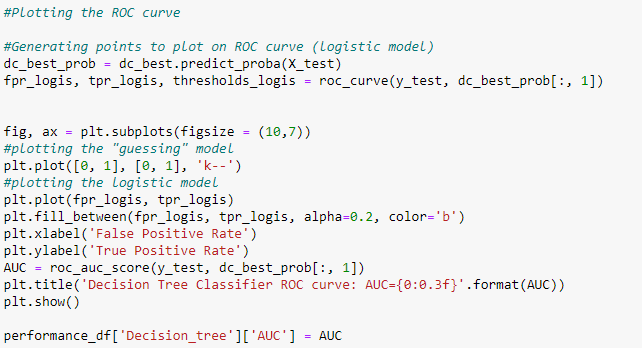


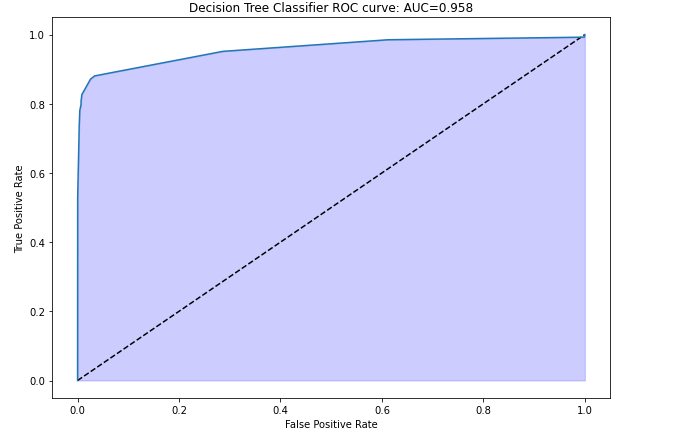


**Evaluation on the basis of most important metric-** Finally we evaluate the best models on an important metric- ROC curve and AUC value. ROC (Receiver Operating Characteristics) is a probability curve. It is a plot between TPR and FPR. AUC is the area under the curve. Higher the value of AUC, better is the model at predicting 0’s as 0 and 1’s as 1. It is the degree of separability. Roc\_auc\_score returns the AUC value (between 0.0 and 1.0).

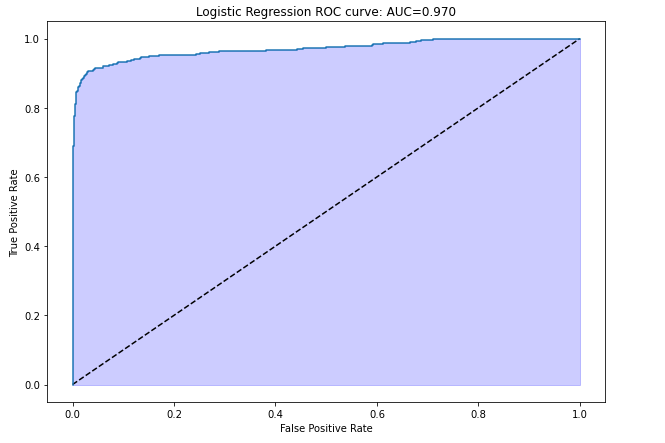
As ROC is a probability curve, so we generate the points to plot using the predict\_proba\_function. For all the data points in the X test set, we predict the probability of that data point to be a pulsar. So it gives a value between 0.0 and 1.0. We pass these values and the Y test set to the roc\_curve function.

**ROC curve for Decision Tree Classifier-**

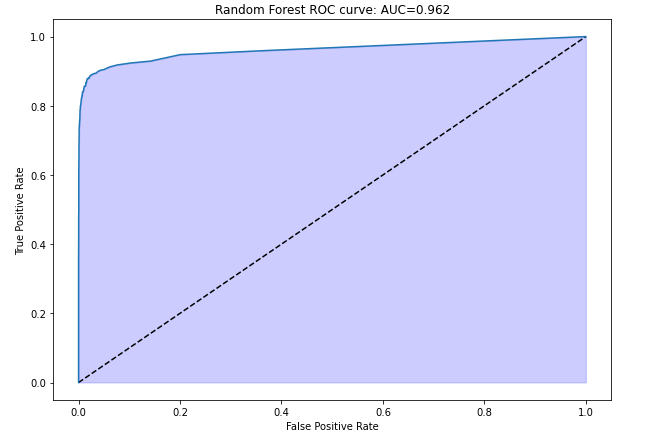




**ROC curve for Logistic Regression-**

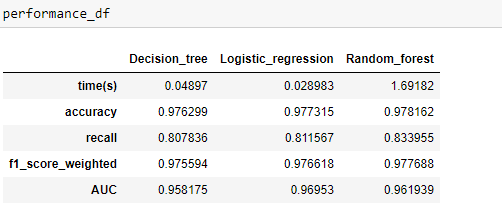


**ROC curve for Random Forest Classifier-**



**3.6. CONCLUSION ON THE BEST MODEL**

We print the performance table for all 3 models which consists of the metrics- Time, Accuracy, Recall, F1 score weighted, AUC.



We see that the time taken to train the models is the lowest for Logistic Regression. So, it is the fastest model which is a very important factor in case of pulsar detection. Accuracy, Recall, F1 score is almost same for all 3 models. The most important metric is AUC and it is almost same for all 3 models but slightly higher in case of Logistic Regression model. So, Logistic Regression is the most efficient model on the basis of HTRU2 data set to detect pulsars with an accuracy of 97.7%.

Complete code with output can be accessed at [Github](https://github.com/BhavnaDora/ML-approach-to-detect-Pulsar).

**FUTURE APPLICATIONS**

Square Kilometer Array (SKA) - An international effort to build the world’s largest radio telescope with over a square kilometer of collecting area is called the square kilometer array. It is going to be constructed in Australia and South Africa. It will be comprised of around 3,000 individual 15 meter antennas, and hundreds of thousands of dipoles. The core aim of SKA is to find pulsars by conducting a cosmic census of the pulsar population, searching for pulsars in the Galactic Centre, and searching for pulsars in Galactic clusters. This is anticipated to proceed in two phases- SKA-1 and SKA-2. SKA-1 will detect between 7000 to 9000 normal pulsars, and between 900 to 1400 MSPs. The completed SKA-2 instrument will be capable of observing all detectable pulsars in our galaxy of which there are estimated to be around 160,000 normal pulsars and 30,000 MSPs

SKA will produce data at a rate of 0.47 - 1.6 TB/s [6]. This data simply cannot be stored because of the huge cost of building storage media of this capacity. Instead it must be processed to extract data in real-time. So, transition from human annotation to machine learning is the need of the hour. Very fast classification algorithms like GH-VFDT (Gaussian Hellinger Very Fast Decision Tress) and artificial neural networks like SPINN are the highly efficient algorithms for a transition from off-line data processing to real-time data processing. A typical SKA observation will generate 6 million candidates per observation, with 500 seconds allocated to process this data. This implies that during SKA pulsar searches, candidates must be classified at a rate of at least 12,000 per second to achieve real-time operation. GH-VFDT classifiers are tested to be capable of classifying approximately 100,000 candidates per second [6]. Thus utilizing such type of machine learning models in data collected by SKA will add many more legitimate pulsars to our existing bank of pulsars.

**REFERENCES**

[1] Lorimer D.R. and Kramer M. Handbook of pulsar astronomy. Cambridge University Press, 2005

[2] Hessels J.W.T. “A Radio Pulsar Spinning at 716 Hz.” Science, 311(5769):1901– 1904, 2006

[3] Kaspi V.M. “Grand unification of neutron stars.” *Proceedings of the National Academy of Sciences*, 107(16):7147–7152, 2010.

[4] Lorimer D.R. “Binary and Millisecond Pulsars.” *Living Reviews in Relativity*, 11(8):1–90, 2008.

[5] Lyne A.G., Smith F.G., & Graham D.A. “Characteristics of the radio pulses from the pulsars.” *Monthly Notices of the Royal Astronomical Society*, Vol. 153, p. 337 (1971)

[6] Lyon R.J. “Why are Pulsars hard to find” The University of Manchester, UK, 2016

[7] Janssen G., Hobbs G., McLaughlin M., et.al. “Gravitational Wave Astronomy with the SKA.” *Advancing Astrophysics with the Square Kilometer Array, Proceedings of Science*, 037, 2014.

[8] 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*, Vol. 459, Issue 1, June 2016

[9] R. J. Lyon, HTRU2, DOI:10.6084/m9.figshare.3080389.v1.

[10] Scikit-learn 1.0.2- PyPI

[11] Javatpoint - <https://www.javatpoint.com>.

[12] <https://archive.ics.uci.edu/ml/datasets/HTRU2>