# Classification of High Time Resolution Universe Survey 2 Data by Machine Learning Technique

**Pratibha Verma[1], Sanat Kumar Sahu[2], A.K. Shrivas[3]**
[1]Ph.D. Scholar, Dr. C.V. Raman University, Bilaspur (C.G.) India.
[2]Dept. of Computer Science, Govt. Kaktiya P.G. College, Jagdalpur (C.G.) India.
[3]Dept. of IT, Dr. C.V. Raman University, Bilaspur (C.G.) India.

**ABSTRACT**
*Machine learning techniques are useful to finding or discovering the new pulsar. We study the candidate filters used to moderate theseproblems for the period of past few years. Pulsars are types of star of huge scientific interest. We have used two classification techniques like Conditional Inference Tree (CTREE) and Classification and Regression Tree (CART) and their ensemble model (CTREE+CART) for classifying the HTRU2 dataset. The ensemble model (CTREE+CART) give the better performance compared to the Individual models of each classifiers. The ensemble model (CTREE+CART) is useful of the candidates must be classified in to pulsar and non-pulsar classes to aid discovery.*

## I INTRODUCTION

Magnetic high-speed neutron stars are known as pulsars. Whose linearly transmitted polarized electromagnetic radiation extends along its magnetic poles? While Pulsar, maintains the variance. It radiates from time to time the journeys towards the observer's vision, such as the rotating beacon, which is the result of a periodic train of narrowband radiation pulses, which was detected by a radio telescope[1], [2]. Pulsar studies involve accurate measurements of the arrival time of the pulse, followed by an appropriate modeling of the observed arrival times to study and understand the various phenomena that may influence arrival times. Machine learning techniques are useful to finding or discovering the new pulsar. Here we describe the classification techniques to decide pulsar and non pulsar candidates. Classification techniques CTREE and CART are tree based classifiers and their ensemble model is CTREE, CART. These are useful to predict pulsar candidate selection.

## II RELATED WORK

Prior to the discovery of pulsar, many researchers have worked in the past. Some of its precise introductions are as follows.

Each candidate must be inspected by an automatic method like machine learning techniques and a human expert to determine its authenticity[3]. The process for deciding which candidates are worth investigating is known as 'selection of candidates' and called as "pulsar candidates"[4]. The authors [4] have presented a new model it selecting promising candidate using a purpose built in tree–based machine learning classifiers. With the help new approaches they have discovered 20 new pulsars.The authors [5]have explained the discovery of a new pulsar survey by using the Parkes Radio Telescope. The high time and frequency resolution of our digital backend system leads to increased sensitivity for short period, high-DM pulsars compared to previous surveys.

## III METHODOLOGY

(a) **Classification:** Classification is a form of data analysis that extracts models describing important data classes. Such models, called classifiers, predict categorical (discrete, unordered) class labels[6]. We can build a classification model to classify the dataset in to different groups or class. Classification has several types of applications, including fraud detection, target marketing, performance prediction, manufacturing, and medical diagnosis.

(i) **Conditional Inference Tree (CTREE):** In addition to traditional decision trees, conditional inference trees are another popular tree-based method. Like traditional decision trees, conditional inference trees also divide the data recursively by performing a univariate division in the dependent variable. However, what makes conditional decision trees have conditional inference trees is that conditional inference trees adapt meaning verification procedures to select variables instead of selecting variables maximizing information measures. In this way, we will present how to adjust a conditional inference tree to construct a classification [7].

(ii) **Classification and Regression Tree (CART):** CART adopt a greedy (i.e., non back tracking) approach in which decision trees are constructed in a top-down recursive divide-and-conquer manner. Most algorithms for decision tree induction also follow a top-down approach, which starts with a training set of tuples and their associated class labels. The training set is recursively partitioned into smaller subsets as the tree is being built [6], [8]. CART is classification and regression tree uses recursive partitioning to split the training records into subdivision with similar target field ideals using Gini index.

**(iii) Ensemble Model:** When two classification techniques like CTREE Tree and CARTcombined it is called hybrid or ensemble model[9].

## IV RESEARCH DATA

The dataset was downloaded from UCI Machine Learning Repository. The HTRU (High Time Resolution Universe Survey) 2 dataset have total number of instance is 17898 with 1639 are positive instances and 16259 are negative instances. The total number of attributes (features) is 8 with an additional class label [10]. 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.

**Table 1**
**Descriptions of HTRU 2 Data Set**

| Sl. No. | Attributes | Details |
|---|---|---|
| 1 | Profile_mean | Mean of the integrated profile |
| 2 | Profile_stdev | Standard deviation of the integrated profile |
| 3 | Profile_skewness | Skewness of the integrated profile |
| 4 | Profile_kurtosis | Excess kurtosis of the integrated profile |
| 5 | DM_mean | Mean of the DM-SNR curve |
| 6 | DM_stdev | Standard deviation of the DM-SNR curve |
| 7 | DM_skewness | Skewness of the DM-SNR curve |
| 8 | DM_kurtosis | Excess kurtosis of the DM-SNR curve |
| 9 | Class | Negative and Positive |

## V CLASSIFICATION FRAMEWORK PROCESS

The Figure 1 shows the process flow the used classifier like CTREE and CART and their ensemble model. The HTRU 2 dataset classify the classifier used under 10 folds cross validation techniques. The finally obtained performance of the classifier is accuracy, sensitivity and specificity.

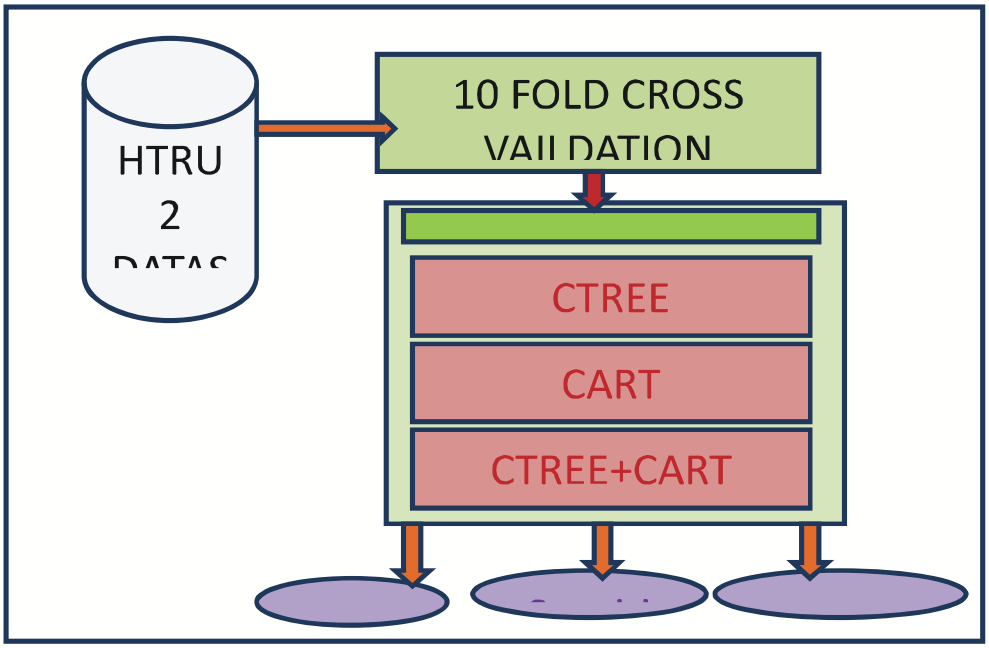


**Fig. 1: Process Flow of Classifier**

## VI RESULT AND DISCUSSION

After loading the HTRU 2 [10]dataset, the input and output attributes are selected and given to the classifier models, one by one, the results are automatically derived and presented in the form of various performances measurement. However, we considered three performances measurement like accuracy, sensitivity and specificity, as these three measures clearly reflect the efficiency of the classification model as shown in Table 2. Among three techniques of the classification ensemble model (CTREE+ CART) is producing remarkable results.

**Table 2**
**Comparison of Classification Performance**

| Sl. no. | Name of Algorithm | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|
| 1 | CTREE | 97.77 | 84.19 | 99.13 |
| 2 | CART | 92.96 | 87.48 | 84.25 |
| 3 | CTREE+CART | **97.84** | 84.56 | 99.18 |

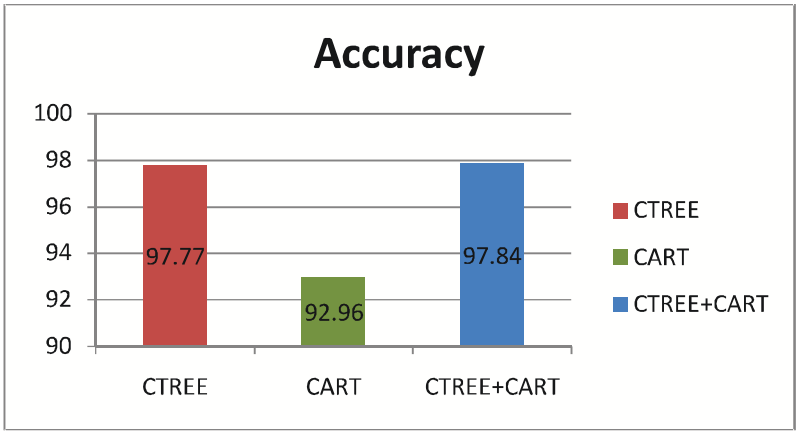


**Fig. 2: Comparative Accuracy graphs of the different Classifier**

The highest accuracy is obtained by ensemble model (CTREE+CART) compared to the individual models CTREE and CART.

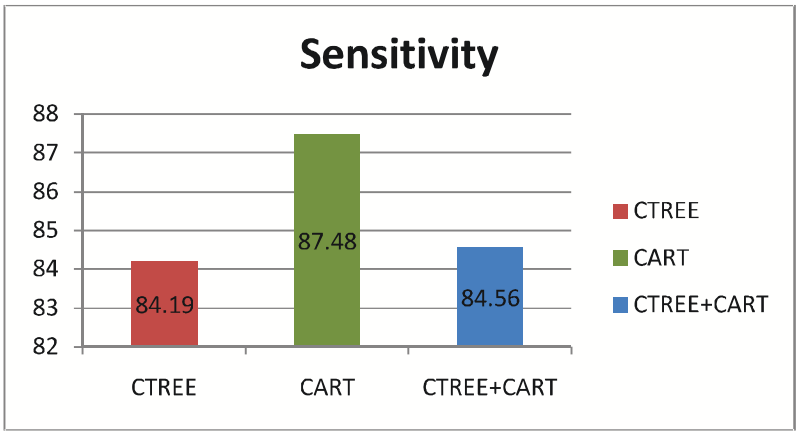


**Fig. 3: Comparative Sensitivity graphs of the different Classifier**

The figure 3 shows a fluctuation in the sensitivity of performance. CTREE gives 84.19% of sensitivity and CART gives 87.48% whereas ensemble model (CTREE+CART) gives 84.56% of sensitivity.

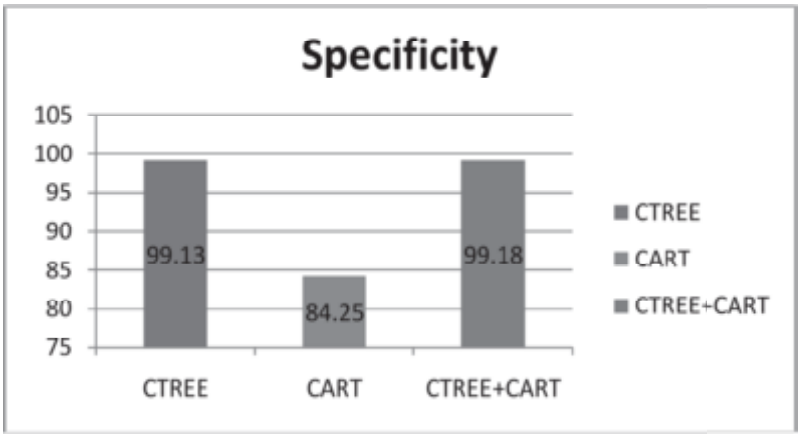


**Fig. 4: Comparative graphs of the different Classifier**

The maximum Specificity is obtained by ensemble model (CTREE+CART) compared to the individual models CTREE and CART.

The pulsar candidate can be very well predicted using many classifiers in data mining and machine learning. The purpose of this study was to analyze the application of data mining algorithms and machine learning in the HTRU2 dataset and to predict pulsar and non pulsar. In this paper, the pulsar candidate is predicted using two types of classification techniques CTREE and CART and their ensemble model. In ensemble model obtained the highest accuracy 97.84 % compared to other classification models.

There are other possible techniques for selecting the features that can be used to obtain better results from the ensemble model. Other classification techniques and the techniques for selecting features can be performed in the future.

## VII CONCLUSION

## REFERENCES

[1] S. . Saha, Aperture Synthesis Methods and Applications to Optical Astronomy. springer, 2011.

[2] P. S. Ramkumar and a. a. Deshpande, "Real-time signal processor for pulsar studies," J. Astrophys. Astron., vol. 22, no. 4, pp. 321–342, 2001.

[3] R. J. Lyon, "WHY ARE PULSARS HARD TO FIND ?," 2016.

[4] R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, and J. D. Knowles, "Fifty Years of Pulsar Candidate Selection : From simple filters to a new principled real-time classification approach," vol. 22, no. March, pp. 1–22, 2016.

[5] M. J. Keith et al., "The High Time Resolution Universe Pulsar Survey – I . System configuration and initial discoveries 1 I N T RO D U C T I O N 2 S I M U L AT I O N A N D S U RV E Y S T R AT E G Y," vol. 627, pp. 619–627, 2010.

[6] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques, Third. Elsevier, 2012.

[7] T. Hothorn, K. Hornik, and A. Zeileis, "Unbiased Recursive Partitioning: A Conditional Inference Framework," J. Comput. Graph. Stat., vol. 15, no. 3, pp. 651–674, 2006.

[8] A. Pujari, Data mining techniques, Third. University press, 2013.

[9] S. Haykin, Neural Networks and Learning Machines, vol. 3. 2008.

[10] R. Lyon, "HTRU2," 2016. [Online]. Available: https://figshare.com/articles/HTRU2/3080389/1. [Accessed: 01-Dec-2017].