Hazardous Asteroid Prediction Using Machine Learning

Dr. Babu Rao M
Dept. of Computer Science and Engineering
SR Gudlavalleru Engineering College
Gudlavalleru, India
baburaompd@gmail.com

Devi Alekhya G
Dept. of Computer Science and Engineering
SR Gudlavalleru Engineering College
Gudlavalleru, India
devialekhya7842@gmail.com

Aakanksha J
Dept. of Computer Science and Engineering
SR Gudlavalleru Engineering College
Gudlavalleru, India
jaruguaakanksha@gmail.com

Naga Jyothi K
Dept. of Computer Science and Engineering
SR Gudlavalleru Engineering College
Gudlavalleru, India
kokkirapatinagajyothi2002@gmail.com

***Abstract*— With the help of machine learning algorithms, make a prediction of potentially hazardous asteroids (PHAs) which are having different combinations of orbital parameters. The suggested method intends to identify all major groupings of PHA-rich near-Earth asteroids (NEAs) and their subgroups. we applied different machine learning methods namely Random Forest, decision tree, k nearest neighbours, Naive bayes, Logistic Regression, Support Vector Machine (SVM) – for identifying potential hazardous and non-hazardous asteroids and features responsible for claiming an asteroid to be hazardous. so far as, asteroids are considered, there are many asteroids which are nearer to the earth, but all are not hazardous. So, our target in this project is to identify the asteroids which are hazardous and classify them with non-hazardous types. For this, we choose machine learning models. We used the data attributes to train these models, and then we compared the results to choose the model that provided the greatest accurate categorization.**

Keywords— ***Hazardous, Non Hazardous, Machine Learning Algorithms, Navye bayes, SVM, Linear Regression, Random Forest, Decision trees, KNN, Potentially Hazardous Asteroids (PHAs).***

I. Introduction

Small, rocky objects that orbit the Sun are called Asteroids. Although asteroids orbit the Sun like planets, they are much smaller than planets. Coming to the image of space, asteroids are one of the most important topic to discuss because space contains large numbers of objects in which some of them are stars, planets, moon, astronomical bodies, asteroids etc. All near earth objects may contains asteroids, they may fall on the earth and cause severe damage on earth.[9] So, our target in this project is to predict the accuracy of hazardous asteroid by applying different machine learning algorithms and compare their accuracies. Larger asteroids sometimes called Planetoid's. The study of asteroids is also crucial as historical events prove some of them being hazardous. [1] Asteroids that cross Earth's orbit are continuously monitored by scientists because they are potentially dangerous and approach within 28 million miles (45 million kilometers) of our planet.

Machine learning models with moving asteroids is not only interesting but also useful. Being data science enthusiasts, we thought of using machine learning to predict whether an asteroid could be hazardous or not[10]. This covers the following aspects of machine learning: Data Exploration, Data Wrangling, Data Preprocessing, ML Model Developing. Several unnecessary features present in the dataset which hardly contribute towards classification were removed and data was pre-processed before building Machine learning models. Various classification models are used to predict hazardous asteroids approaching earth.

II. Related Work

In this paper[1], Vadym Pasko tried to make some predictions over the combinations of orbital parameters for yet undiscovered and potentially hazardous asteroids by machine learning techniques. For this reason, they have used the Support Vector Machine algorithm with the kernel RBF. By this approach, the boundaries of the potentially hazardous asteroid groups in 2-D and 3-D can be easily understood. By this algorithm, they have achieved the accuracy over 90%.

In this paper[2], M. Popescu, J. Licandro, J.M. Carvano has provided classification of asteroids, which are observed by VISTA-VHS survey. They have used some statistical methods to classify the 18,265 asteroids. They used a probabilistic method, KNN method and a statistical method to classify those. Later, they compared the algorithms' accuracy. They have classified the asteroids into V, S and A types. They did it on the 18265 asteroids and test set consists of over 6400 asteroids.

In this Paper[3] , Ihara Csllik have classified the asteroids which are near to the earth and can easily pass through the orbital of Mars. For this purpose they, have used a perceptron-type network. And their output class are three, they are Apollo, Amor and Aten. They have also found that semi-major axis and focal distance are the two important major features, for which the asteroids can be separable linearly.

In this paper[4] , they identified the asteroid family by the help of the supervised hierarchical clustering method. For this method, they have found the distance between an asteroid from any reference objects. Later they compared the results with traditional hierarchical clustering method and found that their result gives 89.5% accuracy, which is

2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN) | 979-8-3503-4798-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/VITECON58111.2023.10157937

979-8-3503-4798-2/23/$31.00 ©2023 IEEE

better. By this approach, they found 6 new families and 13 clumps.

In this paper[5] , they used KNN, gradient boosting, decision trees, logistic regression to classify the family of the asteroids. They have identified main belt asteroids, which are three body resonant. They have identifies 404 new asteroids. They found that gradient boosting methodgives the accuracy of 99.97% for identification purpose which is maximum among all the methods they have applied.

In this paper[6] , they have also classified the asteroids according to their family such as Apollo, Amor, Aten etc. They have done it only over the asteroids, who pass the mars orbit. They can be easily separable by the semi major axis and the focal distances.

In this paper[7], the author has used random foresttechniques to classify the asteroids. They have done it on the 48642 asteroids. They have classified those asteroids among 8 classes such as C, X, S, B, D, K, L and V.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[8] , 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 paper[9] The naive Bayes classifier is surprisingly efficient in practise despite its erroneous independence assumption since its classification determination may frequently be accurate even if its probability estimations are incorrect.

In this paper [10] They have provided a creative remedy for coping with DNN's drawbacks in this study. They choose a few representatives from the training dataset with some additional information to represent the entire training dataset in order to get around the issues of low efficiency and dependence on k.

Author	Year	Methodology	Advantages	Disadvantages	Remarks
Rishi	2001	Naive Bayes	It handles both continuous and discrete data	Faces the 'zero frequency problem'	It is highly scalable
Gongde Guo, Hui Wang, David Bell Yaxin Bi, Kieran Greer	2002	KNN	Quick calculation and high accuracy	Does not work with very large dataset.	Distance is calculated on all attributes-irrelevant attributes are a problem curse of dimensionality
Chaoying joanne peng, Kuk lida lee, Gary Ingersoll	2008	Logistic Regression	Easier to implement, Interpret and very efficient to train	It is bound to discrete number set	It constructs only linear boundaries
Geard Biau	2012	Random Forest	Flexible to Both Regression and Classification Problem	Requires more time to train compared to decision tree	It does not overfit and powerful algorithm.
Harsh Patel Purvi Prajapti	2018	Decision Tree	Does not require normalization and scaling of data	Relatively expensive and complexity of time is more	Unstable, small change in data leads to large change in decision tree

Fig1: Literature Review

III. Proposed Work

At first the data need to be collected, then after data preprocessing is done to clean the data. Next some data is used for training the model, some data is used for testing the model. Finally Machine Learning algorithms are applied.

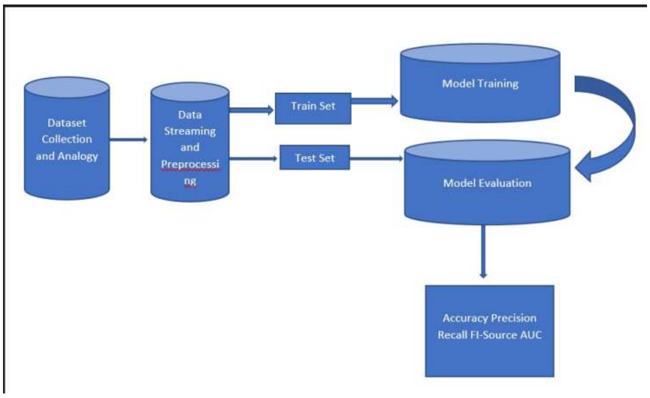

Fig2: Methodology

Machine Learning Models: We intend to use Supervised Machine Learning[8] models to classify Hazardous Asteroids and Non-Hazardous Asteroids. The following models are used:

- *Naive Bayes:* is one of the fast and easy Classification algorithms in Machine Learning which is used to predict a class of datasets. Both binary and multi-class classifications can be done using it.

Types of Naïve Bayes Models are:

- *Gaussian:* It assumes all the features to follow a normal distribution. The predictor takes continuous values instead of discrete values.
- *Multinomial:* When the data is multinomially distributed, it is employed. It is mostly used to categorise document issues, indicating to which category a specific document belongs, such as politics, education, weather, etc.
- Word frequency is used by the classifier as a predictor.
- *Bernoulli:* Similar to a multinomial classifier, but with independent Boolean variables as predictor variables. such as determining whether a word is used or not in a document. For tasks involving document classification, this model is renowned. Here we use Gaussian model.

- *Logistic Regression:* A statistical model known as logistic regression allows one to predict categorical dependent variables from a given collection of independent variables. The probabilistic values are always between zero and one.[12]
- *Random Forest:* In ML, it is applied to problems involving both regression and classification. To increase the dataset's predictive accuracy, it uses a number of decision trees on various subsets of the given dataset and averages the results[11][14].

- *Decision tree:* Although classification and regression issues can both be addressed, classification issues are usually resolved. It is a tree-structured classifier in which internal nodes represent dataset features, branches represent decision rules, and leaf nodes represent results.

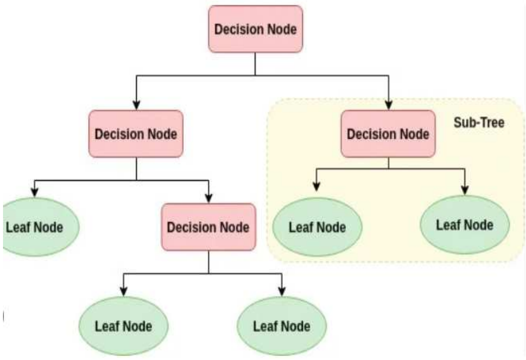

Fig3: Decision Tree

- *K-Nearest Neighbours* (KNN) is Classification Algorithm which classifies the data point on how its neighbour is classified[2][5].
- *Support Vector Machine:* The SVM algorithm's objective is to establish the best line or decision boundary that can divide n-dimensional space into classes, allowing us to quickly classify fresh data points in the future. A hyperplane is the name given to this optimal decision boundary.[13]

Accuracy: Accuracy is Calculated and Compared and best one should be noticed.
Precision: It counts the number of predictions from the positive class that are actually in that class.
Recall: It calculates how many positive class predictions were made using all of the dataset's positive examples.
F-Measure: It offers a single score that evenly weighs the issues of precision and recall.
Confusion Matrix: It is used to determine the classification models performance for a set of test data.

Data Collection: This data has been collected from the official website of Jet Propulsion Labouratory of California Institute of Technology which is an Organization under NASA, and maintained by them. Dataset contains 4688 rows and 40 features.
The columns consists of orbital details of the asteroid like name, semi-major axis(a), eccentricity(e), absolute magnitude, orbiting body, orbit id, EST dia in km, est dia in m, close approach date, minimum orbit intersection, relative velocity km per hr, miles per hour, orbit determination date, orbit uncertainty, relative velocity km per sec, accentricity, Orbital period etc features are there in the data set.The whole dataset[9] is available in official website of NASA's Jet Propulsion Laboratory.

Neo Refer	Name	Absolute N	Est Dia in K	Est Dia in K	Est Dia in M
3703080	3703080	21.6	0.12722	0.284472	127.2199
3723955	3723955	21.3	0.146068	0.326618	146.068
2446862	2446862	20.3	0.231502	0.517654	231.5021
3092506	3092506	27.4	0.008801	0.019681	8.801465
3514799	3514799	21.6	0.12722	0.284472	127.2199
3671135	3671135	19.6	0.319562	0.714562	319.5619
2495323	2495323	19.6	0.319562	0.714562	319.5619
2153315	2153315	19.2	0.384198	0.859093	384.1979
2162463	2162463	17.8	0.732074	1.636967	732.074
2306383	2306383	21.5	0.133216	0.297879	133.2156

Fig4: Dataset in Excel sheet

Data Visualization: Here, for data visualization purpose we mainly use heat map, which is done by the help of the pearson correlation coefficient. From this we can easily identify the highly correlated columns and take necessary steps to remove the dependency. After checking of duplicate values and missing values, we found that the count of hazardous and non-hazardous asteroids. There are only 755 asteroids are hazardous while remaining 3932 are non-hazardous. So, here we need to delete redundant features correlated with each other and after checking of duplicate values and missing values, we need to proceed for Feature Selection.

Features selection: For this classification selection of 16 features. They are absolute magnitude, Est Dia in M(min), Miles per hour, orbit uncertainty, Jupiter Tisserand Invarian, epoch Osculation, eccentricity, minimum orbit intersection, Semi Major Axis, Inclination, Aphekion Dist, Perihelion Time, Mean Anomaly, Mean Motion, label , Asc Node Longitude, Orbital Period, Perihelion Distance, Perihelion Arg,. These are the important features. Here the features are ordered from high to less important. Some of the less important features are also depended on the other highly important and independent features.
Here, some insignificant features need to be identified and removed before splitting training and testing dataset.

Split dataset: Then, dataset is divided into test and train set. Test set has 938 data while training set comprise of 3749 data. We are precisely identifying baseline accuracy and trying to create split dataset as dataset is imbalanced. Now the whole dataset is suitable for using the machine learning techniques. Splitting is done in this by taking the random state value as 10.

Techniques and their accuracy: For implementing python has been used. Because it has many libraries such as numpy, scikit-learn, pandas etc. Here used matplotlib, seaborn, os and xgboost with the above-mentioned three libraries. Here choose eight machine learning techniques[6]. They are logistic regression, K nearest neighbor[7], Support Vector Machine, decision tree, random forest, Naïve bayes, adaboost and xgboost method[3].
From this table we can see that random forest (with the number of tree is 15) and xgbclassifier can classify this hazardous and non-hazardous asteroid most accurately, but between them random forest takes minimum training time. Random forest takes the time 0.174722 ms, while XGBclassifier takes 2.400512 ms. So, random forest with

n_estimators of 15 should be the better approach to classify the asteroids.

Implementation: To implement this model python programming language is used. The libraries used here are NumPy and Pandas for data analysis, sklearn for pre-processing, importing matplotlib and seaborn for data visualization,
It can be implemented in Google Colab.
At first we import all the necessary libraries.
Then load the dataset.

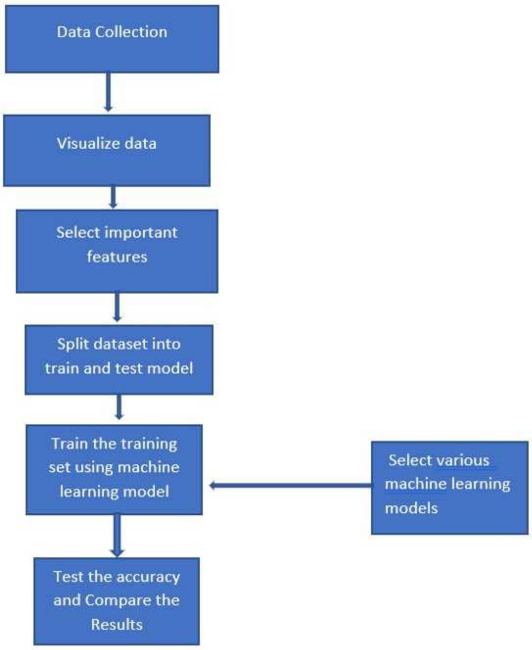

Fig5: Flow Chart for Implementation

The implementation can be done in following steps:
Step1: Data Prepocessing: It is the processing of identifying the missing values, null values, unknown values, noisy data, and other inconsistent data- and cleaned before using it.
We graphically represent the data by using heat maps, that uses color coded system. The main aim is to visualize the volume of locations/events in a better way assist in directing viewers towards areas on data visualizations that matter most.

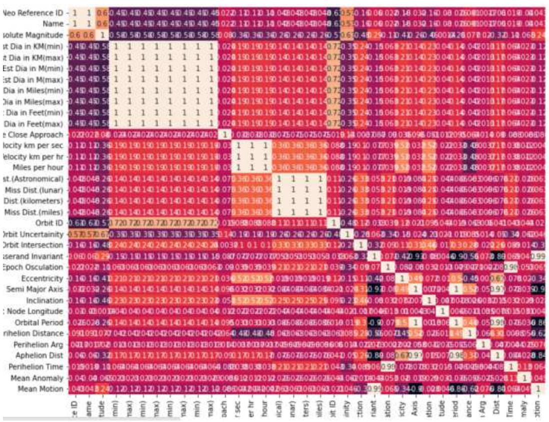

Fig6: Heat Map for all features

Step2: Data Preparation
Here we extract only the necessary data attributes in the data set. So that all the attributes which are having some noise, null etc. data will be unselected. And hence the new data is suitable for further process and analysis.

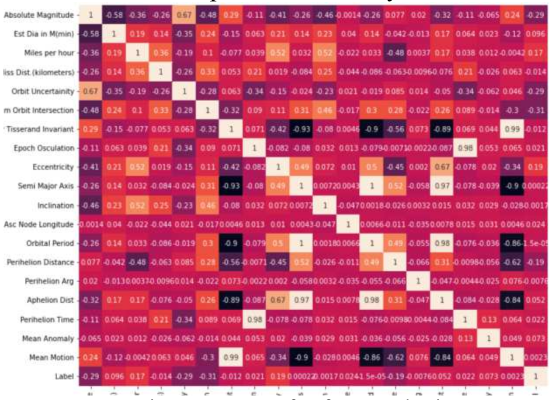

Fig7: Heat Map after feature selection

Step3: Training and Testing the Model
After selecting the necessary data, we use some data for training and some data for testing part.
Step4: Applying algorithms on the dataset.

I) Naïve bayes algorithm:

By applying this algorithm we got the following values[4]

```
Accuracy of train: 0.9478658536585366
Accuracy of test: 0.9573560767590619
```

Fig8: Accuracy of training and test dataset

```
Precision = 0.9194246373248094
Recall = 0.93061776479498
Accuracy = 0.9573560767590619
F1 Score = 0.9248948421030156
```

Fig9: Precision, Recall, Accuracy, F1 Score of Naïve bayes algorithm

Confusion Matrix: Here 1347 are correct predictions and 60 are incorrect predictions.

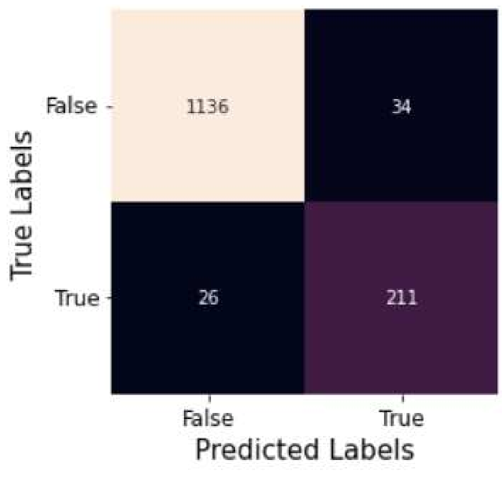

Fig10: Confusion Matrix

II) Logistic Regression:

By applying this algorithm we got the following value:

```
Accuracy of train: 0.9524390243902439
Accuracy of test: 0.9530916844349681
```

Fig11: Accuracy of training and test dataset

```
Precision = 0.9209766386666766
Recall = 0.9095477658768798
Accuracy = 0.9530916844349681
F1 Score = 0.9151300348758142
```

Fig12: Precision, Recall, Accuracy, F1 Score of Logistic Regression

Confusion Matrix: Here 1341 are correct predictions , 66 are incorrect predictions.

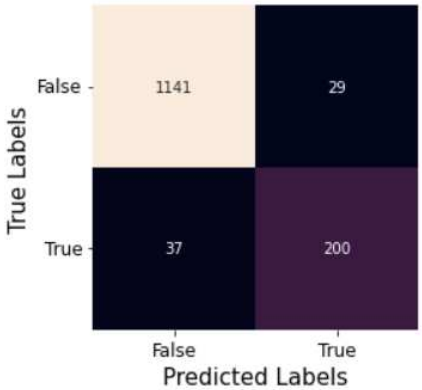

Fig13: Confusion Matrix

III) Random Forest

By applying this algorithm we got the following values[15]

```
Accuracy of train: 1.0
Accuracy of test: 0.9992892679459844
```

Fig14: Accuracy of training and test dataset

```
Precision = 0.9978991596638656
Recall = 0.9995726495726496
Accuracy = 0.9992892679459844
F1 Score = 0.9987336018541437
```

Fig15: Precision, Recall, Accuracy, F1 Score of Random Forest

Confusion Matrix: Here 1406 are correct predictions , 1 is incorrect predictions

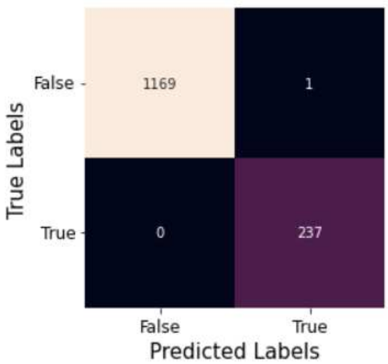

Fig16: Confusion Matrix

IV) Decision Tree

By applying this algorithm we got the following values

```
Accuracy of train: 1.0
Accuracy of test: 0.992181947405828
```

Fig17: Accuracy of training and test dataset

```
Precision = 0.9821835207320966
Recall = 0.9902520826571459
Accuracy = 0.992181947405828
F1 Score = 0.9861622849172751
```

Fig18: Precision, Recall, Accuracy, F1 Score of Decision Tree

Confusion Matrix: Here 1396 are correct predictions, 11 are incorrect predictions

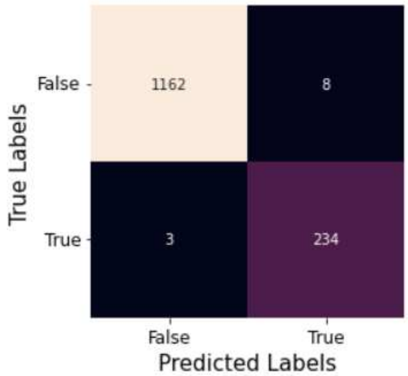

Fig19: Confusion Matrix

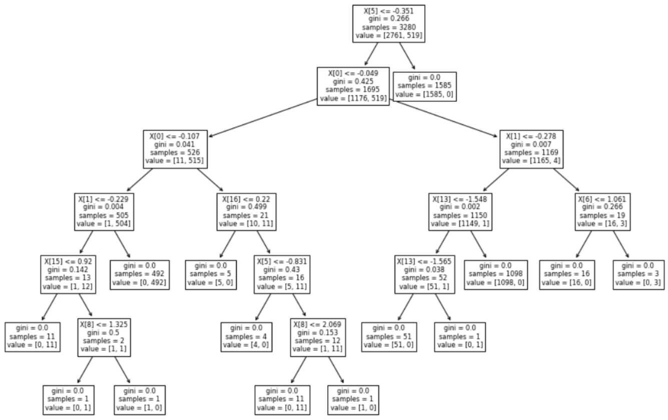

Fig20: Decision Tree

V) KNN

By applying this algorithm we got the following values

```
Accuracy of train: 0.911890243902439
Accuracy of test: 0.8997867803837953
```

Fig21: Accuracy of training and test dataset

```
Precision = 0.868388319772305
Recall = 0.7445905009196148
Accuracy = 0.8997867803837953
F1 Score = 0.7869264815970392
```

Fig22: Precision, Recall, Accuracy, F1 Score of KNN

Confusion Matrix: Here 1266 are correct predictions and 141 are incorrect predictions.

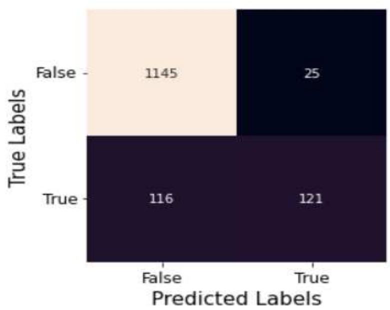

Fig23: Confusion Matrix

VI) Support Vector Machine

By applying this algorithm we got the following values

```
Accuracy of train: 0.9634146341463414
Accuracy of test: 0.9481165600568585
```

Fig24: Accuracy of training and test dataset

```
Precision = 0.9243525480367585
Recall = 0.8846857081034296
Accuracy = 0.9481165600568585
F1 Score = 0.9029241556865514
```

Fig25: Precision, Recall, Accuracy, F1 Score of Logistic Regression

Confusion Matrix: Here 1334 are correct predictions and 73 are incorrect predictions.

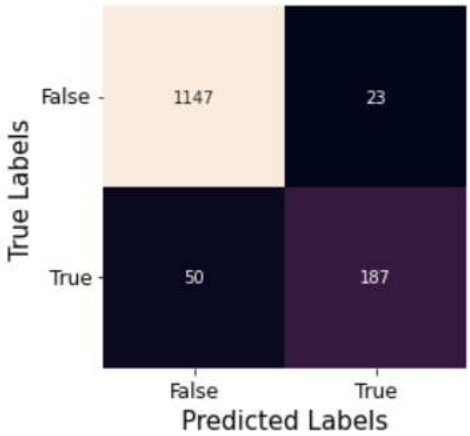

Fig26: Confusion Matrix

RESULT AND ANALYSIS

```
Naive_Bayes_accuracy           0.957356
Logistic_Regression_accuracy   0.953092
Random_Forest_accuracy         0.999289
Decision_Tree_accuracy         0.992182
KNN_accuracy                   0.899787
SVM_accuracy                   0.948117
dtype: float64
```

Fig27: Comparision

By comparing all the accuracies we found Random Forest gives the most accuracy value, then after Decision Tree accuracy followed by Naïve Bayes, Logistic Regression, SVM, KNN accuracies.

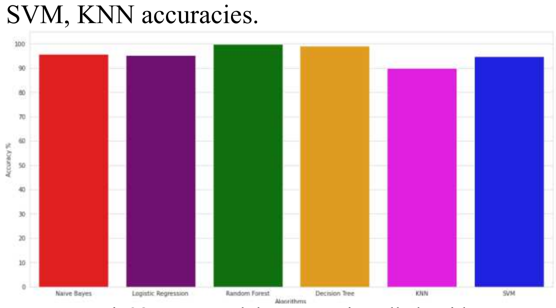

Fig28: Bar graph by comparing all algorithms

CONCLUSION

In this paper, classification of non-hazardous and hazardous has been done through various machine learning techniques. After using many methods, it has been seen that a random forest tree gives the best optimal solution interms of training time and accuracy. So, this paper will help to identify the newly discovered near earth asteroids if that is hazardous or not by finding their accuracies.

REFERENCES

[1] Vadym Pasko, Prediction of Orbital Parameters for Undiscovered Potentially Hazardous Asteroids Using Machine Learning, Stardust Final Conference. Advances in Asteroids and Space Debris Engineering and Science. Edited by Massimiliano Vasile, Edmondo Minisci, Leopold Summerer and Peter McGinty. Astrophysics and Space Science Proceedings, Vol. 52. Berlin: Springer Verlag, 2018

[2] M. Popescu, J. Licandro, J.M. Carvano, R. Stoicescu, D. Morate, Taxonomic classification of asteroids based on the Moves near Infrared-Color, A&A Volume 617, 12 September 2018

[3] Nunget, Carry, Machine learning and next generation asteroid surveys, American Astronomical Society, DPS meeting #49, id.103.03, oct 2017

[4] John D. Hefele, Francesco Bortolussi and Simon Portegies Zwart, Identifying Earth-impacting asteroids using an Naiive Bayes, Atronomy & Astropysics, Volume 634, February 2020

[5] Rkmt Nishavi Ranaweera, Tgi Fernando, Prediction of Potentially Hazardous Asteroids using Deep Learning, 2nd International conference on Advanced Research in Computing (ICARC), Feb 2022

[6] Baburao Markapudi, Kavitha Chaduvula, D.N.V.S.L.S. Indira & Meduri V. N. S. S. R. K. Sai Somayajulu, "Content-based video recommendation system (CBVRS): a novel approach to predict videos using multilayer feed forward neural network and Monte Carlo sampling method", published in the journal of Multimedia Tools and Applications (Springer Nature), published on 11 th August 2022. PP:1-27

[7] D.N.V.S.L.S. Indira, Babu Rao Markapudi, Kavitha Chaduvula, Rathna Jyothi Chaduvula, "Visual and buying sequence features-based product image recommendation using optimization based deep residual network", published in the journal of Gene Expression Patterns, Volume 45, 2022, September 2022

[8] V. Carruba , S. Aljbaae and A. Lucchini , Machine-learning identification of asteroid groups, Advance Access, 1 July 2019

[9] Sirnov, Markov, Identification of asteroid trapped inside three- body motion resonance: a machine learning technique, Monthly Notices of the Royal Astronomical Society, Volume 469, Issue 2, August 2017.

[10] Anish Si, Hazardous Asteroid classification through various Machine Learning Techniques, International Research Journal of Engineering and Technology, Mar 2020

[11] Spectral Classification of Asteroids by Random Forest, Huang Chao, Chinese Astronomy and Astrophysics, 06 oct 2017

[12] LaValley, M. P. (2008). Logistic regression. *Circulation*, 117(18), 2395-2399.

[13] Suthaharan, S. (2016). Support vector machine. Machine learning models and algorithms for big data classification, pp. 207-235.

[14] Scikit Learn (2022). sklearn.ensemble.RandomForestClassifier.

[15] Pramoditha, R. (2022). Why do we set a random state in machine learning models?. tds.