Artificial & Computational Intelligence/Published Online: July 2021
https://acors.org/ijacoi/VOL2_ISSUE2_6.pdf

Artificial & Computational Intelligence

Applied Ensemble Learning in Classification of Diverse Celestial Bodies Imaged Using SDSS

[1]Anukriti, [2]Vandana Niranjan
Department of Electronics and Communication Engineering
Indira Gandhi Delhi Technical University for Women, Kashmere Gate, Delhi, IN
[1]anukriti0009@gmail.com, [2]vandana_niranjan@yahoo.com

Abstract

Space exploration is a growing and dense pool for research ever since the human civilization has started understanding the cosmos. Researchers have been trying to identify various wonders of nature that exists in the cosmos and in recent years, with the development of high intensity telescopes, abundant celestial objects are found by researchers. In this research, we have queried the galactic data representation shared for public use by Sloan Digital Sky Survey to understand some features and took the advantage of Ensemble Learning Techniques to classify variety of celestial entities based on their features learnt by the models. Further we have improvised the classification accuracy using the ensemble boosting techniques to produce a much better classification result. As the conclusion of this research, the model is able to classify a variety of objects with very high accuracy.

Keywords: Astrophysics, Object Classification, Ensemble Learning, Boosting, SDSS-DR16, Cosmos

1. Introduction

The celestial space is an astonishing realm for research since early humans and this antique division of physics consists of most diverse entities. Human civilization had been trying to understand and classify the celestial bodies which they can observe based on their peculiarities and effects on human life. As the field of science advances with the use of lens and scopes along with other optical instrumentation for sensing bodies, resulting in more knowledge on celestial life. Cosmological Classifiers like Annie J. Canon[1]and Williamina Fleming in the early 20th century used to classify the entities based on various parameters. However the process of classification done manually by researchers is arduous and time consuming. With the development of Intelligent Machine Learning Systems, the efficiency of classifying new celestial entity has grown considerably as shown in figure 1. Classification of cosmic objects using photometric data is a rapidly developing domain of research. Since there is a huge number of celestial object detection, tools like SExtractor[3]are becoming known for distinguishing

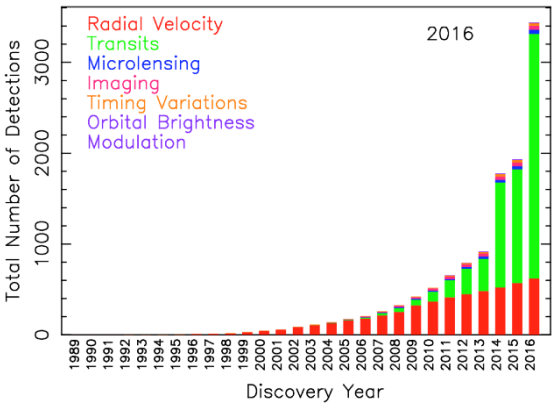

Figure 1 shows the NASA Exoplanet Archive of 2016 showing the growth of discovery inexoplanets from year 1989 to 2016.

theisolated stars and galaxies appearing of the same size. The objective of the paper is to present an efficient machine learning model using the ensemble supervised learning procedure that could be used for the classification of huge dataset with an expense of significantly low computational power.

2. Detection of the cosmic objects

The detection of cosmic object is a historic process and there has been a lot of improvement in the techniques of detecting the object. There are various historic ways for observing cosmic objects.

2.1 By Visual Observation

In order to approximate the visibility accurately some of the factors are specified for brightness and probability of detection of objects such as Physical characteristics including, shape, luminosity and spectral nature. The brightness and character of the background or field against which the object is being observed. The visual acuity of the observer and the magnifying power of his optical aids. The spectral absorption characteristics and the light-transmitting ability of the medium between object and observer. The apparent motion of the object, and the change in its appearance as a function of time. In normal scenario extraterrestrial objects through the Earth's atmosphere, which scatters light, is often cloudy, having constant motion[4],[5]. The movement of the atmosphere imposes a severe limit on the resolving power of large telescopes. A large telescope employed above the atmospheres such as on the surface of the Moon-would have a greatly enhanced ability to resolve details on the surfaces of the planets and the Sun and would tremendously improve man's ability to explore the entire visible universe[7].

2.2 By Infrared Observation

One of the most fundamental physical phenomena which is responsible for this is the radiations in the infrared portion of the electromagnetic spectrum that interact strongly with the molecular structure of matter as physical objects emits thermal radiation whose intensity is a rapidly increasing function of its temperature.

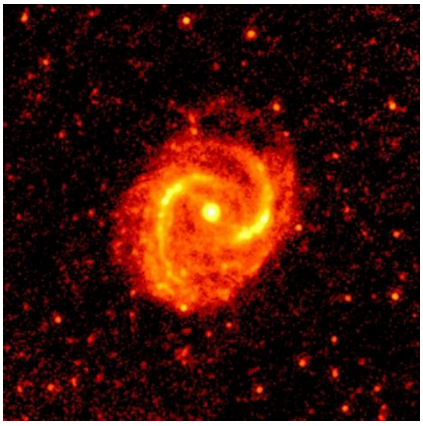

Figure 2 shows a mid-infrared image of M91 as seen by the Spitzer Space Telescope (Image credit: G. J. Bendo/JBCA/NASA)

Transmission of infrared radiation serves as a probe of the structure and composition of chemical and biological matter[8]. Infrared probes have made is easier to understand the composition of atmospheres for the planets and other celestial objects. The principal hindrance to such extraterrestrial observations has been the confusion offered by the intervening Earth's atmosphere. Infrared observation of the Earth's atmosphere from an orbiting vehicle permits measurements of methane clouds, water-vapor, and carbon-dioxide concentrations, and the like, which is valuable for meteorological observations.

2.3 Electromagnetic Detection

Electromagnetic radiation detection has been most popular way to detect astronomical objects while today modern astrophysicists are also making observations using neutrinos, cosmic rays or gravitational waves. Observing a source using multiple methods is known as multi-messenger astronomy[9]. Optical and radio astronomical methods are used by situating them at heights to minimize absorption and distortion caused by earth’s surface, and are therefore performed with ground-based observatories.

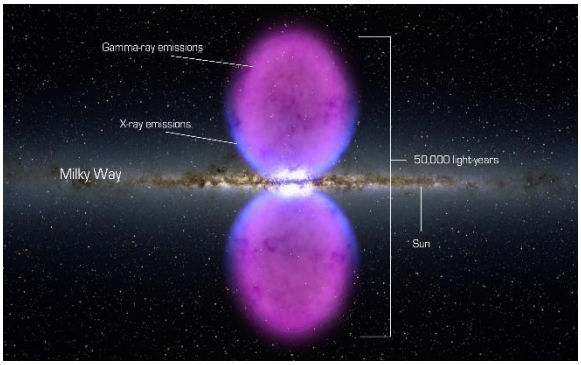

Figure 3 shows electromagnetic elements of milky-way (Image credit: NASA's Goddard Space-Flight Center)

Infrared observatories are well suited at dry places as wavelengths of infrared light are heavily absorbed by vapor. Observations are carried out in space or balloons for X-ray astronomy, gamma-ray astronomy, UV astronomy shown in figure 3 due to atmosphere being opaque at the wavelengths used by astronomy. Powerful gamma rays are detected by their large air showers therefore the study of cosmic rays is a rapidly expanding branch of astronomy.

2.4 By Sloan Digital Sky Survey

The Sloan Digital Sky Survey or SDSS is a major multi-spectral imaging and spectroscopic redshift survey using a dedicated 2.5-m wide-angle optical telescope at Apache Point Observatory shown in figure 4. Images were taken using a photometric system of

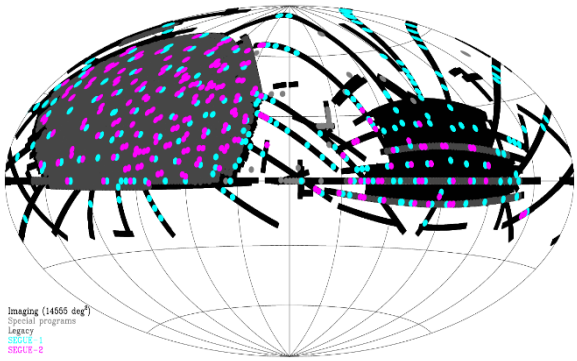

Figure 4 shows digital sky survey of 2016-2020[16] (Image credit: SDSS)

five filters. These images are processed to produce lists of objects observed and various parameters, such as whether they seem point-like or extended (as a galaxy might) and how the brightness on the CCDs relates to various kinds of astronomical magnitude[24].

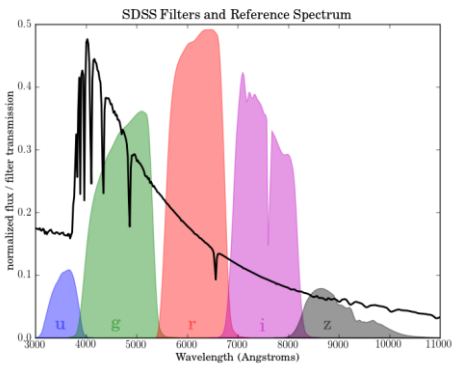

Figure 5 shows SDSS filter and spectrum with five different wavelength filters namely U, G, R, I, Z and the normalized flux transmitted for the wavelength to calculate the redshifts[14].

In the SDSS DR-16[14],[17] database the telescope observes the space using various optical filters having distinct central wavelengths ranging from 3351Å to 8932Å[19]-[30]. Using these photometric data, stars, galaxies, and quasars are also selected for spectroscopy[13].

3. The machine learning approach

3.1 Ensemble learning

In recent years, the multi-class classification procedure recognized as ensemble learning has gained much popularity. The sanity being its wide domain of applications and its efficacy in decision making[6]. The use of ensemble learning is in multifariousness of machine learning conundrum, such as selection of features, error correction, imbalanced classification of data, and a variation of learning problems drifted from non-stationary distribution.

The ideology of partitioning the sample space using multi-class classification is the key for preliminary ensemble learning technique[10]. Further work in a optimization procedure called boosting by[11] exhibit that a classifier model functional on a two-category classification problem with high accuracy, can also be constructed from an ensemble of classifiers with even significant accuracy. This concept of boosting provided a substructure for AdaBoost algorithms which spread out the boosting concept to multi-class and regression problems[12].

For constructing an effective Ensemble Learning Model, three verticals are required to be identified. The first vertical is sampling or selection of data. The most prevalent methodology for data selection is using different subset of training data. The bootstrap method is a methodology for quantifying about data population by averaging estimates from numerous tiny data samples[13]. The next vertical is training the member classifiers. Each of the ensemble learning system is maneuvered to train individual ensemble members. Bagging, Boosting and hierarchical Mixture of Experts are the most commonly employed approaches. The final vertical is combining all the individual classifiers. The variation in strategy used in this step depends on the type of classifiers used as ensemble members. Classifiers which provide only discrete-valued label outputs such as SVM uses majority

voting. Other classifiers which provide continuous valued class-specific outputs, such as Naive Bayes uses arithmetic combiners or any different urbane decision making, in addition to majority voting[6].

3.2 Proposed Model

The model that we are proposing is based entirely on ensemble learning methodology with improvised optimization. The following is the algorithm for the model development:

1. Getting the FITS Data.
2. Feature Selection
3. Splitting the Data Set
4. Training the Classifier(s)
5. Validating the Model for Test Data Set
6. Getting the Accuracy of the respective model(s)

3.3 Querying the Data

The dataset for this research is obtained from the Solan Digital Sky Survey (SDSS)-IV:2014–2020[17] DR-16, which is a project that provides public database on observations of celestial objects. The latest dataset extends the precision of cosmological measurements and its infrared spectroscopic survey of the galaxy.

SDSS DR-16 database consists of terabytes of query data, however our intended data resides in imaging and spectra tables. From these tables, we queried first 400,000 entries from the SAS (Science Archive Server) for FITS file retrieval.

In python, working directly on a FITS file is rather another challenge. Hence for working on a FITS file, the astropy library is employed.

3.4 Feature Selection

Monitoring the bulk of the data set, filtering out important feature is a crucial step for a successful model development. In the SDSS DR-16 database the telescope observes the space using various optical filters having distinct central wavelengths namely Ultraviolet at 3551Å, Visible Green at 4686Å, Visible Red at 6166Å, Infrared at 7480Å, Gamma at 8932Å [14],[17],[19],[24] as shown in figure 5 and are the features filtered for model development. Location is also an important feature for identification of location of the object in celestial space. Hence, the Right Ascension and Declination of fiber as shown in figure 6 are filtered.

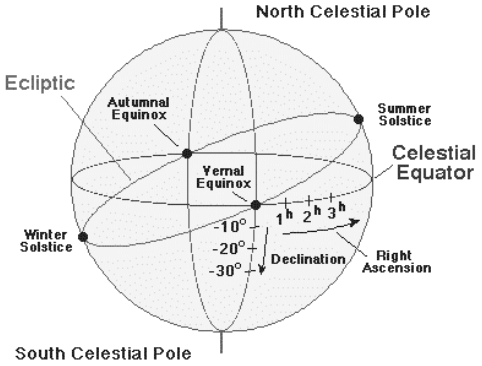

Figure 6 shows the Right Ascention (hours) and Declination (degrees) for a celestial object.

Another important feature selected is the value of Red shifts. A value less than zero is for stars, whereas value greater than 5 represent some QSOs.

The Petrosian is another important feature used to compare the radial profiles of galaxies at different distances. Concentration is similar to the luminosity profile of the galaxy, which measures what proportion of a galaxy's total light is emitted within what radius. A simplified way to represent concentration is to take the ratio of the radii containing 50\% and 90\% of the Petrosian flux.

$$\text{Concentration} = \frac{\text{Petr } o_{50}}{\text{Petr } o_{90}} \quad (1)$$

Additional features that has to be created are the color index which is the difference between two measurements of the magnitude made at different wavelengths, the value found at the longer wavelength being subtracted from that found at the shorter, i.e. U-G, G-R, R-I, I-Z[14]. The index is a measure of a star's color, an indication of its temperature, and a fairly crude description of the distribution of its radiated energy through the electromagnetic spectrum. Hot, blue stars have negative color indices, as they radiate most strongly and therefore have numerically lower magnitudes at short wavelengths, and those of cool, red stars are positive. The color index of a star is increased by the passage of its light through interstellar matter; the amount by which it exceeds the normal value for its spectral type is called the color excess. The final table looks similar to table 1-5

Table-1 shows right ascendant and declination of the celestial object along with the color filter data from the SDSS Capturing Device

ra	dec	u	g	r	i
85.9392	0.3754	24.5006	22.3374	20.86	19.2349
85.9711	0.3593	23.9677	22.2227	20.5468	19.2685
85.9729	0.405	23.8119	22.0235	20.2776	19.0042
..	..	..	..	..	..
..	..	..	..	..	..
86.044	0.363	23.6639	21.4424	19.9678	19.2619
86.1064	0.31	23.0465	21.1507	19.6875	18.9339
86.1134	0.3045	24.1607	22.0223	20.3422	19.1197

Table-2 shows the redshift data and petrosian 90% flux data which will be required to deduce the concentration

z	redshift(E-05)	petroR90_u	petroR90_g	petroR90_r	petroR90_i
18.2909	9.5664	1.8148	1.2178	0.9402	1.2109
18.5083	3.3769	0.1356	0.8662	1.1183	1.0464
18.302	3.6198	1.9317	1.1924	1.1011	1.0510
..	..	..	..	..	..
..	..	..	..	..	..
17.8517	7.0037	0.8078	1.0400	1.1566	1.1612
18.8419	15.042	0.4343	1.1319	1.1649	1.1464
18.4065	14.489	0.7915	1.1810	1.2015	1.2244

Table-3 shows petrosian 50% flux data

petroR90_z	petroR50_u	petroR50_g	petroR50_r	petroR50_i	petroR50_z
1.0244	1.2612	0.5307	0.4869	0.5061	0.4581
1.0696	0.0872	0.4670	0.5286	0.4876	0.4585
1.0472	1.4965	0.4727	0.5266	0.5009	0.5028
..	..	..	..	..	..
..	..	..	..	..	..
1.0878	0.6648	0.5111	0.5103	0.5071	0.4762
0.7549	0.3280	0.5136	0.5204	0.4972	0.4082
1.0032	0.6156	0.5082	0.5272	0.5100	0.4803

Table-4 shows the concentration per color filter

conc_u	conc_g	conc_r	conc_i	conc_z
0.695	0.4358	0.5179	0.4183	0.4472
0.6438	0.5392	0.4727	0.4661	0.4287
0.7747	0.3964	0.4782	0.4766	0.4801
..	..	..	..	..
..	..	..	..	..
0.823	0.4915	0.4412	0.4367	0.4377
0.7553	0.4538	0.4468	0.4337	0.5408
0.7778	0.4303	0.4388	0.4166	0.4788

Table-5 shows the color index deduced from the actual color filter data.

u-g	g-r	r-i	i-z
2.1632	1.4774	1.6252	0.944
1.7451	1.6758	1.2783	0.76
1.7885	1.7459	1.2735	0.702
..	..	..	..
..	..	..	..
2.0583	1.7881	1.4297	0.803
2.0708	0.9289	0.451	0.266
2.1385	1.6801	1.2225	0.713

3.5 Splitting the data

Before the development of our model, the dataset is needed to be randomly separated into two parts such that one part is reserved for training and the other is reserved only for testing and verification of the model developed and accuracy measurement. The data is split in the ratio of 75:25, where 75\% is the training data and remaining is used as the testing data.

3.6 Training the Classifier Model and Validation

In this research, five different ensemble models are developed for comparison in accuracy. The respective models are Decision Tree Classifier, Randomized Forest Classifier, ERT Classifier, Adaptive Boost Classifier, and

Gradient Boost Classifier.

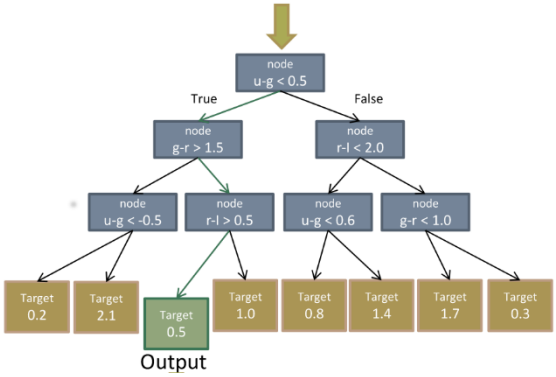

Figure 7 shows the decision tree model fitted over the dataset

A decision tree is a flowchart-like structure as shown in figure 7 in which each internal node represents a test on a feature, each branch represents the outcome of the test, and each leaf node represents a decision taken after going through all features known as label. The paths from root to leaf represent classification rules.

The Random forest is based on measuring the normal of various decision tree results. Random forest is trained over different parts of the same training set such that the overall variance is reduced[15]. The overall performance is boosted, however, this also result in a small increase in the bias and some loss of interpretability. Forests could be understood as pulling together all decision tree algorithm efforts, and thus improving the performance of a single random tree generating a K-fold cross validation effect.

The Extremely Randomized Trees (ERT) Classifier is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a forest to output its classification result. It has conceptual similarity to a Random Forest Classifier yet it differs in the manner of construction of the decision trees in the forest.

Boosting reduces error mainly by reducing bias and to some extent variance, by aggregating the output from many models. The Adaptive Boost Classifier is used to boost the performance of decision trees on binary classification problems. These Weak learning models are sequentially trained using the weighted training data. The process continues until a pre-set number of weak learners have been created or no further improvement can be made on the training dataset. To make predictions with Adaptive Boost, calculation is done over the weighted average of the weak classifiers.

Gradient boosting classifiers is the Adaptive Boosting method combined with weighted minimization, after which the classifiers and weighted inputs are recalculated. The objective of Gradient Boosting classifiers is to minimize the loss, or the difference between the actual class value of the training example and the predicted class value. It isn't required to understand the process for reducing the classifier's loss, but it operates similarly to gradient descent in a neural network.

Getting the Accuracy

The models so created are fitted over the thorough training data set and the model is executed for training. The respective trained models are then verified for the accuracy over the testing data set. The accuracy measure is the fraction of objects that are correctly classified. That is

$$accuracy = \frac{\text{number of correct predictions}}{\text{total predictions}}$$

$$accuracy = \frac{\Sigma_{i=1}^{n} \text{predicted}_i = \text{actual}_i}{n}$$

The accuracy measure is often called the model score. While the way of calculating the score can vary depending on the model, the accuracy is the most common for classification problems.

3.7 Confusion Matrix

In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix[32], is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one. Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class, or vice versa. The name stems from the fact that it makes it easy to see whether the system is confusing two classes. It is used to evaluate the performance of a classification model through the calculation of performance metrics like accuracy, precision, recall, and F1-score.

4. Result Metrics

The accuracy of each model, fixing internal parameters of the model is shown below. The results clearly shows that random forest methodology has produced the best results

with an accuracy of 98.4%, but it took a while for random forest to produce the best accuracy result. However, the Extremely Randomized Tree Classifier produced the second best accuracy of 97.96% within the shortest time span. The gradient boosting method is able to obtained 97.73% accuracy, however, it took the longest time for the algorithm to obtain such accuracy. The following confusion matrix in figure 8-12shows how well the model is trained and the calculated Accuracy, Sensitivity, Specificity and F1 Score is shown in table 6-10.

Model-1: Decision Tree Classifier

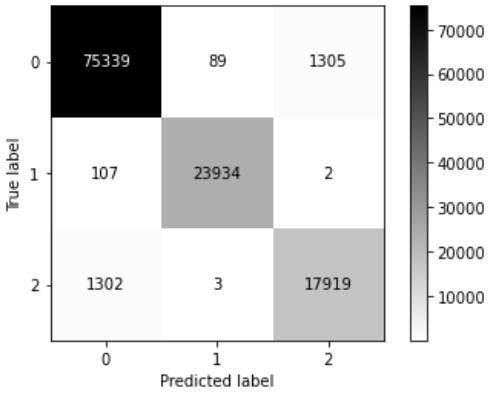

Figure 8 shows the confusion matrix of the Decision Tree Classifier Model

Model / Parameters	DTC
Accuracy (%)	97.6275
Time Taken (sec)	27.7303
Error Rate (%)	2.3725
Sensitivity (%)	97.6275
F1 Score (%)	97.6276

Table 6 shows the Accuracy, Sensitivity and F1 score along with the Error Rate and the Time taken by the model

Model-2: Random Forest Classifier

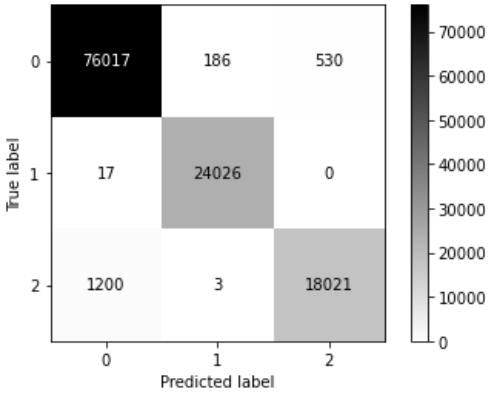

Figure 9 shows the confusion matrix of the Random Forest Classifier Model

Model / Parameters	RFC
Accuracy (%)	98.4050
Time Taken (sec)	53.1260
Error Rate (%)	1.5950
Sensitivity (%)	98.4050
F1 Score (%)	98.3951

Table 7 shows the Accuracy, Sensitivity and F1 score along with the Error Rate and the Time taken by the model

Model-3: Extra Randomized Tree Classifier Model

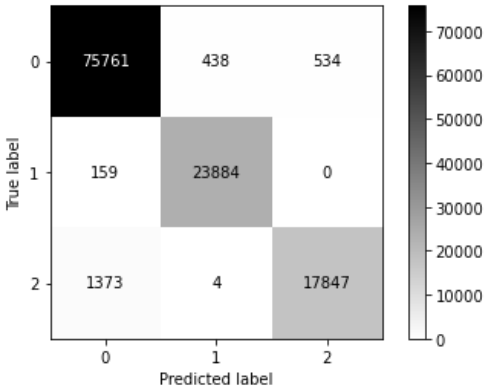

Figure 10 shows the confusion matrix of the Extra Randomized Tree Classifier Model

Model / Parameters	ERTC
Accuracy (%)	97.9633
Time Taken (sec)	8.6176
Error Rate (%)	2.0367
Sensitivity (%)	97.9633
F1 Score (%)	97.9517

Table 8 shows the Accuracy, Sensitivity and F1 score along with the Error Rate and the Time taken by the model

Model-4: Adaptive Boost Classifier

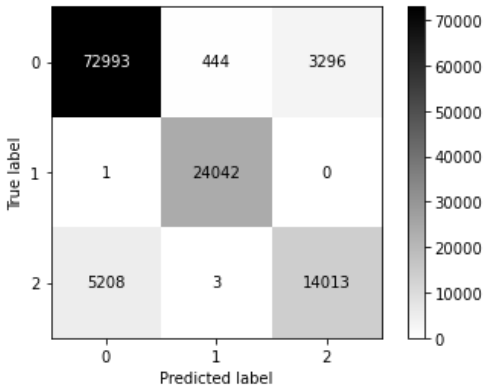

Figure 11 shows the confusion matrix of the Adaptive Boost Classifier Model

Parameters \ Model	ABC
Accuracy (%)	92.5400
Time Taken (sec)	23.4950
Error Rate (%)	7.4600
Sensitivity (%)	92.5400
F1 Score (%)	92.3919

Table 9 shows the Accuracy, Sensitivity and F1 score along with the Error Rate and the Time taken by the model

Model-5: Gradient Boost Classifier

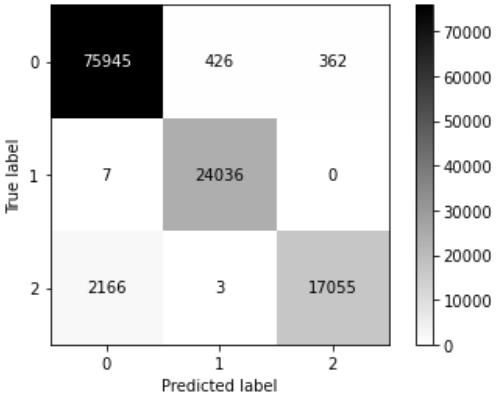

Figure 12 shows the confusion matrix of the Gradient Boost Classifier Model

Parameters \ Model	GBC
Accuracy (%)	**97.7300**
Time Taken (sec)	**123.3292**
Error Rate (%)	**2.2700**
Sensitivity (%)	**97.7300**
F1 Score (%)	**97.5906**

Table 10 shows the Accuracy, Sensitivity and F1 score along with the Error Rate and the Time taken by the model

5. Conclusion

The results clearly show that random forest classifier model has produced the best results with an accuracy of 98.405%, but it took a while for random forest to produce the best accuracy result. However, the Extremely Randomized Tree Classifier produced the second best accuracy of 97.963% within the shortest time span of just 8 seconds. The gradient boosting method is able to obtained 97.73\% accuracy, however, it took the longest time for the algorithm to obtain such accuracy.

Comparing this research with previously done work by Wierzbiński(2021)[18] who had done a similar research but is limited with only 10,000 sets of data and comparatively low feature selections generating accuracy for models, giving them a drawback of higher computational complexity, this research produced a much comparable result with low comparable complexity due to higher features and fixing internal parameters for all models. The result showcased that even for big-data the Random Forest Algorithm works significantly great. Boosting Methods falls few steps backward, but could further be improvised in learning rate to deduce better results. In future, this research could be done using the Quantum Machine Learning algorithm to reduce the time consumption of training of the data for the model.

References

[1]. Britannica, "A. J. Cannon", URL:https://www.britannica.com/biography/, April 2021

[2]. E. Bertin, S. Arnouts, "SExtractor: Software for source extraction", Astronomy andAstrophysics Supplement, vol.117, pp.393-404, 1996

[3]. R. Akeson, "The NASA Exoplanet Archive", California Institute of Technology,NASA Exoplanet Science Institute, DOI:10.1051/aas:1996164,2016

[4]. E. Tozer, "Uncle Sam's New Wonder Weapon This Week", pp. 10, November1958

[5]. R. S. Estey, "Infrared: New Uses for an Old Technique Missiles and Rockets", June1958.

[6]. R. Polikar, "Ensemble Learning", Springer, Boston, MA. DOI:10.1007/978-1-4419-9326-71, ISBN:978-1-4419-9325-0, 2012

[7]. S. H. Dole, “Visual Detection of Light Sources on or Near the Moon”, The RANDCorp Research Memorandum RM-1900, May 1957

[8]. J. T. Mengel, “Tracking the Earth Satellite and Data Transmission by Radio”, Proceedings of the Institute of Radio Engineers, vol. 44, no. 6, pp. 755, June 1956

[9]. W. F. Sampson, H. L. Richter, S. Robertson, “Microlock: A Minimum Weight Radio Instrumentation System for Satellite”, Progress Report No. 20-308, Departmentof the Army, Ordnance Corps, Jet Propulsion Laboratory, California Institute ofTechnology, November 1956

[10]. B. V. Dasarathy, B. V. Sheela, “Composite classifier system design: conceptsand methodology”, Proceedings of the IEEE, vol. 67, no. 5, pp. 708–713, 1979

[11]. R. E. Schapire, “The strength of weak learnability, Machine Learning”, vol. 5, no. 2,pp. 197–227, June 1990

[12]. Y. Freund, R. E. Schapire, “Decisiontheoretic generalization of online learningand an application to boosting”, Journal of Computer and System Sciences, vol. 55,no. 1, pp. 119–139, 1997

[13]. G. James, D. Witten, T. Hastie and R. Tibshirani, “An Introduction to Statistical Learning”, ISBN:978-1-4614-7137-0, DOI:10.1007/978-1-4614-7138-7, pp. 187, 2013

[14]. SDSS Filter and Spectrum, URL:http://abyss.uoregon.edu/ js/glossary/colorindex.html

[15]. T. Hastie, R. Tibshirani, J. Friedman, “The Elements of Statistical Learning (2nd ed.)”. Springer, ISBN:0-387-95284-5, 2008

[16]. J. E. Gunn, W. A. Siegmund, et al. “The 2.5 m Telescope ofthe Sloan Digital Sky Survey”, The Astronomical Journal,vol. 131 no. 4, pp. 2332–2359, arXiv:astro-ph/0602326, DOI:10.1086/500975.17, April 2006

[17]. R. Ahumada, C. A. Prieto, et al., “The Sixteenth Data Release of the Sloan Digital Sky Surveys:First Release from the APOGEE-2 Southern Survey and Full Release of eBOSSSpectra”, arXiv:1912.02905, 2019

[18]. M. Wierzbinski, P. P Lawiak, M. Hammad and U. R. Acharya, “Developmentof accurate classification of heavenly bodies using novel machine learning techniques”, Soft Computing, DOI:10.1007/s00500-021-05687-4, May 2021

[19]. F. D. Albareti, C. A. Prieto, et al., “DR11-12”, SDSS, vol. 219, no. 12, 2015

[20]. F. D. Albareti, C. A. Prieto, et al., “DR13”, SDSS, vol. 233, no. 25, 2017

[21]. M. R. Blanton, M. A. Bershady, et al.,“SDSS-IV Overview”, SDSS, vol. 154, no. 28,2017

[22]. K. Bundy, M. A. Bershady, et al., “MaNGA Overview”, SDSS, vol. 798, no. 7, 2015

[23]. K. S. Dawson, D. J. Schlegel, et al., “BOSS Overview”, SDSS, vol. 145, no. 10, 2013

[24]. M. Doi, M.Tanaka, et al., “SDSS photometry”, SDSS, vol. 139, no. 1628, 2010

[25]. J. E. Gunn, W. A. Siegmund, et al., “APO 2.5m Telescope paper”,SDSS, vol. 131, no. 2332, 2006

[26]. S. R. Majewski, R. P. Schiavon, et al., “APOGEE-1 andAPOGEE-2 overview”, SDSS, vol. 154, no. 94, 2017

[27]. S. A. Smee, J. E. Gunn, et al., “SDSS and BOSS spectrographs”, SDSS, vol. 146, no. 32, 2013

[28]. C. Stoughton, R. H. Lupton, et al., “EDR”, SDSS, vol. 123, no. 485, 2002

[29]. M. A. Strauss, D. H. Weinberg, et al., “SDSS Main Galaxy SampleTargetting”, SDSS, vol. 124, no. 1810, 2002

[30]. J. C. Wilson, F. Hearty, et al., “APOGEE instrument”,SDSS, vol. 7735, no. 77351C, 2010

[31]. B. Yanny, C. Rockosi, et al., “SEGUE overview”, SDSS, vol. 137, pp. 4377-4399,2009

[32]. S. V.Stehman, “Selecting and interpreting measures of thematic classification accuracy”, Remote Sensing of Environment, vol. 62. no. 1, pp. 77–89,DOI:10.1016/S0034-4257(97)00083-7, 1997