CLASSIFICATION OF THE ASTEROID BY DIFFERERNT MACHINE LEARNING TECHNIQUES

SUB- ITA 5007- DATA MINING AND BUSINESS INTELLIGENCE

ANISH SI **PANKAJ MALLICK** **SOUMITRA DAS**
19MCA0199 **19MCA0223** **19MCA0233**
Vellore Institute of Technology, Vellore, Tamil Nadu

***Abstract*- The purpose of this study is to identify the Near-Earth Asteroids also to predict the new asteroid as the Near Earth or not. Basically, there are four asteroid families is known as Near Earth asteroids. Our motto is to classify those asteroids by various machine learning and deep learning techniques. We will try to implement those models by extracting features by doing the PCA analysis of the dataset and compare those models. Later, we build a neural network model to achieve the grates accuracy, where we gain the accuracy over 99.4% by that deep neural network. We can easily say that this accuracy is far better from those best machine learning model random forest, gradient boosting etc. which gave over 95% accuracy. Our main motto of this project is to classify the asteroids by the PCA analysis and various techniques.**

Keywords: asteroids, machine learning, neural network, gradient boosting, PCA

1. *INTRODUCTION*

Asteroids are one of the most important Space objects. It has a great scientific value as they allow to study the early solar system. It has many different classes. Those classes are divided with the help of the principal materials of the asteroids. Sometimes, the classes are divided for some orbit classes. As they rotate through the different orbits, it is obvious that they will be divided into many categories according to that. So, they also have different orbital periods also. But the main fact is all those rotation period, orbital periods depends on the semi-major axis, eccentricity and largely depends on the inclination of the elliptic plane. So, naturally all the classes have different length of axis and also the different eccentricity measure. These axis, eccentricity and inclination curve are some of the basic features of the orbital parameters.

Asteroid classification is very important as some of the asteroids are also the Near-Earth Object(NEO) among them some are also potentially hazardous in nature. So, they may hit earth any times. So, by this classification, we can easily identify whether the asteroid is the near-earth asteroid and that will help to determine the PHA (Potentially Hazardous Asteroid). Basically, there are three asteroid classes. The classification is also needed to obtain larger family with many members, and it also helps to find rotation and acceleration of the members of a particular class.

There are only four asteroid families which are called as a Near Earth Object. Those families are Atiras, Atens, Apollos, Amors. The other families of the asteroids are the Trojan asteroid, main belt asteroids, Mars Crossing Asteroids etc. So, our main motto in this project is to classify those asteroid families and build a model which can predict whether the new found asteroid is near earth asteroid, and also try to predict that is hazardous or not. Now-a-days, NEA (Near

Earth Asteroids) are the major threats to the earth, so it is necessary to have a good algorithm to identify those asteroids that could hit the earth and take pre-cautions.

Recent times, machine learning is a very famous and one of the best tools to classify the data and to build the prediction model. So, we want to use those machine learning techniques to classify the asteroids from a large dataset and predict the new asteroid according to their attributes of classification.

2. ***LITTERATURE SURVEY***

In this chapter[1], the author gives a review of recent developments in our astronomical understanding of small bodies. To keep the discussion manageable, they focus mainly on comets and asteroids that are not in orbit around major planets; i.e., they are excluding those small satellites that are likely captured comets and asteroids. Such satellites can inform and have informed our understanding of the general comet and asteroid population. So, they are restricted to small bodies of asteroids and comets. They are keen to know about the nature of our protoplanetary disk, and how did the process of planetary accretion occur.

In this paper[2], the author has used k-Nearest Neighbor (KNN) in combination with the current Bus-DeMeo (DeMeo, et al. 2009) taxonomic classification schema to test if machine learning can take the place of Principal Component Analysis. Using a dataset of spectrophotometric color indices derived from combined visible and near-infrared (NIR) observations and paired with Bus-DeMeo taxonomic class, I created a training dataset for the model to learn. The results support the visible wavelength region as more diagnostic of spectral slope and the NIR wavelength region as more diagnostic for surface mineralogy. The overall accuracy scores (>80%) of the machine learning test dataset validate the methodology, but fall short of the threshold necessary to replace current methods of classification (>95%). The overall robustness of the Bus-DeMeo taxonomy is corroborated through the relatively similar grouping structure between the C-, S-, and X-complexes in both wavelength ranges, suggesting an overall relationship between slope and qualities present across multiple wavelength regimes. This is possibly due to spectral features being closely tied to surface mineralogy and spectral reddening of the slope believed to be tied to the effects of space weathering.

In this paper[3], the author has evaluated the main techniques and technologies being developed and tested to explore, prospect and harvest near-Earth asteroids estimated to be worth billions of dollars. Further, the two major space-related companies Planetary Resources and Deep Space Industries and their respective approaches to asteroid mining have been reviewed as well. Ultimately, the aspects of financial feasibility and environmental and legal issues have been approached. The author wanted to made to provide the reader with a general understanding of what processes and theories are currently being developed, what methods are possible to use and combine and how sustainability and legal perspectives are applied to the space business in general.

In this paper[4], the author has used Hierarchical Clustering Method (HCM) to cope up with the addition of new members to existing families of asteroids, based on their physical observations and their orbital catalogs. Their work was based on a large catalog of high accuracy synthetic proper elements (available from AstDyS), containing data for >330,000 numbered asteroids. By selecting from the catalog a much smaller number of large asteroids, they first identify a number of core families; to these they attribute the next layer of smaller objects. Then, they removed all the family members from the catalog, and reapply the HCM to the rest. This gives both satellite families which extend the core families and new independent families, consisting mainly of small asteroids. These two cases are discriminated by another step of attribution of new members and by merging intersecting families. This leads to a classification with 128 families and currently 87,095 members. The number of members can be increased automatically with each update of the proper elements catalog; changes in the list of families are not automated.

In this paper[5], the author assumes the near-Earth asteroids are likely targets for resources to support space industrialization, as they appear to be the least expensive source of certain needed raw materials. Furthermore, exploitation of asteroids for precious metals and semiconducting elements is a possible environmentally friendly remedy for impending terrestrial shortages of these resources. This paper discusses the resources available from NEAs, as well as the technical engineering aspects of possible mining project designs, including a survey of mission plans, mining and extraction techniques that may be used. The author is briefly introducing the concept of Net Present Value as the appropriate measure to determine the technical economic feasibility of a hypothetical near earth asteroid (NEA) mining operation.

In this research paper[6], we are going to discuss how the concept of artificial neural network could be utilized to estimate the diameter of an asteroid. In this research, we have used the Multilayer Perceptron algorithm as the base algorithm to predict the diameter. We have used different algorithms to test and evaluate the performance of the model with the same dataset but Multilayer Perceptron algorithm performed best in these type of situations with higher accuracy and least error while prediction. The dataset is we have used is officially maintained by NASA Jet Propulsion Laboratory. In this dataset we have considered all types of asteroids such as asteroids which are grouped as Near-Earth Objects(NEO), Potentially Hazardous Objects(PHA), we have also considered all the possible asteroid orbit classes as mentioned in the official website of JPL(Jet Propulsion Laboratory). The columns of the dataset also contain all the physical and basic properties of an asteroid. We have used Mean Absolute Error, Mean Squared Error, Median Absolute Error, Explained Variance Score and R2-Score as metrics to evaluate and compare the performance of different regression algorithm against the same dataset.

In this paper[7] they have used various machine learning techniques for classify Quasar-star from a given dataset. Data are taken from 6 and 7 (DR6 and DR7) of the Sloan Digital Sky Survey (SDSS) which are primarily distinguishes the stars and the Quasars. Their aim is to investigate the appropriateness of the application by the help of some certain Machine Learning method. For this they need to classify the photometric data. They has been seen that asymmetric AdaBoost method is the most efficient machine learning technique to classify these stars and quasar from the dataset.

In this paper[8], they tried to propose an machine learning technique, to easily solve the principal component analysis. There they selected the KNN algorithm or K-Nearest Neighbor to classify the asteroid using Spectrophometry. They created the dataset by using a dataset of spectrophotometric color indices, which has been derived from combined visible and near-infrared (NIR) observations. After that, they paired that with the Bus-DeMeo taxonomy class. Thus the total training dataset is made. They also achieved the accuracy of 94% on the visible dataset and the 91% on the NIR dataset. This has been done only for the KNN algorithm while fitted on the only spectrophotometric dataset.

In this paper[9], they have proposed the classification of the asteroids by a machine learning technique called random forest. They have taken the data from the observations of the Sloan Digital Sky Survey(SDSS) and Moving Object Catalouge(MOC). They have worked on the 48642 asteroids. With the combination of the four taxonomy of asteroid such as Tholen, Bus etc. and the principal component anlysis asteroids are divided into 8 classes such as C, X, S, B, D, K, L and V. For these they have used the random forest method. This division are done according to their SDSS magnitudes at the wavebands of g, r, I and z. They also reached the higher accuracy compare to the many proposed methods.

In this paper[10], they used supervised learning algorithm to detect asteroid from a large dataset. For this case they have used vetted NEOWISE dataset (E. L. Wright et. Al,2010). According to them, the metrics they have used can be easily done as it can be easily associated with the extracted sources. They also used the python SKLEARN package. After doing this also gave the report on the reliability, feature set selection, and also the suitability of the various machine learning algorithms. At the end they also compares the results.

In this article[11], they presented an machine learning analysis of five level galaxy catalogues from the Galaxy And Mass Assembly better known as GAMA. In this paper, they have used vector quantization and the random forest methods. They faced a problem that neither data from the single catalogues nor the combining of the five catalogues can give the dataset in the support of the inspection based galaxy classification scheme. To overcome this and to implement the nature of the employed visual based classification scheme, they presented the galaxy classification parameters, which are discriminative to achieve the class distinction with respect to physical and morphological features.

In this paper[12], they proposed a new methodology for variable star classification by combining two machine learning algorithm. Those are Self-Organizing Map(SOM) which is a unsupervised algorithm and other one is the Random Forest, a supervised algorithm. They took the data from the K2 mission fields 0-4, finding 183 detached eclipsing binaries, many Scuti pulsator, Doradus pulsator etc. They have also shown the light-curve features for all the K2 stellar targets, which includes three strongest detected frequencies. They have also shown that the graphical representation of the period and frequency emitted by the different targets. They have also achieved the 92% accuracy rate using the training set. They have also stated that their methodology will give more accuracy for the K2 field 3-4.

In this paper[13], they have classified the Near Earth Asteroid(NEA), which can easily pass the orbital of the Mars. They are mainly three types such as Amor, Apollo and Aten. According to the authors, they can be linearly seperable due to focal distance and the semi major axis. In this paper, they have shown that also. So they easily concluded that a perceptron-type artificial neural network is enough to classify those three types of the asteroid. They concluded that above two features are more convenient to classify the NEA as in any plane, they can be linearly separable among those 3 types.

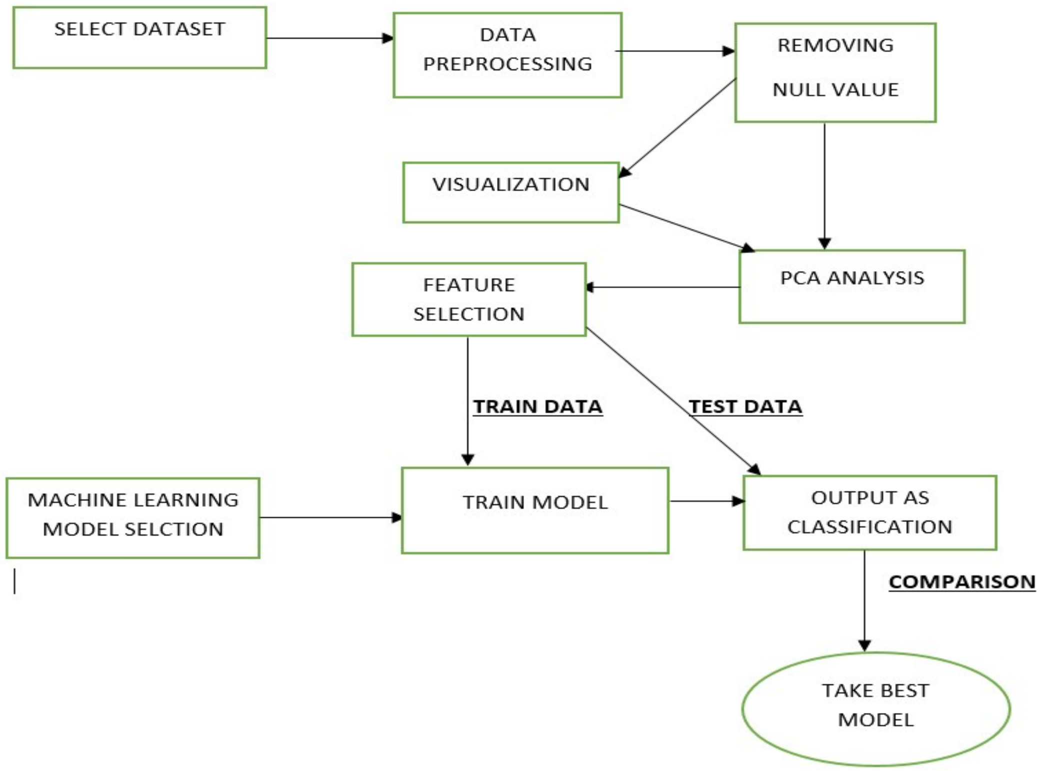
In this paper[14], they categorized the observed targets by machine learning algorithm. They have taken data from the observatory and can easily classify the observed targets among four asteroid taxonomies such as S-, C-, X- or D- type. They also given the light curve data for each Near Earth Asteroids. Of 39 targets, they are able to resolve the complete rotation periods and amplitudes of six observed NEA. They have also shown the graphical representation of the classified data and also find a greater accuracy of 80% and it has been also seen that "stony" types NEA and "non-stony" types NEA has ratio of approximately 1:1.

In this thesis[15], they have proposed a new way of dynamically classification of the asteroids. At first they have summarized the basic characteristics of the Near Earth Asteroids. There they classify the NEAs by the new model which is build by the means of fuzzy sets. They proposed that the one of the main features of the NEA dynamics is their possibility of colliding with earth, mars or Venus. But some of them are not classified among those, as they are not so hazardous. In this paper, they were interested in the values of the collision probability. As the result of this thesis, they concluded that four groups classification called G1, G2, G3, G4 according to the non hazardous, collide with mars, and collide with Venus.

3. ***PROBLEM DESCRIPTION***

Our motto is to classify the near earth asteroids by various machine learning and deep learning techniques. We will try to implement those models by extracting features by doing the PCA analysis of the dataset. Later we compare those results. Here we selected 13 features from the given 31 features. Some of those features are semi-major axis, eccentricity, magnitude slope parameter, aphelion, perihelion distance etc. We use different machine learning model such as support vector machine, random forest, decision tree, gradient boosting. Later, for increasing accuracy, we use neural network model and compare that with the best machine learning model.

4. ***ARCHITECTURAL DIAGRAM***



5. ***METHODOLOGY***

A. ***DATA COLLECTION:***

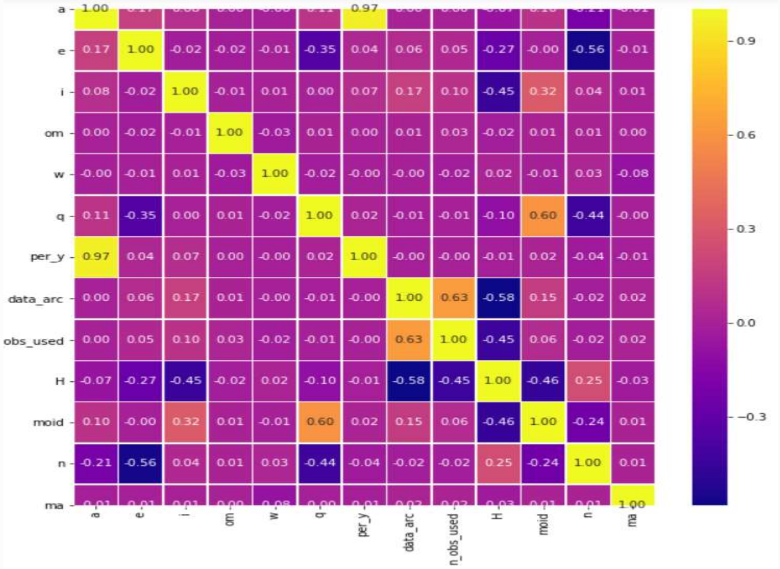
This dataset is for the asteroid is taken from https://ssd.jpl.nasa.gov/sbdb_query.cgi. It is mainly an open asteroid dataset which is updated in regular interval. The main source of this data is a Nasa observatory. It is also found in the Kaggle. This dataset has many type of asteroids. From that dataset we extract only 4 near earth asteroids labelled as AMO, AST, APO, ATE. After the extraction of data, the final dataset for taken for our project is comprised of 21464 columns and 31 rows.

B. ***DATA PREPROCESSING and VISUALIZATION:***

- In this pre-processing purpose, we first change the object type values to the float values and remove the null values from the columns, which have more than 75% of the NULL value. After doing this, the remaining null values are filled with the mean values of each columns.

- We also removed some unnecessary columns such as hazardous, astronomical type, near earth attribute etc. as those columns are unnecessary for this classification.

- In the visualization purpose we use heat map, count plot, boxplot, pair plot. While using heatmap, it has been seen that the two columns are highly correlated with the other two columns. We have to remove the any two of them.



After removing those two columns, the number of attributes in the final set is 13.

C. ***FEARTURE ENGINEERING AND SELECTION:***

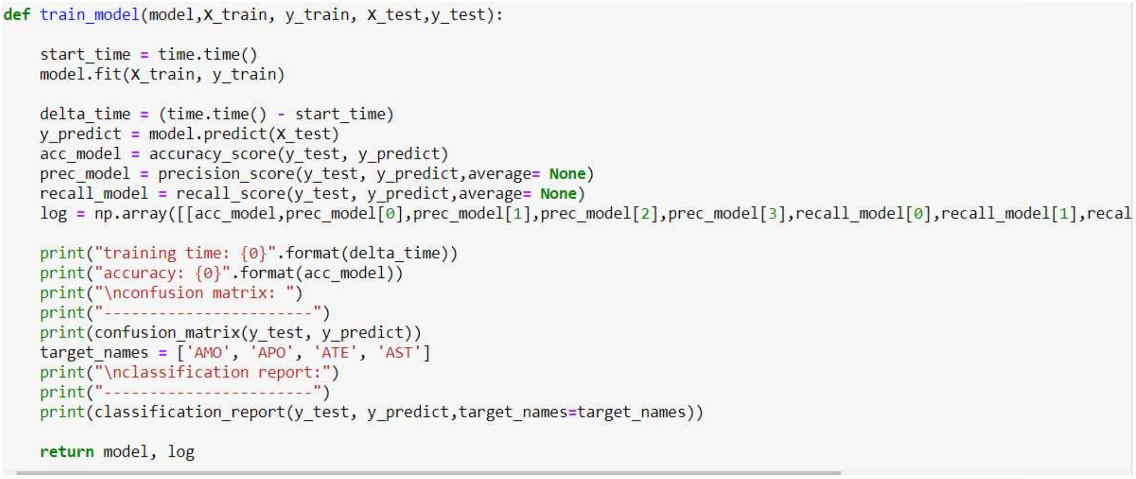
For this purpose, we use Principal component analysis. In the boxplot scenario, we have seen that, many values are in the outliers of the dataset. So we cannot reject those huge amount of values. So we calculate PCA and then check the variance measure of the attributes. After seeing that there is no sudden drop of variance among the attributes, we take all the 13 attributes for the classification purpose.

Those attributes are semi major axis, eccentricity, inclination with respect to plane, longitude of ascending node, perihelion distance, its argument, distance from the earth orbit, orbital period, observation, absolute magnitude parameter, magnitude distance from sun etc. Then using those attributes and the PCA value, we classify the near earth asteroids.

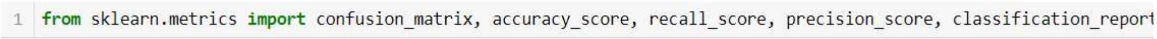
D. *MODEL SELECTION:*

For this classification purpose, we have select 8 different machine learning algorithms. They are logistic regression model, support vector machine, decision tree, naïve Bayes model, random forest, gradient boosting, xgboost and KNN classifier model. Later, we also used an artificial neural network comprised of 5 dense layer and 2548 parameters to classify the asteroids.

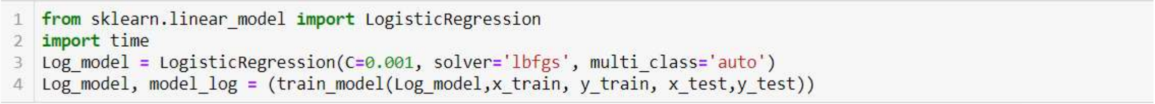
Later we split the dataset in the test and train purpose. Here 80% of the dataset means, 17171 number of data used for training set while other 20% of the dataset is used for the testing purpose. Now, the model are trained by the training set and classify those classes.



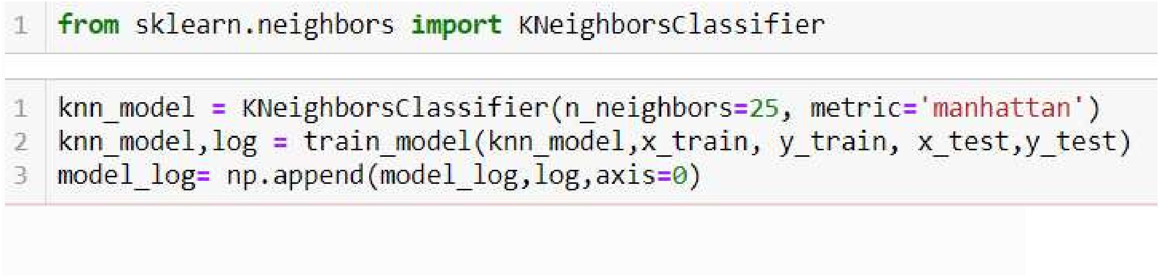
```
def train_model(model,X_train, y_train, X_test,y_test):

    start_time = time.time()
    model.fit(X_train, y_train)

    delta_time = (time.time() - start_time)
    y_predict = model.predict(X_test)
    acc_model = accuracy_score(y_test, y_predict)
    prec_model = precision_score(y_test, y_predict,average= None)
    recall_model = recall_score(y_test, y_predict,average= None)
    log = np.array([[acc_model,prec_model[0],prec_model[1],prec_model[2],prec_model[3],recall_model[0],recall_model[1],recal

    print("training time: {0}".format(delta_time))
    print("accuracy: {0}".format(acc_model))
    print("\nconfusion matrix: ")
    print("----------------------")
    print(confusion_matrix(y_test, y_predict))
    target_names = ['AMO', 'APO', 'ATE', 'AST']
    print("\nclassification report:")
    print("----------------------")
    print(classification_report(y_test, y_predict,target_names=target_names))

    return model, log
```



```
from sklearn.metrics import confusion_matrix, accuracy_score, recall_score, precision_score, classification_report
```



```
from sklearn.linear_model import LogisticRegression
import time
Log_model = LogisticRegression(C=0.001, solver='lbfgs', multi_class='auto')
Log_model, model_log = (train_model(Log_model,x_train, y_train, x_test,y_test))
```

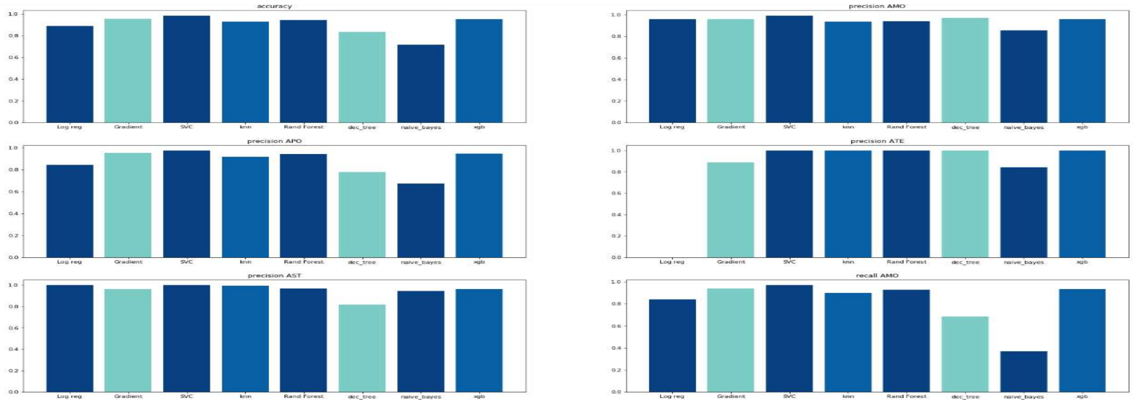


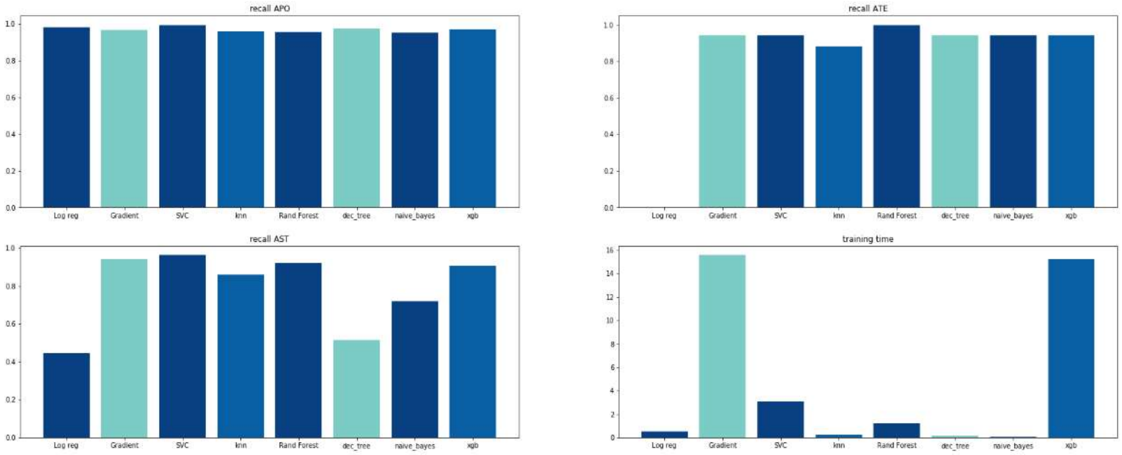
```
from sklearn.neighbors import KNeighborsClassifier
```

```
knn_model = KNeighborsClassifier(n_neighbors=25, metric='manhattan')
knn_model,log = train_model(knn_model,x_train, y_train, x_test,y_test)
model_log= np.append(model_log,log,axis=0)
```

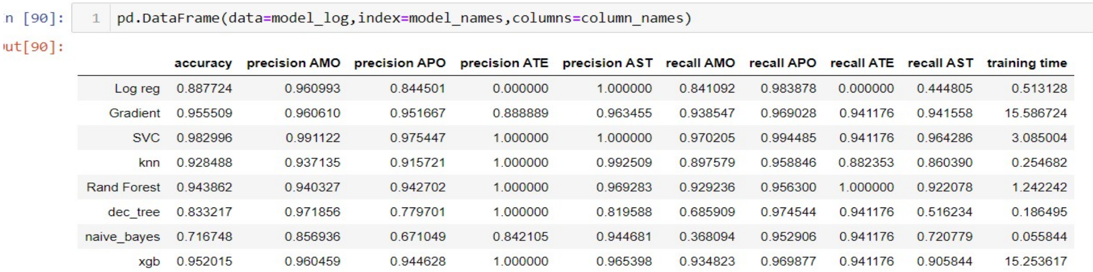
6. ***RESULTS and OUTPUT***

Here we can after testing the data, gradient boosting method gives the highest accuracy of 95.5% among the machine learning techniques, while Xgboost are next best with accuracy of 95.2%. The lowest accuracy percentage is Naïve Bayes classifier model having 71.6% only.





Above graphical representation of the results of different machine learning techniques is shown below by the tabular format.

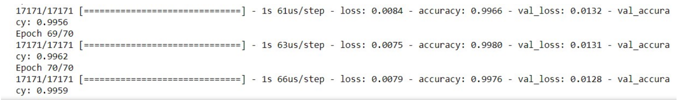


```
n [90]:  pd.DataFrame(data=model_log,index=model_names,columns=column_names)
ut[90]:
```

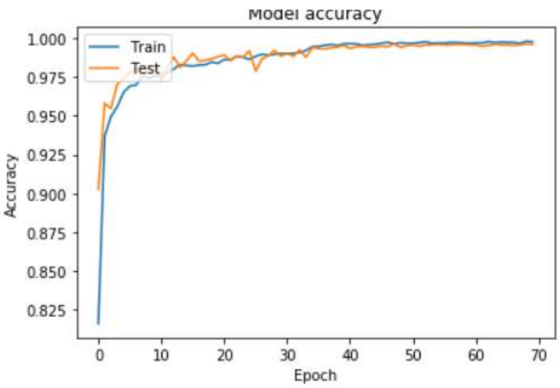
	accuracy	precision AMO	precision APO	precision ATE	precision AST	recall AMO	recall APO	recall ATE	recall AST	training time
Log reg	0.887724	0.960993	0.844501	0.000000	1.000000	0.841092	0.983878	0.000000	0.444805	0.513128
Gradient	0.955509	0.960610	0.951667	0.888889	0.963455	0.938547	0.969028	0.941176	0.941558	15.586724
SVC	0.982996	0.991122	0.975447	1.000000	1.000000	0.970205	0.994485	0.941176	0.964286	3.085004
knn	0.928488	0.937135	0.915721	1.000000	0.992509	0.897579	0.958846	0.882353	0.860390	0.254682
Rand Forest	0.943862	0.940327	0.942702	1.000000	0.969283	0.929236	0.956300	1.000000	0.922078	1.242242
dec_tree	0.833217	0.971856	0.779701	1.000000	0.819588	0.685909	0.974544	0.941176	0.516234	0.186495
naive_bayes	0.716748	0.856936	0.671049	0.842105	0.944681	0.368094	0.952906	0.941176	0.720779	0.055844
xgb	0.952015	0.960459	0.944628	1.000000	0.965398	0.934823	0.969877	0.941176	0.905844	15.253617

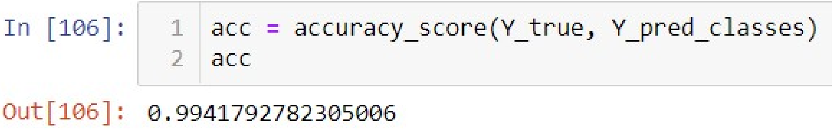
From that above table we can easily reach the conclusion of which one is the best model and also it can be seen that the which takes the less time for testing the test values.

The accuracy of that artificial neural network is the best one after 70 epochs executed it has shown the accuracy of 99.41%, which is the much better than the best machine learning model
named gradient boosting algorithm.



```
17171/17171 [==============================] - 1s 61us/step - loss: 0.0084 - accuracy: 0.9966 - val_loss: 0.0132 - val_accura
cy: 0.9956
Epoch 69/70
17171/17171 [==============================] - 1s 63us/step - loss: 0.0075 - accuracy: 0.9980 - val_loss: 0.0131 - val_accura
cy: 0.9962
Epoch 70/70
17171/17171 [==============================] - 1s 66us/step - loss: 0.0079 - accuracy: 0.9976 - val_loss: 0.0128 - val_accura
cy: 0.9959
```





```
In [106]:  acc = accuracy_score(Y_true, Y_pred_classes)
           acc

Out[106]: 0.9941792782305006
```

7. *CONCLUSION*

In this paper, we want to compare the accuracy of the various model for the classification of the near earth asteroid with the various machine learning techniques. Later we compare the accuracy percentage of the model by the basis of various features. Here we take 13 important features. Then for the better accuracy we use the PCA for the principal component analysis. Then we use those models. With the help of the accuracy, we choose the best one compare a neural network model. Neural model give the better accuracy of 99.4%, but with the change of various parameters and analysis we may achieve greater accuracy.

8. REFERENCES

[1] *Asteroids and comets, Y. R. Fernandez, J.-Y. Li, E.S. Howell, L.M. Woodney*

[2] *Applying Machine Learning to Asteroid Classification Utilizing Spectroscopically Derived Spectrophotometry, Kathleen Jacinda Mcintyre*

[3] *Asteroid Mining- A review of Methods and Aspects, Vide Hellgren*

[4] *Asteroid families classification: Exploiting very large datasets, Andrea Milani, Alberto Cellino, Zoran Knezevic, Bojan Novakovic, Federica Spoto, Paolo Paolocchi*

[5] *Near Earth asteroid mining, Shane D. Ross*

[6] *Prediction of Asteroid Diameter, Victor Basu*

[7] *Machine Learning in Astronomy: A Case-study in Quasar-star classification, Mahammod Viquor, Suryoday Basak, Arunina Dasgupta*

[8] *Applying Machine Learning to Asteroid Classification Utilizing Spectroscopically Derived Spectrophotometry, Kathleen Macntry*

[9] *Spectral Classification of Asteroids by Random Forest, Huang Chao*

[10] *Machine learning and next generation asteroid surveys, Nunget, Carry*

[11] *Galaxy classification: A machine learning analysis of GAMA catalogue data, Aleke Nolte*

[12] *Machine learning classification of variable stars and eclipsing binaries in K2 fields 0-4, D. J. Armstrong*

[13] *Classifications of the near earth asteroids with artificial neural network, Zoltan Mako*

[14] *Characterization of the near earth asteroids using KMTNET-SAAO, N. Erasmus*

[15] *A new dynamical classification of asteroid, Florian Friestter*