Stellar and Pulsar Classification using Machine Learning

Jishant Talwar[1], Parul Jain[2], Chetana Jain[3*], Baljeet Kaur[4*]

[1]B.Sc. (H) Physics II year, Hansraj College, University of Delhi
[2]B. Sc. (H) Physics III year, Hansraj College, University of Delhi
[3]Department of Physics, Hansraj College, University of Delhi
[4]Department of Computer Science, Hansraj College, University of Delhi
**Corresponding Authors*: chetanajain11@gmail.com and baljeetkaur26@hotmail.com

Abstract

This work presents a review of results from a study undertaken as part of a course on *Machine Learning for Physics and Electronics*. The course was designed to give an overview of applications of machine learning to the undergraduate students of Physics and Electronics. This article discusses the results from two projects undertaken by students on application of supervised machine learning in astrophysics. One project is based on stellar classification while the second one focuses on pulsar classification. For both the projects, publicly available data was used and supervised machine learning techniques such as, the *k*-Nearest Neighbour classifier, the decision tree classifier and, the random forest classifier, were tested to predict the efficacy of the models. The models formulated give encouraging results which are comparable to the state of the art results.

Keywords: Astronomy, Stellar Classification, Pulsar Classification, Supervised Machine Learning

Introduction

Astrophysics is one of the most fascinating themes in physics. And an ever-increasing knowledge of our Universe is driven by large amount of data, which is regularly released by various surveys such as the Sloan Digital Sky Survey (SDSS, York et al. 2000) and the High Time Resolution Universe Survey (HTRU, Keith et al. 2010; Levin et al. 2013). The technological advancement over the past decade has led the astronomers to look for scalable and judicious data handling. In this context, machine learning and deep learning techniques have become increasingly popular, and are now being used to understand various astrophysical models (Bailer-Jones et al. 1998; Ball et al. 2004; Miller et al. 2015).

This review focuses on two areas of astrophysics. The first one deals with stellar classification schemes which are useful in astronomical studies involving Galaxy formation (Seeds and Backman 2011). The stellar classification is based on the Hertzsprung-Russel diagram (Morison 2008) which classifies stars based on their absolute magnitude (or luminosity) and their effective temperature (or spectral type).

The second study involves classification of pulsars based on their characteristic pulse profile. The statistical measures of the integrated profile and the Dispersion Measure - Signal to Noise Ratio curve (DM curve) enhances the understanding of emission mechanism from compact objects, the structures in interstellar medium and verification of the existing astrophysical models (Lyon 2016).

Machine learning techniques enable a decision system to learn and improve its prediction with experience. There are several algorithms of machine learning which are used to achieve classification, clustering or regression, based on the problem domain. The *k*-Nearest Neighbour (*k*-NN) classifier, the decision tree classifier and, the random forest classifier have been used in this article. The *k*-NN algorithm works on the principle of feature similarity and classifies a data point on how closely it matches the points in the training set. The decision tree classifies instances by sorting them from the root of a tree structure down to some leaf node. The nodes of the tree test some specific attribute of the instance and each branch descends from that node based on the outcome of the test. This process is repeated for each sub tree rooted at a new node. Random Forest consists of an ensemble of individual decision trees that are generated using a random selection of attributes at each node to determine the split. To classify and instance, each tree votes and the majority class is chosen (James et al. 2013; Mitchell 1997).

In this work, for the stellar classification as well as for the study of pulsars, data analysis based on the characteristic features of the dataset has been carried out extensively and reported. Decision models based on different classifiers have also been obtained and compared with other reported results.

DATA ANALYSIS AND DECISION MODELS

Stellar Data

The publicly available data used for understanding stellar classification has been taken from the *Kaggle* repository for machine learning (https://www.kaggle.com/deepu1109/star-dataset). This data is based on a series of astrophysical equations, such as, those governing the blackbody radiation, the stellar magnitude and stellar parallax. The data comprises of 240 stars, each having seven attributes – absolute temperature, luminosity (in solar units, L/Lsun), radius (in solar units, R/Rsun), absolute magnitude, stellar type, stellar colour, and spectral class (O, B, A, F, G, K, M). Stellar types consist of brown dwarfs, red dwarfs, white dwarfs, main sequence, super-giants and hyper-giants. Each stellar type has 40 stars, thereby making this a balanced classification problem. The temperatures of all the stars lie in the range of 2000 K to 40000 K with a majority of stars having temperatures less than 10000 K.

Data Analysis

Figure I shows the distribution plot of the entire data as a function of temperature. It shows that most of the stars in the stellar data lie in a temperature range of 0 K – 10000 K. This range corresponds to the stellar spectral class M.

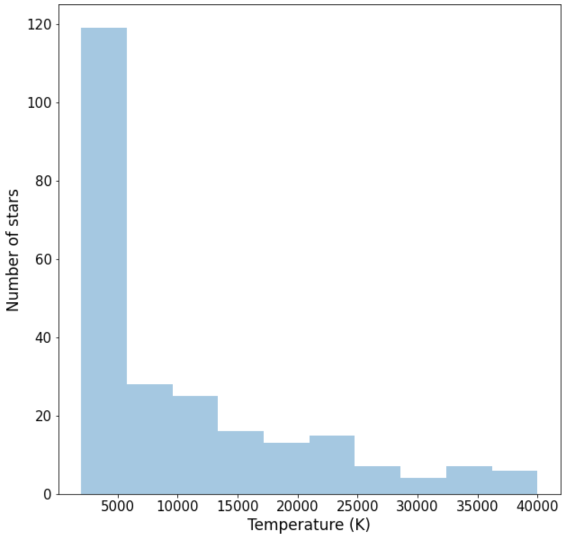


Figure I: Distribution of data over the temperature range

Figure II: Heatmap of correlation coefficients for the stellar data

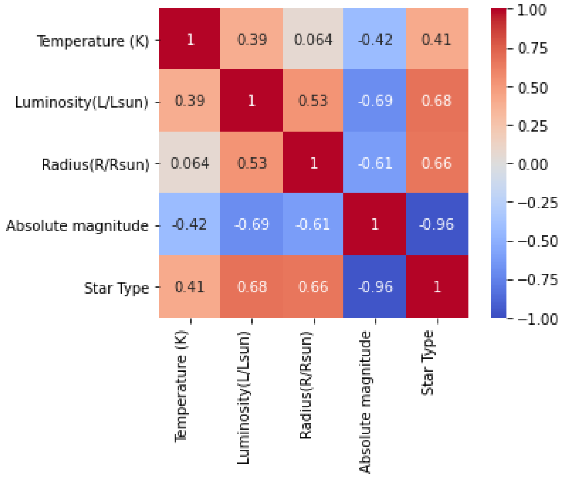


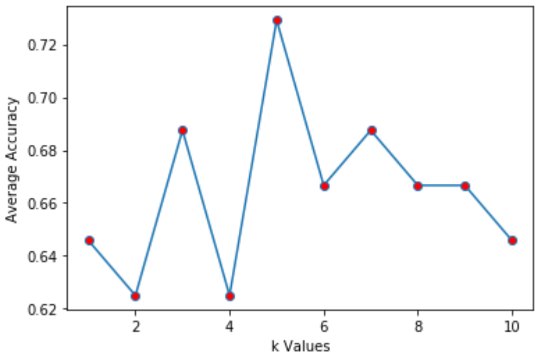
Figure II is the heatmap of the correlation matrix of different attributes given in the data. It is observed that absolute magnitude is maximally correlated (-0.96) to the target classes. Correlation is also observed with luminosity and radius.

Classification Models

For our experiments, we worked with the *k*-NN, the decision tree, and the random forest classifiers. The performance measures that evaluate these classifiers are the accuracy, the recall rate, the precision rate and the F1 - score. The accuracy is defined as the ratio of correctly classified objects to the total number of data points, recall measures the ratio of number of correctly classified objects to their actual number in the data, precision measures the correctness of the classification, and F1 - score is the weighted average of precision and recall.

In the *k*-NN classification, the experiments were performed for the *k* value ranging from 1 to 10 and the average accuracy over 10 runs for each of the *k* values is shown in Figure III. In order to compare with the known results (https://www.kaggle.com/sachinsharma1123/accuracy-score-of-98-using-random-forest), 20% of the entire data was taken as the test data and unique train-test splits were chosen in all the runs. As can be observed in this figure, five nearest neighbours give the best result with an average accuracy of 0.729.

Figure III: Average accuracy at different values of *k*



In the decision tree classification, the experiments were performed for 10 runs, each with a unique train-test split, with test data comprising of 30% of the total data for comparing with the state of the art results (https://www.kaggle.com/taghredsalah199/journey-beyond-the-stars-with-accuracy-100) available for the same dataset. The results of the decision tree classification scheme have been discussed in the following section.

In the random forest classification, the experiments were performed for the number of decision trees ranging from 100 to 200. The average performance accuracy over 10 runs was 1.0. The test data comprised of 20% of the entire data.

Results and Observations

The confusion matrix for each classifier mentioned in Section 2.1.2 is given in Table I. In this table, the letters A, B, C, D, E, and F respectively represent brown dwarfs, red dwarfs, white dwarfs, main sequence, super-giants, and hyper-giants.

Table I: Confusion matrices for stellar data

	k-NN classifier						Decision-Tree classifier					
	A	B	C	D	E	F	A	B	C	D	E	F
A	10	4	0	0	0	0	13	0	0	0	0	0
B	2	4	0	0	0	0	0	8	0	0	0	0
C	0	0	10	0	0	0	0	0	12	0	0	0
D	0	0	3	4	0	0	0	0	0	17	0	0
E	0	0	0	0	2	3	0	0	0	0	7	0
F	0	0	0	0	5	1	0	0	0	0	0	15

	Random Forest classifier					
	A	B	C	D	E	F
A	14	0	0	0	0	0
B	0	6	0	0	0	0
C	0	0	10	0	0	0
D	0	0	0	7	0	0
E	0	0	0	0	5	0
F	0	0	0	0	0	6

Table II: Classification report for the stellar data

	k-NN classifier				Decision Tree classifier			
	Precision	Recall	F1 Score	Support	Precision	Recall	F1 Score	Support
A	0.83	0.71	0.77	14	1.00	1.00	1.00	13
B	0.50	0.67	0.57	6	0.94	1.00	0.97	8
C	0.77	1.00	0.87	10	1.00	1.00	1.00	12
D	1.00	0.57	0.73	7	1.00	0.93	0.96	17
E	0.29	0.40	0.33	5	1.00	1.00	1.00	7
F	0.25	0.17	0.20	6	1.00	1.00	1.00	15
Accuracy			0.65	48			0.99	72

	Random Forest classifier			
	Precision	Recall	F1 Score	Support
A	1.00	1.00	1.00	14
B	1.00	1.00	1.00	6
C	1.00	1.00	1.00	10
D	1.00	1.00	1.00	7
E	1.00	1.00	1.00	5
F	1.00	1.00	1.00	6
Accuracy			1.00	48

A comparison of the three confusion matrices for each classifier used, indicates that the *k*-NN classification scheme yields the highest number of incorrect prediction across the stellar types. For example, out of a total of 14 objects identified as brown dwarfs, only 10 have been correctly classified. The decision tree classifier and the random forest classifier give more accurate results.

Figure IV: H-R diagram of predicted stars by each classifier

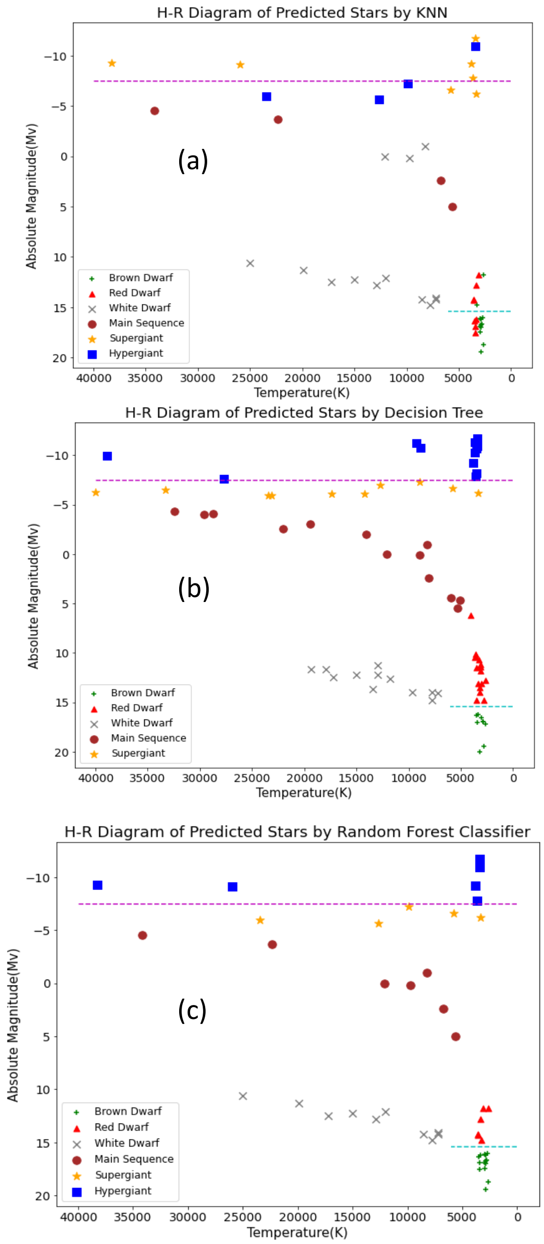


Table II shows the classification report for each classifier. Here also, the letters A, B, C, D, E, and F respectively represent brown dwarfs, red dwarfs, white dwarfs, main sequence, super-giants, and hyper-giants. The performance measures have been mentioned in each case. It can be clearly seen from Table II that the best performance measures are obtained for the decision tree classifier, and the random forest classifier. As compared to the *k*-NN classifier, there is a considerable improvement in the performance metrics (precision, recall and F1 -score) for decision tree and random forest classifiers.

Table III: Comparison of results of stellar data with the state of the art

	***k*-NN classifier**		**Decision Tree classifier**		**Random forest classifier**	
	Present work	**Link 1**	**Present work**	**Link 2**	**Present work**	**Link 1**
	F1 score	**F1 score**	**F1 score**	**F1 score**	**F1 score**	**F1 score**
A	0.77	-	1.00	1.0	1.00	-
B	0.57	-	0.97	1.0	1.00	-
C	0.87	-	1.00	1.0	1.00	-
D	0.73	-	0.96	1.0	1.00	-
E	0.33	-	1.00	1.0	1.00	-
F	0.20	-	1.00	1.0	1.00	-
Maximum Accuracy	0.73	0.70	0.99	1.0	1.00	0.979

Link 1: https://www.kaggle.com/sachinsharma1123/accuracy-score-of-98-using-random-forest
Link 2: https://www.kaggle.com/taghredsalah199/journey-beyond-the-stars-with-accuracy-100

In order to support the metrics given in Table II, the H-R diagram was also generated from results of each of the classifier mentioned above. This is shown in Figure IV. In this figure, the panel marked (a) corresponds to the *k*-NN classification scheme, panel marked (b) shows the results from the decision tree classifier, and the panel marked (c) gives the results from the random forest classification scheme. It can be clearly seen from Figure IV (a), that the *k*-NN classification scheme is not as accurate as the other two classifiers. As an example, there is an obvious overlap between the red dwarfs (symbolised by solid triangle) and the brown dwarfs (symbolised by plus sign). However, the other two classifiers show a clear distinction between these two stellar types.

It was observed that for decision tree and random forest classifiers, the most relevant differentiating features for the seven stellar types were absolute magnitude and radius of the stars.

Table III shows a comparison of results from this work with the state of the art results which are available at Link 1 and Link 2 (details of these links have been mentioned in Table III). The results are comparable. For the *k*-NN classifier, the maximum accuracy obtained is 0.73 which is better than that of the compared result. For the decision tree classifier, the F1 - score of our model is same for four stellar classes and accuracy is marginally less than the reported value. For the random forest classifier, the accuracy obtained is 1.0.

Pulsar Data

The pulsar classification problem has been studied by using the HTRU data which is publicly available from the University of California, Irvine (UCI) machine learning repository (https://archive.ics.uci.edu/ml/datasets/HTRU2). This data consists of an imbalanced distribution of 17,898 data points comprising of 9.2% pulsar candidates and 90.8% non-pulsars. The attributes in this data include the mean, standard deviation, excess kurtosis, and skewness of the integrated pulse profile; and the mean, standard deviation, excess kurtosis and, skewness of the DM curve.

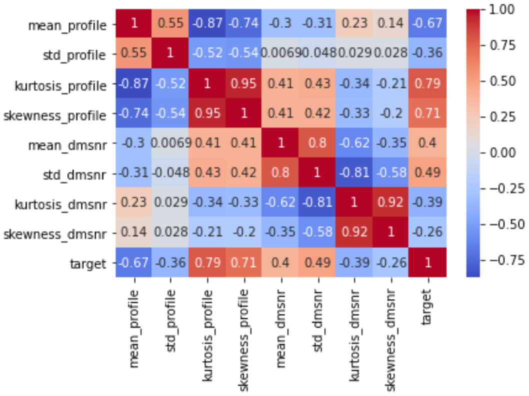


Figure V: Heatmap of correlation coefficients for the pulsar data

Data Analysis

Figure V shows the heatmap of the correlation matrix amongst the attributes and the target class (pulsars/non-pulsars). A strong correlation of the target class with the excess kurtosis of the integrated profile (0.79) and skewness of the integrated profile (0.71) was observed. A moderate correlation with the standard deviation of the DM curve (0.49) and mean of the DM curve (0.4) was also found. This heatmap is also indicative of the within-attribute correlations.

Figure VI shows the violin plots of all the attributes separately for both the classes (pulsars: 1 and non-pulsars: 0). It was plotted to examine the differences between the attribute values of pulsars and non-pulsars. The main observations from Figure VI are:

1. The mean of the integrated profile of non-pulsars is higher than that of pulsar candidates.
2. The distribution of all the statistical measures is distinctive in all the eight violin plots.
3. In the case of kurtosis of the integrated

profile, the distribution of non-pulsars is peaked about its mean whereas in the case of pulsars, a uniform spread can be seen.

4. In the case of the skewness of the DM, the pulsars exhibit a peaked distribution in comparison to a smoother distribution of non- pulsars.

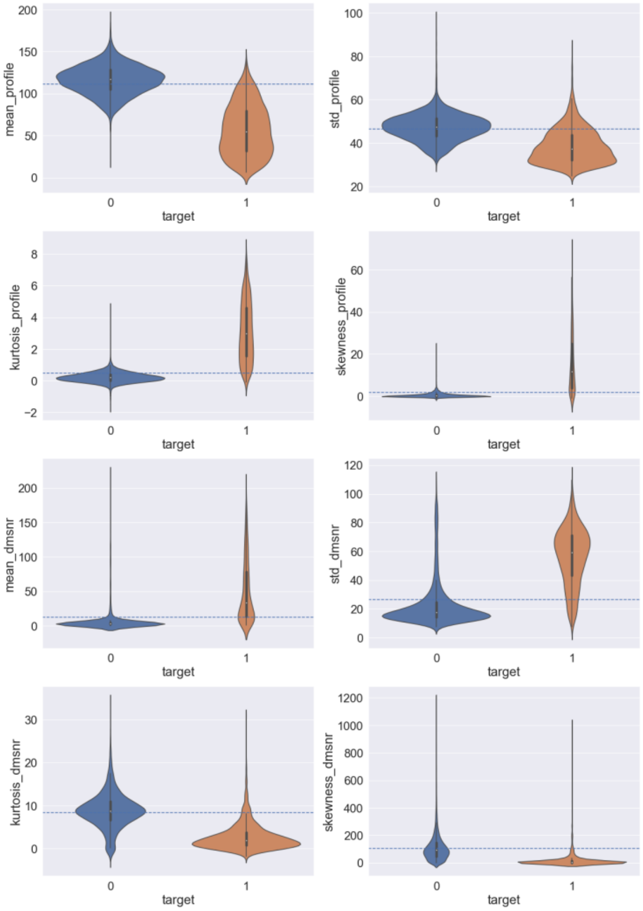


Figure VI: Violin plots for pulsars and non-pulsars on the basis of statistical parameters

Classification Models

For our experiments, we worked with the *k*-NN, the decision tree, and the random forest classifiers. The performance measures that evaluate these classifiers are the accuracy, the recall rate, the precision rate and the F1 - score. The accuracy is defined as the ratio of correctly classified objects to the total number of data points. The recall measures the ratio of correctly identified pulsars to the total number of real pulsars in the data. The precision gives the number of real pulsars that can be predicted correctly out of the total candidates identified as pulsars. The F1 - score is a trade-off between the recall and the precision and it is defined as the harmonic mean of recall and precision.

In the *k*-NN classification, the experiments were performed for the *k* value ranging from 1 to 10 and the average accuracy over the 10 runs for each of the *k* values is shown in Figure VII. 30% of the entire data was taken as the test data and unique train-test splits were chosen in all the runs empirically. As can be observed in this figure, nine nearest neighbours give the best result with an average accuracy of 0.973.

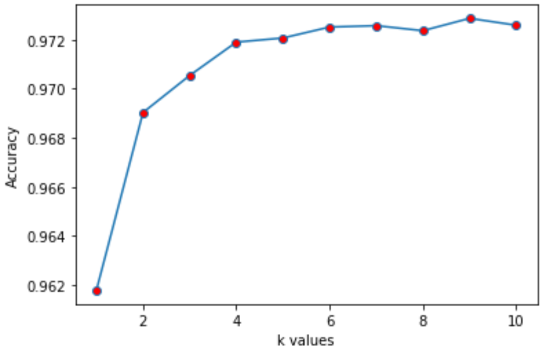


Figure VII: Average accuracy of pulsar data at different values of *k*

In the decision tree classification, the experiments were performed for 10 runs, each with a unique train-test split, with test data comprising of 30% of the total data chosen empirically. The results of the decision tree classification scheme have been discussed in the following section.

In the random forest classification, the experiments were performed for the number of decision trees ranging from 100 to 200. The average performance accuracy over 10 runs was 0.983. The test data comprised of 30% of the total data.

Table IV: Confusion matrix for pulsar data

	Predicted					
	k-NN classifier		Decision Tree classifier		Random Forest classifier	
Actual	Non Pulsars	Pulsars	Non Pulsars	Pulsars	Non Pulsars	Pulsars
Non Pulsars	4859	35	4802	89	4829	38
Pulsars	108	368	96	383	82	421

Results and Observations

Table IV gives the confusion matrix for the pulsar data for each classifier. Based on the results quoted in Table IV, a comparison between the three classifiers is given below.

1. The number of true positives for the real pulsars (correct prediction of real pulsars) is marginally more (421/503) in case of the random forest classifier as compared to that obtained in the decision tree (383/479) and the *k*-NN classifiers (368/476).
2. The number of false positives (wrongly classified as pulsars) is significantly less in case of the *k*-NN classifier (35/4894) and the random forest classifier (38/4867) as compared to the decision tree classifier (89/4891).

Table V gives the classification report of pulsar data for each classifier. Since the pulsar data is unbalanced, it is imperative to report the efficacy of the models in terms of precision, recall and F1 score. It can be inferred from this table, that the precision (0.92) and recall (0.84) for pulsars is higher for the random forest in comparison to the *k*-NN and the decision tree classifiers. Moreover, the precision from the *k*-NN classifier is comparable with the precision from the random forest classifier.

Table V: Classification report for the pulsar data

	k-NN classifier				Decision-Tree classifier				Random Forest classifier			
	Precision	Recall	F1 Score	Support	Precision	Recall	F1 Score	Support	Precision	Recall	F1 Score	Support
Non-Pulsars	0.98	0.99	0.99	4894	0.98	0.98	0.98	4891	0.98	0.99	0.99	4867
Pulsars	0.91	0.77	0.84	476	0.81	0.80	0.81	479	0.92	0.84	0.88	503
Accuracy			0.97	5370			0.97	5370			0.98	5370

It was observed that in the case of decision tree classifier, the most relevant classification feature was the excess kurtosis of the integrated pulse profile. In the case of random forest classifier, excess kurtosis of the integrated pulse profile along with the skewness of the integrated pulse profile was found to be the important distinguishing feature.

Table VI shows a comparison of the results of this work with the state of the art. The results have come out to be comparable, for the *k*-NN, the decision tree, as well as the random forest classifier.

Table VI: Comparison of Results of HTRU 2 with the state of the art

	***k*-NN classifier**		**Decision-Tree classifier**		**Random forest classifier**	
	Results	**(Sardana et al. 2017)**	**Results**	**(Sardana et al. 2017)**	**Results**	**(Sardana et al. 2017)**
	F1 Score	**F1 Score**	**F1 Score**	**F1 Score**	**F1 Score**	**F1 Score**
Non Pulsars	0.99	0.99	0.98	0.98	0.99	0.98
Pulsars	0.84	0.86	0.81	0.82	0.88	0.87
Accuracy	0.97	0.98	0.97	0.97	0.98	0.98

CONCLUSION AND FUTURE DIRECTIONS

The machine learning techniques on the stellar and the pulsar datasets have yielded encouraging results. The stellar data is a balanced classification problem whereas the pulsar data is unbalanced. The experiments have been averaged over numerous runs of train and test splits and have resulted in high accuracy scores for both the data sets. The correlation of the attributes is depicted in terms of heatmaps, which illustrate the strength of relationship amongst the attributes and the target class. The promising models by *k*-Nearest Neighbour classifier, the decision tree classifier and, the random forest classifier, give us impetus to carry out further experiments with other stellar and pulsar data which are considered to be more challenging. Intensive cross validation can be carried out for the data. We shall also look into pre-processing, feature selection, feature extraction and transformation strategies on these data.

This paper is an attempt to assimilate the understanding of the undergraduate students of the machine learning techniques for Physics and Electronics. It will be further substantiated with an in-depth and better machine learning techniques for data preparation, transformation and prediction with larger and more challenging databases. In the case of pulsar data, methods to balance the unbalanced dataset may be explored, so as to reflect the true accuracy of the performance of each classifier. Having built a strong foundation of machine learning techniques, this work can be extended to include larger database involving more stellar types such as sub-giants and giants; and more spectral classes.

REFERENCES

1. Bailer-Jones, C. A. L., Irwin, M., & von Hippel, T. (1998). Automated classification of stellar spectra - II. Two-dimensional classification with neural networks and principal components analysis. *Monthly Notices of the Royal Astronomical Society*, 298(2), 361-377.
2. Ball, N. M., Loveday, J., Fukugita, M., et al., (2004). Galaxy types in the Sloan Digital Sky Survey using supervised artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, 348, 1038-1046.
3. James, R., Witten, G., Hastie, D., et al., (2013). *An Introduction to Statistical Learning with Applications*. Springer.
4. Keith, M., Jameson, A., Van Straten, W., et

al., (2010). The High Time Resolution Universe Pulsar Survey - I. System configuration and initial discoveries. *Monthly Notices of the Royal Astronomical Society*, 409, 619-628.

5. Levin, L., Bailes, M., Barsdell, B., et al., (2013). The High Time Resolution Universe Pulsar Survey -VIII. The Galactic millisecond pulsar population. *Monthly Notices of the Royal Astronomical Society*, 434, L1387-L1397.
6. Lyon, R. J. (2016). *Why Are Pulsars Hard to Find?*. PhD Thesis, University of Manchester.
7. Mitchell, T. (1997). *Machine Learning*. McGraw Hill.
8. Miller, A. A. (2015). VizieR Online Data Catalog: Machine learning metallicity predictions using SDSS. *Astrophysical Journal*, 811, 30
9. Morison, I. (2011), *Introduction to Astronomy and Cosmology*. Wiley Publishers
10. Seeds, M. A. & Backman, D. (2011). *Universe: Solar System, Stars, and Galaxies*.
11. York, D. G., Adelman, J., Anderson, J. E., et al. (2000). The Sloan Digital Sky Survey: Technical Summary. *The Astronomical Journal*, 120(3), 1579-1587.
12. https://github.com/AshishSardana/pulsars-candidate-classifier