Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Supervised Machine Learning Algorithm for Stellar Classification

Dr. Kavitha Subramani[1], Saranya. R[1], Pushpavalli K[3], G B Renuka[3], Mr. S. Subburaj[4]

[1]Department of Computer Science and Engineering, Panimalar Engineering College, Chennai, India. kavitha.pec2022@gmail.com, saranyar645@gmail.com
[2]Assistant Professor, Jerusalem College of Engineering, Chennai, India, pushpavalli.k@gmail.com
[3]Computer science and engineering, Madanapalle Institute of Technology & Science , Madanapalle, Andra Pradesh ,India, renukagollabala@gmail.com
[4]Department of mathematics, Amrita vishwa vidyapeetham, Coimbatore, India, subburajs87@gmail.com

Astronomical object such as stellar,quasar,galaxy are very important key for the study of universe and galaxy. We all know stars emit light likewise quasar and galaxy do. The light from these astronomical object has a kind of radiation know as electromagnetic radiation. when we split that electromagnetic radiation we get spectrum.Spectrum is defined as the lights of seven colors and the spectrum is used to identify the chemical composition and temperature of each star.Each light ray indicates a particular chemical element or molecule. Due to the amount of chemical element present in the each light ray the temperature of each light ray vary. We will be able to get that temperature using Sloan Digital Sky Survey (sdss) telescope which is located in mexico. As these spectral characteristics contains significant information about the astronomical object which is very much useful for making better classification of object. For processing enormous amounts of data, data mining is a common technique.Various supervised machine learning algorithm like naive bayes,random forest, decision tree and multi layer perceptron is used and the results are compared with each other. Random forest has vast advantage like by averaging a number of decision trees, a Random Forest diminishes overfitting and is less vulnerable to noise and outliers in the data. Hence accuracy percentage in random forest is high as compared with other existing algorithms.

Keywords: stellar spectra, astronomical objects,machine learning, multi layer perceptron.

1. Introduction

Space exploration greatly benefits from machine learning. Accurately classifying astronomical objects manually takes a lot of time and is highly challenging. Data science, a rapidly expanding area, is fundamental to machine learning. Algorithms are instructed in data mining activities to create classifications or predictions using statistical methods. The suggested method classifies astronomical objects using supervised machine learning algorithms.

Algorithms are taught by supervised learning from tagged data. After gaining an understanding of the data, the algorithm decides which label fresh data should receive based on patterns and associations with unlabeled new data.Stars are the building of galaxies; they are widely spread across the sky and easily recognised. Stars are responsible for the formation of chemical substances like carbon, nitrogen, oxygen. Stars emit a spectrum of light .Spectrum light is helpful in finding the temperature each object has a unique temperature and spectral characteristics with which we will be able to classify astronomical objects. stellar spectra are obtained using a spectroscope. Data science, a rapidly expanding area, is fundamental to machine learning. Algorithms are instructed in data mining activities to generate classifications or predictions using statistical methods.The resulting spectrum is focussed by another lens and can either be viewed directly or photographed.

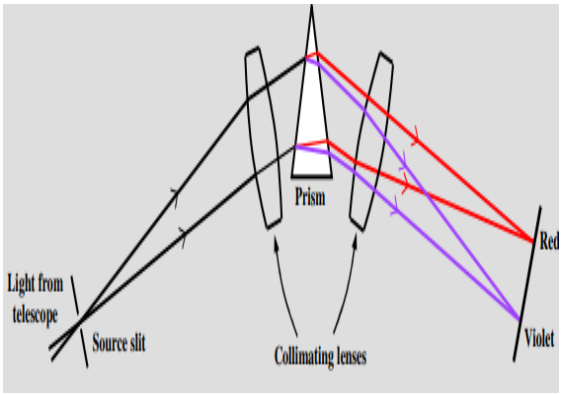

Fig 1: Dispersion of white light

A quasar is a tremendously bright astronomical object that is fueled by gas spiralling rapidly into a massive black hole in the centre of some galaxies. The metric expansion of space causes any light from quasars, which are far away but incredibly bright objects, to be redshifted when it reaches Earth. An enormous cluster of gas, dust, and billions of stars and their solar systems is known as a galaxy. A galaxy is held together by gravity.Four machine learning classification approaches are used to differentiate between star, quasar, and galaxy. The aim of the proposed method is to predict whether the received light rays from the space are stars or other. This can be predicted by a machine learning method. The process starts from collecting the data. After collecting the dataset, it is pre-processed for removing the unwanted data from the dataset. The application of machine learning is now widespread across all departments since it lowers error rates. The best method is employed when guessing whether light rays are coming from stars or anything else. The objective is to propose a machine learning model for anticipating Stellar classification using supervised machine learning classification models by forecasting outcomes in the form of best truthfulness by differentiate supervised algorithms.

2. RELATED WORKS:

Object explore using low-S/N stellar spectrum in SDSS was proposed by Minglei MU Et al [1]. drawn dwarfs, carbon lines, magnetic white dwarfs can all be detected. It will be competent. Using stacking ensemble learning with SDSS DR 7 and the 10-fold nested cross-

validation approach, LI Chao suggested the concept of classifying stars and galaxies in 2020. As a result, the ensemble model's training duration will be significantly extended [2]. As a result, investigations in the future will try to use a distributed strategy for training. A practical overview of supervised machine learning in astronomy was provided by Dalya Baron. It covered the selection and pre-processing of the input dataset, evaluation techniques, and three well-known supervised learning algorithms. Cluster analysis, dimensionality reduction, visualisation, and outlier detection are performed using unsupervised machine learning methods. Galaxy Image Classification is a concept that Manuel Jimenez et al. discovered in 2020. They build the FE of galaxy images using autoencoder architectures based on citizen science data, and they also employ the CNN suggested for comparing ML methods for galaxy image classification[4]. In order to find an unsupervised star, galaxy, and quasar categorization function, C. H. A. Logan and S. Fotopoulou employed hierarchical density-based spatial clustering of applications with noise (HDBSCAN). To do this, the scientist performed three different HDBSCAN run, each of which picked a distinct object group, then optimised the input characteristics and hyperparameters of each run, using the result of each run's optimization as a binary classifier[5]. Twin hypersphere model for categorising stellar spectra urbanized by Zhongbao Liu. The pair of hyperspheres created by THM has greater anti-noise capabilities in comparison to the hyperplane-based classification algorithms SVM and TWSVM, which are sensitive to outliers close to the hyperplane[6].The classification of stellar spectra using SVM by Zhong-bao Liu and Fang-xiao Zhou was based on within-class scatter and between-class scatter, ignoring internal training dataset information such within-class structure and between-class structure. Finding the best hyperplane to divide two classes is the goal of WBS-SVM[7]. Makhija, M. Das, S. Saha, and S. Basak Using photometric data, a machine learning study is conducted to distinguish stars from quasars. The issue of classifying matching components in the GALEX and SDSS (Sloan Digital Sky Survey) inventories into stars and quasars based on color-color plots can be resolved using machine learning. [8] An Imbalance learning method for classifying variable stars was proposed by ZafiirahHosenie et al. It is challenging to correctly computerize the classification of variable stars into their corresponding sub-types. Machine learning-based systems frequently experience the imbalanced learning problem, which leads to subpar generalisation performance, especially for rare variable star sub-types[9]. A multitask residual neural network is proposed in a study by Yuxiang Lu, Jingchang Pan, and Zhenping Yi titled "Study on Stellar Spectra Classification Based on Multitask Residual Neural Network" (2020). This network uses the relationship between the brightness class and phantom subtype of stars to process two tasks simultaneously [10].Quasar Detection by Aniruddh by means of Linear Support Vector Machine and Learning From Mistakes HerleEtal employs A Quasar prediction with a significantly lower FNR is produced by training a Linear SVM using Sloan Digital Sky Survey (SDSS) Data and cascade it with an Ensemble Bagged Tree Algorithm. Additionally, FPR can be decreased, which is a possibility that can be investigated in future research[11]. Asad Muhammad Usman, Mansoor Khan A Machine Learning Technique to Classify LSST was developed by AkramEtal. Astronomical Observations Based on Photometric Information As previously established, there are six distinct bands that represent the light curves: u, g, r, I z, and y. Based on the light's wavelengths, these bands are distinguished from one another. Methodology is adaptable since it draws features from photometric light curves[12]

Rajdeep Banerjee and Jaswinder Singh suggested A learning on solo and Multi-layer

Perceptron Neural set-up in 2019. In terms of artificial neural networks, the perceptron model is the most fundamental. The multiple improvements have been significantly impacted, and the field of neural networks has been greatly boosted. It has been clear from the start that it is the secret to how machines perceive, and via intensive training processes, it has been possible to generate intelligence[13] machines.Additional two-dimensional spectral facts can be supplementary to the method, as suggested by Yakun Lu Etal in his proposal for a stellar spectral classification using a 2D spectrum and fully connected neural network. The two-dimensional spectral image's quality can also be increased with the help of the preprocessing operation[14].A astrophysical spectral classification and characteristic assessment based on a random forest was proposed by Xiang-Ru Li Etal. A spectrum is first normalised using a 17th order polynomial fitting, and then. Research on four stellar spectrum libraries demonstrates that the RF performs well in terms of categorization. The evaluation of spectral features based on RF was also examined in this work[15].Yosry A. Azzam created an artificial neural network for predicting the fundamental characteristics of the atmosphere from star spectra. To automatically classify stellar spectra, create an Artificial Neural Network (ANN) method. To extract the essential parameters, the method has been used. In order to better understand if the current algorithm is accurate, the forecasted atmosphere attributes for the two samples were compared. It was found that they are in excellent accordance for about 50% of the samples[16]. The idea for the paper Machine Learning: Techniques, Real-World Applications, and Future Research was put forth by Iqbal H. Sarker. explains the foundations of various machine learning techniques and how they can be applied in a range of real-world application domains, such as defense systems, smart buildings, medical, e-commerce, farmers, and many more[17].The use of logistic regression for star classification with an application to color-color diagrams was suggested by Leire Betia-Antero and Javier Yanez. examined a different supervised classification method, Logistic Regression, which is little used in astronomy, in order to obtain membership probabilities for possible T Tauri star candidates from UV-IR color-color diagrams[18].

3. METHODS

In recent years much research is going on astronomical data. It is very hard to classify celestial objects manually.It is very difficult to make highly accurate and efficient results from such a huge and complex data set. So an automatic classification mode is built.Spectral features from the Sloan Digital Sky Survey dataset is used to train the model (SDSS). The information is made up of 100,000 samples of stars with 17 attributes. The green,ultraviolet,red and Infrared filter bands, as well as the alpha, Delta, and redshift of stars, are indeed the key features concentrated. The data set must then be cleaned using pre-processing methods. To transform unprocessed data into a format that is both useful and effective, data pre-processing is performed.In the proposed system data preprocessing techniques like find the missing values, duplicate values, find the unique value of a data frame, deleting a particular column are used. To handle an imbalance in the data set which leads to oversampling to overcome this SMOTE technique is used. SMOTE is an oversampling technique in which the minority class is provided with synthetic samples. This approach helps resolve the overfitting problem.Suppose a model build with imbalanced data then the accuracy of the model will not be more accurate. It is important to equilibrate the data before modelling.

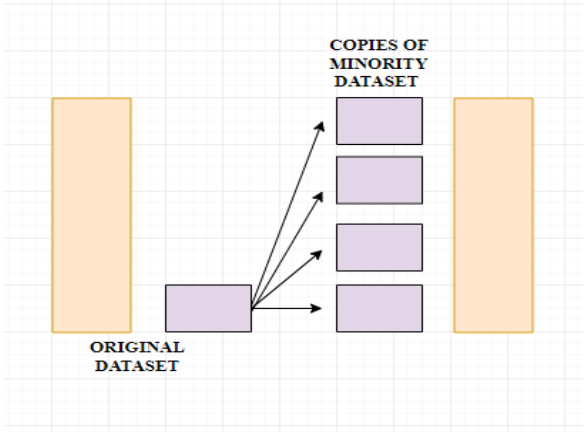

Fig 2 : Synthetic Minority Oversampling Technique

One of the major machine learning approaches, feature selection is applied to take out features which edge the integer of input columns and effectively remove noise.Feature extraction techniques are used to improve the accuracy of the model.We use feature selection techniques likeUni-variate, bi-variate and multivariate analysis. Data visualisation is employed and is an essential skill in contemporary machine learning and data mining. Actually, the focus of statistics is mostly on quantitative estimates and data descriptions. A vital set of tools are provided by data visualisation for obtaining a qualitative insight. This might be helpful when exploring and understanding about a dataset to identify patterns, corrupt data, outliers, and much more.

The data will then be divided into a learning set of data and test set of data set. The supervised learning model is developed using the train dataset. In order to assess the model, the experimental data set is used and to measure performance. Machine learning models are used to make classification on test data based on the trained data set .Multiple machine learning algorithms are used for classification of star, galaxy, quasar.

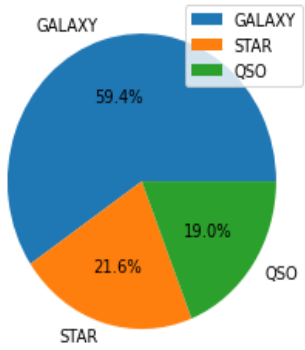

Fig 3: Proportion of different class

Then we find the accuracy for each model.

Model with the best accuracy is used for the deployment. a simple UI is created where the user

input the data as an output it will predict the correct astronomical object.

A. System Architecture

The proposed method aims to determine if the received light rays from space are coming from a star or anything else. A machine learning approach can predict this

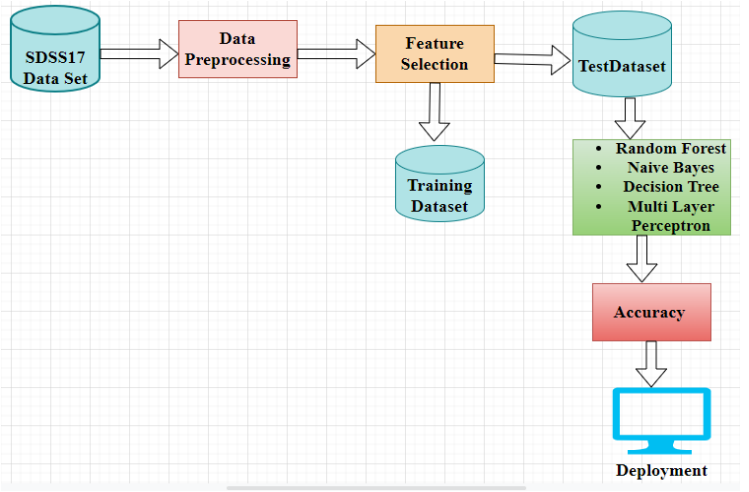

Fig 4: System architecture

Data collection is the first step in the process. The dataset is pre-processed after collection to get rid of any unnecessary data. In all departments where it decreases error, machine learning is currently applied extensively. For identifying light beams as coming from a star or another object, the best method is used among many others. For deployment, a model with great precision is used.

B. Supervised Learning

Simply because it employs annotated datasets to coach algorithms that precisely identify data or make predictions, supervised machine learning stands out. The weights of the model are improved as input data is fed into it, and this process continues until the model is built-in, which takes place during the cross confirmation phase. Businesses can build sound solutions to a range of practical issues with the aid of supervised learning.

Supervised learning in machine learning enables you to gather data or generate output from previous experiences. helps in boosting performance criteria via using experience. By supervised machine learning, you can deal with a variety of compute problems that arise in real-world situations.

C. Naïve Bayes:

The Class Conditional Independence hypothesize of the Bayes Theorem is followed by the Naive Bayes classification method. When one attribute is present, the likelihood of a specific event does not change, proving that each forecaster has an equal impact on the result. The three different types of Naive Bayes classifiers are multinomial, Bernoulli, and Gaussian. This method is often used by manuscript categorization, spam recognition, and suggestion systems.

D. Decision Tree:

As a non-parametric supervised learning method, decision trees are utilised in classification and regression applications. A origin node, branches, interior nodes, and leaf nodes make up its hierarchically organised structure.First node in decision tree is known as the root node, has zero extending branches. The inner nodes, also known as conclusion nodes, are nourished by the root node's egressing branches. mutual node type perform evaluations based on the known attributes to produce homogenous subsets, which are represented by leaf nodes or terminal nodes. All of the results are represented by the leaf nodes in the dataset.

E. Multilayer Perceptron:

A multi-layer neural network with fully connected layers is referred to as a "multilayer perceptron" (MLP). Tuning is required for a neural network's hyperparameters, including the number of layers and neurons. Cross-validation methods must be used to choose the most appropriate values for these. Backpropagation is used to spread training for weight correction. Deeper neural networks do better at data processing.

The following is the MLP learning process:

Data should be pass on from the enter level to the output level in a forward trend. This is the stage of forward propagation.

Analyze the outcomes and identify the error The error needs to be minimised.

Correct your errors. By calculating the model's derivative with respect to each network weight, the model can be updated.

The three procedures mentioned above should be repeated over a number of epochs to learn the ideal weights.

To extract the data and generate the anticipated class labels, a threshold function is then employed.

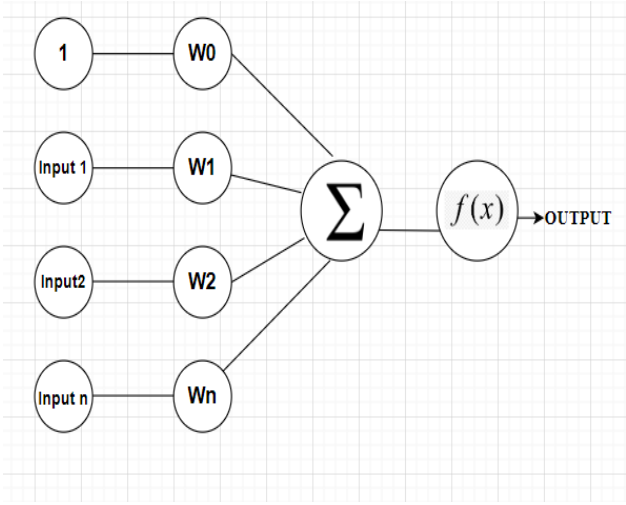

Fig 5: Multilayer perceptron architecture

F. Random Forest

Both classification and regression are implemented using random forest, a robust supervised machine learning method.The term "forest" refers to a collection of discrete decision trees that are joined to diminish variation and produce more accurate data forecasts.

Random forest classification strikes the result using an ensemble process. Different decision trees were constructed utilising the training data. This dataset is composed of up of findings and characteristics that will be chosen at random when the nodes are being split apart.Divergent decision trees are used in a random forest system. Each decision tree has a core node, leaf nodes, and assessment nodes. The result generated by a scrupulous decision tree is characterized by the leaf node of each tree. The majority voting method is used to choose the final product.

4. RESULT

The next stage is to determine the model's efficacy based on some metric using test datasets after executing the model and receiving output in the form of a probability or a class. Machine learning algorithms are compared using a variety of performance criteria.machine learning model's evaluation metrics are very important. How machine learning algorithm performed is rated and compared depends on the metrics employed.The confusion matrix is one of the clearest and most straightforward metrics for assessing the precision and correctness of the model. For classification issues where there are two or more potential class outputs, a confusion matrix is used. Precision, recall, and f1 score can be determined amid the support of the confusion matrix.

MODEL	ACCURACY
Naive bayes	76.31
Decision tree	96.12
Multilayer Perceptron	96.35
Random forest	97.71

Table 1: models wih accuracy

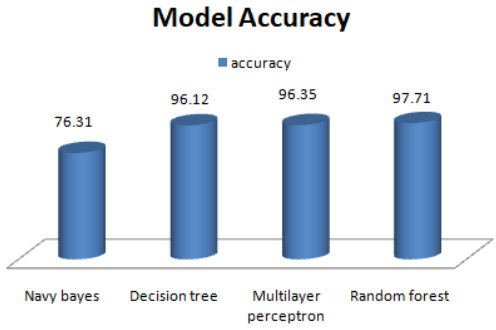

Fig 6: Model Accuracy

From the above diagrams we can conclude that random forest algorithm has the highest accuracy.hence random forest is best suited for the classification problem

5. DISCUSSION

SMOTE technique is used which helped o remove the imbalance in the dataset and the made the model an efficient model.

the below table depicts the performance metric graph of naive bayes algorithm while classifying galaxy we get a pricision score of 0.73, recall 0.96 and f1-score 0.83.very low percentage of 0.26 in F1-score.therefore the accuracy of naive bayes is very low

Classification report of naive Bayes:

	precision	recall	f1-score	support
0	0.73	0.96	0.83	11889
1	0.86	0.85	0.86	3792
2	0.98	0.15	0.26	4319
accuracy			0.76	20000
macro avg	0.86	0.65	0.65	20000
weighted avg	0.81	0.76	0.71	20000

Fig 7: Performance metric for naive bayes algorithm

Figure 8 shows the performance metric graph using multilayer percrptron.the precision ,recall and f1 score while classifying quassar is very low.while the classification of galaxy and star results an average accuracy.multi layer perceptron is not best suited for the classification

Classification of Multilayer perceptron report :

	precision	recall	f1-score	support
0	0.95	0.98	0.97	11889
1	0.97	0.84	0.90	3792
2	0.97	1.00	0.98	4319
accuracy			0.96	20000
macro avg	0.96	0.94	0.95	20000
weighted avg	0.96	0.96	0.96	20000

Fig 8: Performance metric of multilayer perceptron

Classification report of decision tree :

	precision	recall	f1-score	support
0	0.97	0.97	0.97	11889
1	0.91	0.91	0.91	3792
2	1.00	1.00	1.00	4319
accuracy			0.96	20000
macro avg	0.96	0.96	0.96	20000
weighted avg	0.96	0.96	0.96	20000

Fig 9: Performance metric of decision tree

The fig 10 shows the performance metrics of random forest algorithm in random forest we are able to get high acccuracy for the classification because of its majority voting method

Classification report of Random forest:

	precision	recall	f1-score	support
0	0.98	0.99	0.98	11889
1	0.96	0.92	0.94	3792
2	1.00	1.00	1.00	4319
accuracy			0.98	20000
macro avg	0.98	0.97	0.97	20000
weighted avg	0.98	0.98	0.98	20000

Fig 10: Performance metric of random forest algorithm

6. CONCLUSION

The analytical procedure included the preparation and processing of the data, blank analysis, exploratory analysis, and eventually the construction and assessment of the model. Data is made efficient and valuable by following a good data cleaning process. The accuracy of the model is increased by using the right pre-processing and feature extraction techniques. As a result, the model undertakes effective training.Able to overcome issues with flocking by using this, such as to replace the time-consuming, traditional method of manually classifying celestial objects. To swiftly and accurately discriminate between several heavenly entities using enormous spectral data.

References

1. Wu, Minglei, et al. "Rare Object Search From Low-S/N Stellar Spectra in SDSS." IEEE Access 8 (2020): 66475-66488.
2. Chao, L. I., et al. "Research on star/galaxy classification based on stacking ensemble learning." Chinese Astronomy and Astrophysics 44.3 (2020): 345-355.
3. Baron, Dalya. "Machine learning in astronomy: A practical overview." arXiv preprint arXiv:1904.07248 (2019).
4. Jiménez, Manuel, et al. "Galaxy image classification based on citizen science data: A comparative study." IEEE Access 8 (2020): 47232-47246.
5. Logan, C. H. A., and Sotiria Fotopoulou. "Unsupervised star, galaxy, QSO classification-Application of HDBSCAN." Astronomy & Astrophysics 633 (2020): A154.
6. Liu, Zhongbao. "Stellar spectra classification with twin hypersphere models." New Astronomy 88 (2021): 101613.
7. Liu, Zhong-bao, et al. "Classification of stellar spectra with SVM based on within-class scatter and between-class scatter." Astrophysics and Space Science 363.7 (2018): 1-6.
8. Makhija, Simran, et al. "Separating stars from quasars: Machine learning investigation using photometric data." Astronomy and Computing 29 (2019): 100313.
9. Hosenie, Zafiirah, et al. "Imbalance learning for variable star classification." Monthly Notices of the Royal Astronomical Society 493.4 (2020): 6050-6059.
10. Lu, Yuxiang, Jingchang Pan, and Zhenping Yi. "Study on Stellar Spectra Classification Based on Multitask Residual Neural Network." 2020 Prognostics and Health Management Conference (PHM-Besançon). IEEE, 2020.
11. Herle, Aniruddh, Janamejaya Channegowda, and Dinakar Prabhu. "Quasar Detection using Linear Support Vector Machine with Learning From Mistakes Methodology." 2020 IEEE

International Conference on Electronics, Computing and Communication Technologies (CONNECT). IEEE, 2020.
12. Khan, Asad Mansoor, et al. "A Machine Learning Technique to Classify LSST Observed Astronomical Objects Based on Photometric Data." 2019 6th Swiss Conference on Data Science (SDS). IEEE, 2019.
13. Singh, Jaswinder, and Rajdeep Banerjee. "A study on single and multi-layer perceptron neural networks." 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2019.
14. Lu, Yakun, et al. "Stellar Spectral Classification with 2D Spectrum and Fully Connected Neural Network." Journal of Physics: Conference Series. Vol. 1626. No. 1. IOP Publishing, 2020.
15. Li, Xiang-Ru, Yang-Tao Lin, and Kai-Bin Qiu. "Stellar spectral classification and feature evaluation based on a random forest." Research in Astronomy and Astrophysics 19.8 (2019): 111.
16. Azzam, Yosry A., M. I. Nouh, and A. A. Shaker. "Prediction of the atmospheric fundamental parameters from stellar spectra using artificial neural networks." NRIAG Journal of Astronomy and Geophysics 10.1 (2021): 23-34.
17. Sarker, Iqbal H. "Machine learning: Algorithms, real-world applications and research directions." SN Computer Science 2.3 (2021): 1-21.
18. Beitia-Antero, Leire, Javier Yáñez, and Ana I. Gómez de Castro. "On the use of logistic regression for stellar classification." Experimental Astronomy 45.3 (2018): 379-395.