Nasa Hazardous Asteroid Classification using various Classification Techniques

Prathith T.A   
Dept of Computer Science and Engg  
PES University (Great Learning)  
Bengaluru, Karnataka

[prathith.ta@gmail.com](mailto:prathith.ta@gmail.com)

Rishabh Kumar  
Dept of Computer Science and Engg  
PES University (Great Learning)Bengaluru, Karnataka  
[rishabh11081994@gmail.com](mailto:rishabh11081994@gmail.com)

Rutuja Shendkar   
Dept of Computer Science and Engg  
PES University (Great Learning)Bengaluru, Karnatakarutujashendkar@gmail.com

***Abstract*—The focus of this research is to find a precise model for predicting hazardous asteroids. Machine Learning techniques such as Logistic Regression, Naïve Bayes, KNN, Decision Tree etc. are used. Boosting, Bagging and Stacking are also explored in this paper. Furthermore, hyperparameter tuning and feature selection is also done to get the most optimum result. 10 models were built using these classification techniques. After comparison, it was found that many models were performing well and picked the best out of these using accuracy, precision and recall.**

***Keywords — Logistic Regression, Naïve Bayes, KNN, Decision Tree, Machine Learning, Supervised, Classification, Bagging, Boosting, Stacking, Hazardous, Non-Hazardous, Asteroids***

# Introduction

66 million years ago, a large asteroid colliding with earth caused a calamitous end to 75% of the Earth’s animals including dinosaurs [1]. Asteroids, the irregularly shaped rocky bodies that orbit the sun, have a very remote probability of striking the Earth and causing damage. However, when these collisions occur, the devastating consequences can last for decades. Hence, detecting and mitigating the threat in time is essential. In addition, asteroids closer to the surface of the earth can be exploited for extraordinary supply of minerals and raw materials [2]. These observations have triggered the scientific interest around asteroids.

This paper aims at classifying the hazardous and non-hazardous asteroids also known as Near Earth Objects (NEO) using Machine Learning techniques. The dataset is taken from Nasa's CNEOS which is responsible for computing asteroids and their orbit and their likelihood of impacting the Earth. The data has 40 attributes with 4687 records. The dataset contains a boolean variable called Hazardous which classifies each record into hazardous or non-hazardous categories. The data contains descriptions of the asteroid such as minimum and maximum estimated diameter. These are approximate values as the objects are irregularly shaped.

A detailed description of each feature is given below:

**Neo Reference ID** - Near Earth Object (NEO) reference ID number for an asteroid

**Name** - 'Name' of asteroid

**Absolute Magnitude** - A measure of the asteroid's luminosity (in H) (the brightness of an asteroid if it is 1 astronomical unit away from both the Sun and the observer, and the angle between the Sun, asteroid, and Earth is 0 degrees)

**Est Dia in (in KM, M, Miles, and Feet) (min)** - Minimum estimated diameter of the asteroid

**Est Dia in (in KM, M, Miles, and Feet) (max)** - Maximum estimated diameter of the asteroid

**Close Approach Date** - Date at which the asteroid approaches close to Earth

**Epoch Date Close Approach** - Date at which the asteroid approaches close to Earth (in epoch time)

**Relative Velocity (in km per sec, km per hr, and miles per hour)** - Asteroid's velocity relative to earth

**Miss Dist. (in Astronomical, lunar, km, and miles)** - Distance by which the asteroid misses Earth

**Orbiting Body** - Planet which the asteroid orbits

**Orbit ID** - An ID of JPL NEA orbit that JPL Nasa uses in its analysis

**Orbit Determination Date** - Date at which the asteroid's orbit was determined

**Orbit Uncertainty** - A measure of the uncertainty ('measurement errors') in the calculated orbit

**Minimum Orbit Intersection** - The closest distance between Earth and the asteroid in their respective orbits (in astronomical units)

**Jupiter Tisserand Invariant** - A value used to differentiate between asteroids and Jupiter-family comets

**Epoch Osculation** - The instance of time at which the asteroid's position and velocity vectors (from which its osculating orbit is calculated) is specified

**Eccentricity** - A value which specifies by how much the asteroid's orbit deviates from a perfect circle

**Semi Major Axis** - The longest radius of an elliptical orbit; a measure of the asteroid's average distance from the Sun (asteroids orbit the Sun)

**Inclination** - Measures the tilt of the asteroid's orbit around the Sun

**Asc Node Longitude** - Angle in the ecliptic plane between the inertial-frame x-axis and the line through the ascending node

**Orbital Period** - Time taken for asteroid to complete a single orbit around the Sun

**Perihelion Distance** - Distance of point in asteroid's orbit which is closest to the Sun

**Perihelion Arg** - The angle (in the body's orbit plane) between the ascending node line and perihelion measured in the direction of the body's orbit

**Aphelion Dist** - Distance of point in asteroid's orbit which is farthest from the Sun

**Perihelion Time** - Length of time of asteroid's passage through the perihelion stage

**Mean Anomaly** - The product of an orbiting body's mean motion and time past perihelion passage

**Mean Motion** - The angular speed required for a body to make one orbit around an ideal ellipse with a specific semi-major axis

**Equinox** - An astronomical standard to measure against (currently 'J2000.0')

**Hazardous** - Is the asteroid hazardous? (True or False)

This paper intends to use various classification algorithms such as Logistic Regression, K Nearest Neighbour, Decision Tree. Ensemble methods such as Bagging, Boosting and Stacking are also used to improve model performance.

# literature Review

In one of the articles ‘NASA Asteroid Classification’ Classifying whether an asteroid is hazardous or not, the author Shubhankar Rawat applies Naïve Bayes, SVM, Decision Tree and LightGBM (Light Gradient Boosting Machine – framework developed by Microsoft) on the dataset. Along with precision, accuracy, error rate the author also uses confusion matrix to analyse the performance of each model. The Naïve Bayes model is found to be of 0 significance as it does not predict any record as hazardous. For SVM, the author finds that the computational speed slows. Moreover, little to no improvement is observed over Naïve Bayes. The article concludes by stating Decision Tree Classifier to be the best model with 99.4% accuracy followed by LightGBM with 99.3% [3].

In another research paper ‘Hazardous Asteroid Classification through Various Machine Learning Techniques’ by Anish S, the author develops 8 models using Machine Learning techniques. These are Logistic Regression, Support Vector Machine, Decision Tree, K Nearest Neighbour, Random Forest, Naïve Bayes, Adaboost and XGBoost. The author also performs feature selection. From this only 15 features of 40 are selected. The lowest accuracy achieved in this research is 80.70% by Naïve Bayes and highest of 100% by Random Forest and XGBoost. In conclusion, the author finds that random forest with n\_estimators as 15 provides the optimum result according to accuracy and training. However, the author has not explored the issue of overfitting [4].

The author of ‘Machine Learning identification of asteroid groups’ using different dataset compare the use of traditional Hierarchical Clustering Method (HCM) to machine learning HCM. The distance of each asteroid is computed. If this distance is found to be less than a critical value, the asteroid is then classified into a family. Thus, 6 new families were identified. In comparison of the two methods, it was found that the machine learning HCM was able to find family members with an accuracy of 89.5% [5].

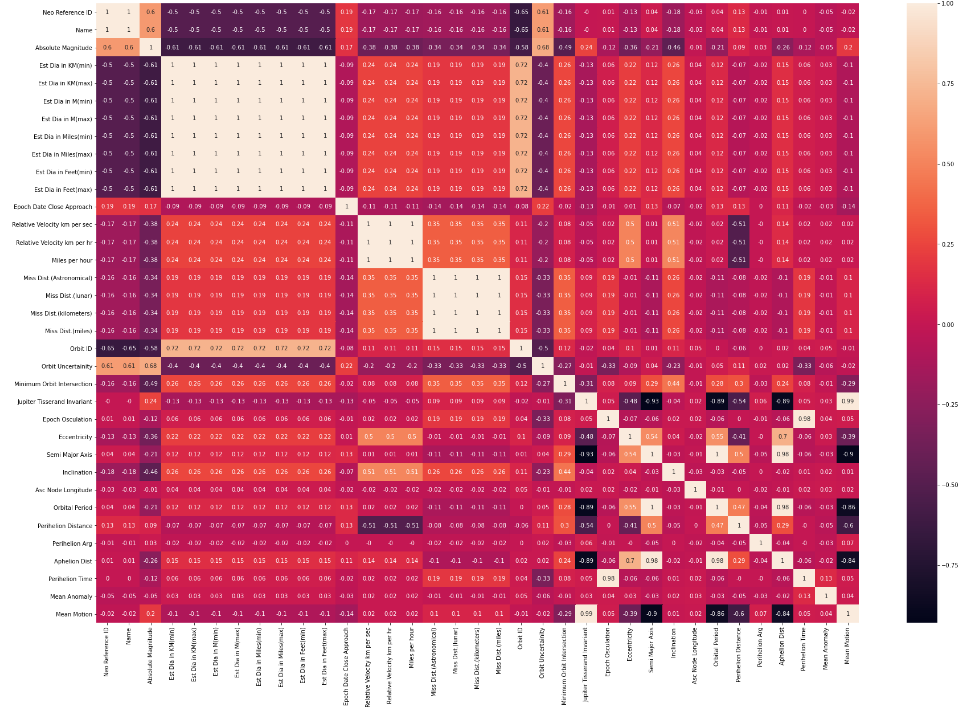
In the paper, ‘Identifying Earth-impacting asteroids using an artificial neural network’ by John D et. al, the authors developed and trained an artificial neural network in order to classify asteroids which might cause an impact to the earth. This developed instrument was named as Hazardous Object Identifier (HOI) and was able to identify 95.25% of hazardous objects positively [6].

# Data Preparation

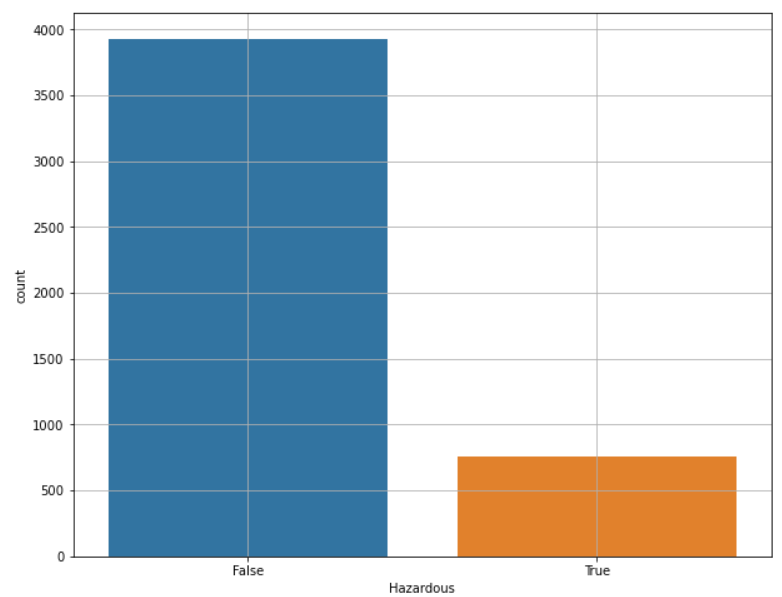
The dataset was checked for any null, missing and duplicate values. However, none were found. Irrelevant columns such as ‘Neo Reference ID’, ‘Name’, ‘Orbiting Body’, ‘Equinox’, ‘Orbit ID’ etc were dropped. The ‘Close Approach Date’ was converted to an interval. a 5-point summary was obtained of the data which confirmed no missing values and gave the count, mean, min, max and percentile values.

# Exploratory Data analysis

A heatmap using Pearson’s correlation coefficient was plotted to obtain the correlation between the features. There is collinearity in some of the features.



In terms of the count for hazardous and non-hazardous asteroids, it was found that 3932 asteroids are non-hazardous whereas 755 are hazardous. A barplot was used to convey the same.



# Train-Test Split

The dataset was split into a train and test set in order to make it suitable for Machine Learning techniques. The data was split at random state of 21 in the size of 70-30 where 70% of the data would be considered as train set and the rest 30% would be test set. Hence, 3280 records were in the train set and 1407 in the test set. X was taken as the dependent variable i.e. Hazardous and the rest of the features were stored in y.

# Model Development

In this section of the paper, the models have been explained briefly. For all the models, first a preliminary model was built. This model was then tuned using hyperparameters, feature selection etc. Flow chart of the approach is provided below:

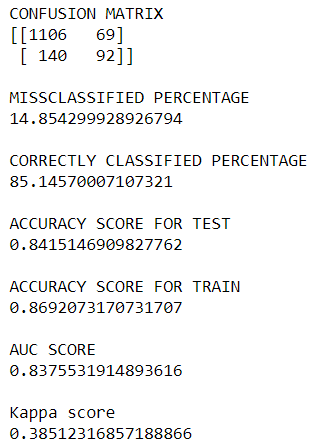
Preliminary Model

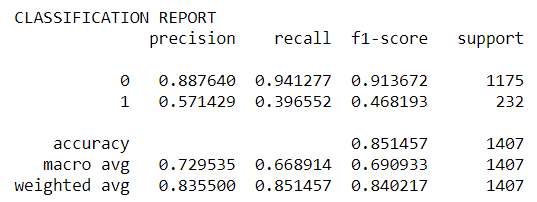
Model Tunning

Final Model

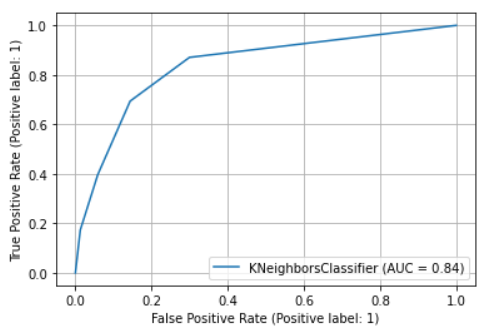
1. KNN

The first model run was KNN where the number of neighbours used were 5. For model evaluation, measures such as Confusion Matrix, misclassified percentage, accuracy score for test and train, AUC score Kappa score along with classification report for test set was computed.





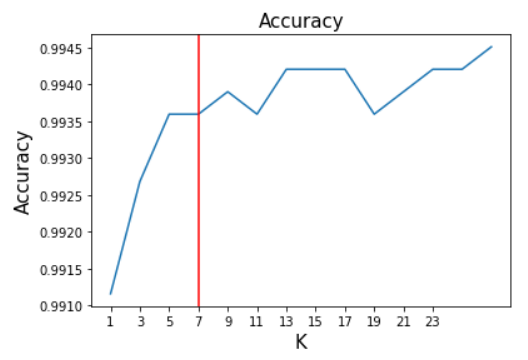
From the confusion matrix it can be seen that 1198 records are correctly identified. From the classification report, the accuracy of the model is 85.1 %. However, the recall value is only 0.39. The weighted avg of f1-score is 0.84 and the AUC score (see graph below) for the test set is 0.83 which is decent. From these parameters, it can be said that this is a decent model, however, improvements can be done to optimise its performance.



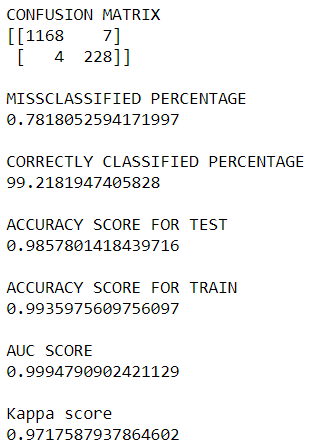
Therefore, the model was further tuned using hyper parameters. A dictionary with various types of metrics and weights and a range for neighbours was passed through gridsearchcv where the cross-validation (cv) value was taken as 5. The best parameters that were found after this were {'metric': 'manhattan', 'n\_neighbors': 3, 'weights': 'distance'}.

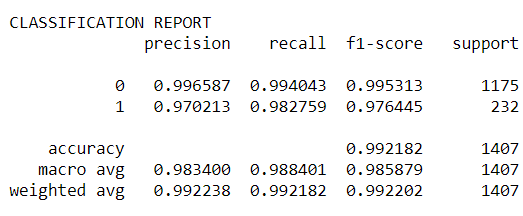
Furthermore, feature selection was done to ensure the model achieves optimum performance level. The Sequential Feature Selection was imported from Machine Learning Extension (mlxtend) library. The model developed was passed through this feature selector. The default setting of forward selection and cv=5 was used. The following feature subset was obtained:

('Est Dia in Miles(min)', 'Minimum Orbit Intersection'). Lastly, the optimum K value for the model was obtained. The graph given below portrays the change in accuracy when different values of K are taken by the model. The optimum accuracy obtained was when K = 7.



A model with tuned hyperparameters, optimum K value and feature selection gives the following result.

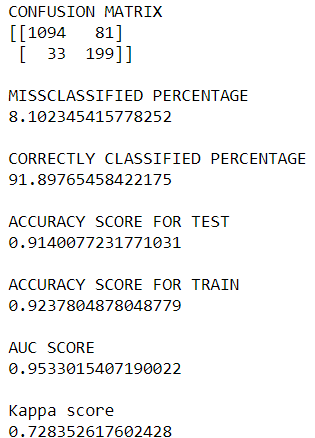


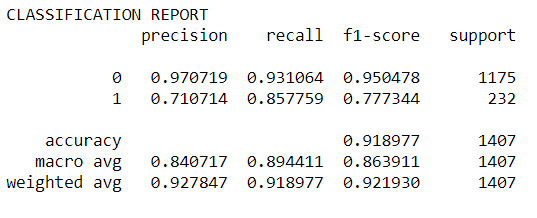


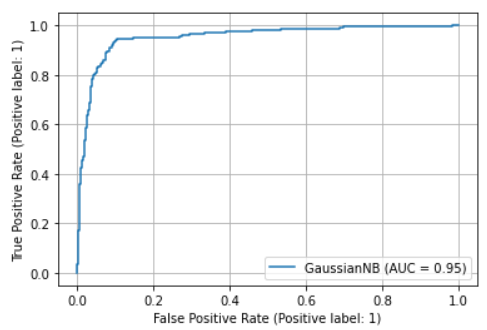
It can be seen that the accuracy has increased to 99.2%. The AUC score has grown to 0.99 which is considered the best. The recall value has improved immensely from 0.39 to 0.98.

Overall, the model has achieved optimum performance level and will not require further tuning.

1. Gaussian Naïve Bayes

A preliminary Naïve Bayes model was constructed. The model performance measure was as given below.



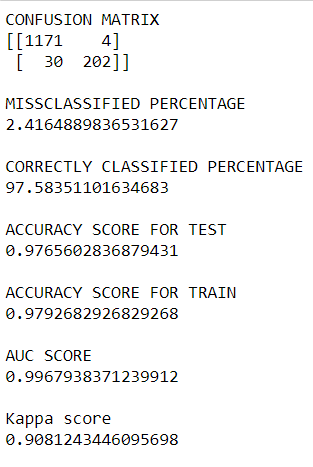


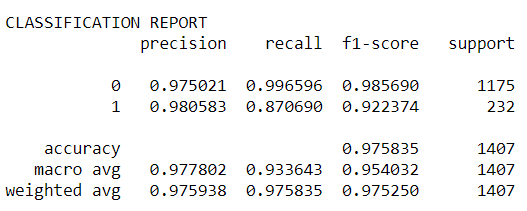
The accuracy of the model is found to be 91.8% which is very good. The AUC score is 0.95 as seen in the above graph. The recall value of 0.85 is also good. This algorithm does not contain hyperparameters to tune further. When the rest of the scores are also taken into account it can be said that the model is performing very well. However, it was improved further, feature selection was performed on this using sequential feature selector along with forward selection and cv = 5(cross validation). The following features were selected:

('Absolute Magnitude', 'Minimum Orbit Intersection',

'Semi Major Axis', 'Asc Node Longitude').

After creating a model with these features the performance of the model was as follows:

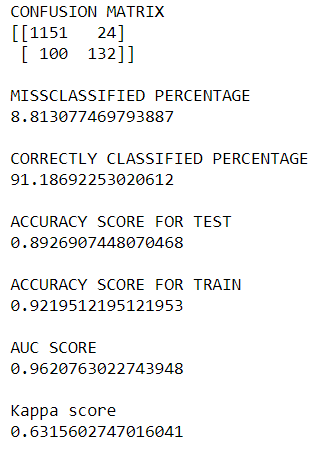


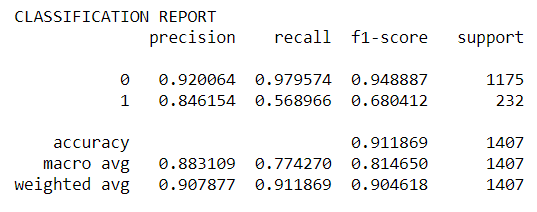


97.5% of accuracy with a recall value of 0.87 was obtained by the model which is excellent. Moreover, the AUC score has improved to 0.99 which is very close to the optimum value of 1. Moreover, the model has achieved a 0.90 Kappa score which is almost perfect. The model can be expected to give excellent results.

1. Logistic Regression

The preliminary model of Logistic Regression gave the following result.





From the classification report above, it can be seen that the accuracy of this model is 91.1% which is high. The AUC score and the weighted avg of f1-score is above 0.90 which is good. To achieve the optimum model, feature selection was performed. The following feature subset was returned as a result:

('Absolute Magnitude', 'Est Dia in KM(min)',

'Est Dia in KM(max)',

'Est Dia in Miles(min)',

'Est Dia in Miles(max)',

'Close Approach Date',

'Orbit Uncertainty',

'Minimum Orbit Intersection',

'Jupiter Tisserand Invariant',

'Eccentricity',

'Inclination',

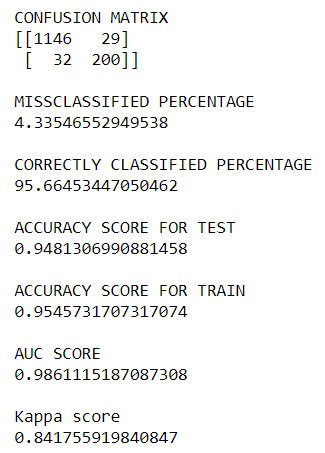
'Perihelion Distance',

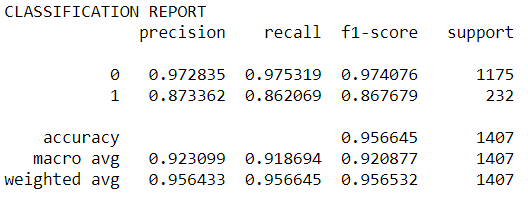
'Mean Motion')

Moreover, the best hyperparameters were found using gridsearchcv with cv = 5, n\_jobs = -1, verbose = 3. The best parameters found were as follows:

{'C': 112.88378916846884, 'penalty': 'l2', 'solver': 'newton-cg'}.

These were then passed to the model to obtain the result given below.





The accuracy has increased to 95.6% with a recall value of 0.86. The weighted average f1-score has increased to 0.95 and a strong kappa score of 0.84 was obtained. Overall, the model has decent performance.

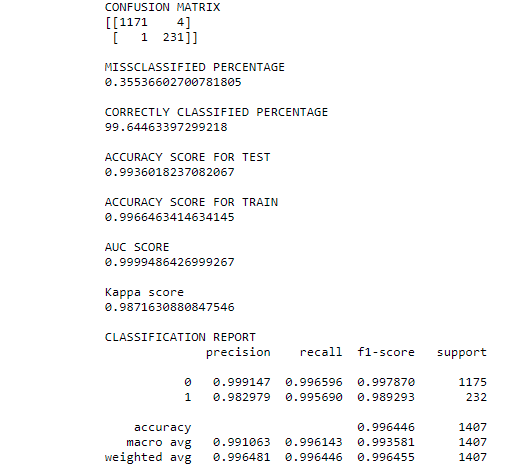
1. Random Forest

Random which uses the decision tree is used here to model our Asteroid Classification Data. Sequential Feature Selector is used with the estimator as a random forest to select the feature which gives the optimum model. The following feature subset were returned as a result:

('Est Dia in Miles(max)',

'Minimum Orbit Intersection')

After the features were selected then we built the random forest model using those features . The primary model gave the following result:



From the classification report above it can be seen that accuracy of this model is 0.99% which is high.

The AUC score and the weighted average of the f1

Score is above 0.95 which is good and the precision and recall values is 0.99 which tells that

The misclassified percentage is very less.

Hyper parameter tuning of the random forest is done

using grid search cv and best parameters are

found which are

{'criterion': 'gini',

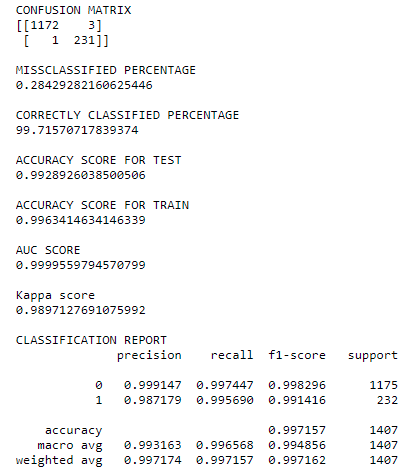
'max\_depth': 13,

'min\_samples\_split': 3,

'n\_estimators': 100}

Using these parameters the random forest model

is tuned and we get the following results :



This is the best model till now, it leads knn tuned models in terms of misclassified percentage. Every parameter whether it is accuracy, f1 score, precision and recall are almost 99 %.The false negative is only 1 in the model, the search for best is still running.

5.AdaBoosting : Ada boost is one of the ensemble technique it boost the weak learner and make sure that there is no overfitting and under fitting. Before applying this algorithm on the data forward feature selection method was used to find the best feature. The best feature are magnitude, minimum orbit intersection, Jupiter tisserand invariant, and eccentricity. The surprising fact is that when this feature were feed to stack classifier , the performance of the model was exceptionally improved, it will discussed separately in the upcoming topic. This algorithm work best for asteroid data when base estimator is Random Forest. After selecting the best feature the model was tuned with base estimator as random forest and following results were obtained.

CONFUSION MATRIX

[[1173 2]

[ 1 231]]

MISSCLASSIFIED PERCENTAGE

0.21321961620469082

CORRECTLY CLASSIFIED PERCENTAGE

99.7867803837953

ACCURACY SCORE FOR TEST

0.9936018777920801

ACCURACY SCORE FOR TRAIN

0.9942073170731707

AUC SCORE

0.9999633162142333

Kappa score

0.9922712549643592

CLASSIFICATION REPORT

precision recall f1-score support

0 0.999148 0.998298 0.998723 1175

1 0.991416 0.995690 0.993548 232

accuracy 0.997868 1407

macro avg 0.995282 0.996994 0.996136 1407

weighted avg 0.997873 0.997868 0.997870 140

The results are great, model is performing excellent in all scores. However, the intent is to find the best model which score almost 100% in recall to avoid type 2 error.

6.Gradient Boost:

Among the boosting technique gradient descent algorithm is one of the best. It often give very good result. The same approach was followed while applying the gradient descent algorithm which is finding best features for the algorithm using forward selection technique and then tuning the hyper parameter with grid search function. The results are mentioned below:

CONFUSION MATRIX

[[1174 1]

[ 4 228]]

MISSCLASSIFIED PERCENTAGE

0.35536602700781805

CORRECTLY CLASSIFIED PERCENTAGE

99.64463397299218

ACCURACY SCORE FOR TEST

0.9928976697061802

ACCURACY SCORE FOR TRAIN

0.9960365853658537

AUC SCORE

0.9999266324284666

Kappa score

0.9870291737574465

CLASSIFICATION REPORT

precision recall f1-score support

0 0.996604 0.999149 0.997875 1175

1 0.995633 0.982759 0.989154 232

accuracy 0.996446 1407

macro avg 0.996119 0.990954 0.993515 1407

weighted avg 0.996444 0.996446 0.996437 1407

The kappa score is .987, recall= .999 for 0 and .9827 for 1 , no doubt the algorithm is giving excellent results. The hunt to search best is still continued.

1. XGBoost

As the name suggest it boosting technique ,it was applied on the data after tuning the following results were obtained:

ONFUSION MATRIX

[[1174 1]

[ 3 229]]

MISSCLASSIFIED PERCENTAGE

0.28429282160625446

CORRECTLY CLASSIFIED PERCENTAGE

99.71570717839374

ACCURACY SCORE FOR TEST

0.9936018777920801

ACCURACY SCORE FOR TRAIN

0.9954268292682927

AUC SCORE

0.99994497432135

Kappa score

0.9896413492013826

CLASSIFICATION REPORT

precision recall f1-score support

0 0.997451 0.999149 0.998299 1175

1 0.995652 0.987069 0.991342 232

accuracy 0.997157 1407

macro avg 0.996552 0.993109 0.994821 1407

weighted avg 0.997155 0.997157 0.997152 1407

In the investigation it is found that the algorithm is doing well to avoid false alarm but not performing competitive with method like random forest to neglect false negative count.

1. Combined Prediction:

In this predictive approach we used the all three advance algorithm, namely ada boost, xgboost and gradient boost. Voting was done to predict the result first using the prediction of each algorithm, later using the average of probability while keeping threshold as 0.5, the performance of the model was good but not very competitive compare to stacking and random forest method.

Results when voting was used for the prediction.

Confusion matrix

[[1173 2]

[ 3 229]]

MISSCLASSIFIED PERCENTAGE

0.35536602700781805

CORRECTLY CLASSIFIED PERCENTAGE

99.64463397299218

CALSSIFICATION REPORT

precision recall f1-score support

0 0.99745 0.99830 0.99787 1175

1 0.99134 0.98707 0.98920 232

accuracy 0.99645 1407

macro avg 0.99440 0.99268 0.99354 1407

weighted avg 0.99644 0.99645 0.99644 1407

Results when average probabilities were used for prediction:

[[1174 1]

[ 4 228]]

MISSCLASSIFIED PERCENTAGE

0.35536602700781805

CORRECTLY CLASSIFIED PERCENTAGE

99.64463397299218

CALSSIFICATION REPORT

precision recall f1-score support

0 0.99745 0.99830 0.99787 1175

1 0.99134 0.98707 0.98920 232

accuracy 0.99645 1407

macro avg 0.99440 0.99268 0.99354 1407

weighted avg 0.99644 0.99645 0.99644 1407

1. Stack model :

In this model base estimator were random forest, ada boost, xg boost, gradient boost with base estimator as naïve bayes. Algorithm gave best performance till now, the ouput

of the python code for this model is mentioned below.

CONFUSION MATRIX

[[1172 3]

[ 0 232]]

MISSCLASSIFIED PERCENTAGE

0.21321961620469082

CORRECTLY CLASSIFIED PERCENTAGE

99.7867803837953

ACCURACY SCORE FOR TEST

0.9936018777920801

ACCURACY SCORE FOR TRAIN

0.9948170731707318

AUC SCORE

0.9987234042553191

Kappa score

0.9922978528508528

CLASSIFICATION REPORT

precision recall f1-score support

0 1.000000 0.997447 0.998722 1175

1 0.987234 1.000000 0.993576 232

accuracy 0.997868 1407

macro avg 0.993617 0.998723 0.996149 1407

weighted avg 0.997895 0.997868 0.997873 1407

The model is performing amazing but used many algorithm which increases computation, however the recall and precision are exceptionally 100% for this model.

1. Save earth:

It is a stack model with base estimator as random forest and final estimator as naïve bayes. The model performs exceptionally well in recall for 1 and precision for 0

, the accuracy , kappa score and other score are competitive with other good scoring algorithm. This should be the most preferable algorithm for prediction of

Hazardous algorithm. If the algorithm is applied on the feature sselected using ada boost in forward selection technique, it performance exceptionally well. The

results are mentioned below:

CONFUSION MATRIX

[[1172 3]

[ 0 232]]

MISSCLASSIFIED PERCENTAGE

0.21321961620469082

CORRECTLY CLASSIFIED PERCENTAGE

99.7867803837953

ACCURACY SCORE FOR TEST

0.9936018777920801

ACCURACY SCORE FOR TRAIN

0.9951219512195122

AUC SCORE

0.9991489361702127

Kappa score

0.9922978528508528

CLASSIFICATION REPORT

precision recall f1-score support

0 1.000000 0.997447 0.998722 1175

1 0.987234 1.000000 0.993576 232

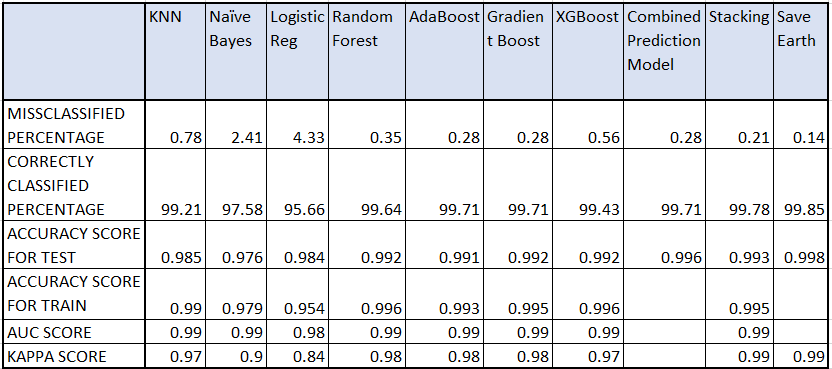
accuracy 0.997868 1407

macro avg 0.993617 0.998723 0.996149 1407

weighted avg 0.997895 0.997868 0.997873 14

# CONCLUSION

The table below summarizes all 10 models that have been developed.



It can be seen that the misclassified percentage is low for the ensemble models except for XGBoost. In terms of accuracy score for the test, the highest is found to be 0.998 of the Save Earth Model followed by Combined Prediction Model.

# References

[1] E. Osterloff, "How an asteroid ended the age of the dinosaurs," Natural History Museum, 18 November 2020. [Online]. Available: https://www.nhm.ac.uk/discover/how-an-asteroid-caused-extinction-of-dinosaurs.html. [Accessed 23 May 2021].

[2] D. K. Yeomans, "Why Study Asteroids?," Nasa - Jet Propulsion Laboratory/California Institute of Technology, [Online]. Available: https://ssd.jpl.nasa.gov/?why\_asteroids#:~:text=Why%20Study%20Asteroids%3F,-Donald%20K.&text=The%20scientific%20interest%20in%20asteroids,Earth's%20biosphere%20in%20the%20past. [Accessed 23 May 23].

[3] S. Rawat, "NASA Asteroid Classification - Classifying whether an asteroid is hazardous or not.," towards data science, 10 December 2019. [Online]. Available: https://towardsdatascience.com/nasa-asteroid-classification-6949bda3b1da. [Accessed 23 May 2021].

[4] A. Si, "Hazardous Asteroid Classification through Various Machine Learning," International Research Journal of Engineering and Technology (IRJET), vol. 07, no. 03, 2020.

[5] V. Carruba, S. Aljbaae and A. Lucchini, "Machine-learning identification of asteroid groups," Monthly Notices of the Royal Astronomical Society, vol. 488, no. 1, pp. 1377-1386, 2019.

[6] J. D. Hefele, F. Bortolussi and S. P. Zwart, "Identifying Earth-impacting asteroids using an artificial neural network," Astronomy and Astrophysics (A&A), vol. 634, 2020.