**Machine Learning Essentials Project**

**Name :** [Harisudhan Ramaswamy](mailto:125018027@sastra.ac.in)

**Register number : 125018027**

**Abstract :**

The project focuses on the detection and classification of asteroids using machine learning techniques. The primary objective is to develop models that can accurately classify asteroids based on their characteristics and orbital parameters. Datasets from Kaggle are being put to use, including the NASA Asteroids Classification dataset. The classification model will be implemented using Scikit-learn. The results indicate a satisfactory accuracy rate, and f1 score demonstrating the use of machine learning in planetary defense and asteroid monitoring.

**Introduction :**

Asteroids come under the spectrum of near earth objects and can pose a significant threat to Earth, with the potential for catastrophic impacts. NASA has been able to identify a number of near-Earth objects (NEOs).

Detecting and Classifying these Asteroids can help predict future catastrophes, help understand better about the universe, and also build our understanding about various machine learning models.

**Project objectives :**

* To create a machine learning model that can identify and categorize asteroids according to their orbital parameters and properties(physical).
* To evaluate and get asteroid datasets for model training by preprocessing and analyzing them from a variety of sources, like the NASA JPL Asteroid Database and datasets from Kaggle.
* To identify the best method for classifying these asteroids, using a variety of machine learning methods. Models including support vector machines, random forests, and decision trees will be put into practice and evaluated.
* To learn more about the key characteristics needed for the classification of asteroids by using methods such as feature significance analysis.
* To guarantee dependable asteroid detection, evaluate the performance of the models that have been trained using performance measures including accuracy, precision, recall, and F1 score.

**Methodology** :

**Selecting Datasets :**

* <https://www.kaggle.com/datasets/sakhawat18/asteroid-dataset/data>
* <https://www.kaggle.com/datasets/shrutimehta/nasa-asteroids-classification>
* <https://www.kaggle.com/code/suhanshubhattacharya/asteroid-classification-on-nearearthobjects/input>
* <https://www.kaggle.com/code/ahmedhassansaqr/dangerous-asteroids-detection-f1-score-97/input>

These datasets contain various attributes, including size, shape, orbital parameters, and classification labels.

Final dataset chosen :

[**https://www.kaggle.com/datasets/sakhawat18/asteroid-dataset**](https://www.kaggle.com/datasets/sakhawat18/asteroid-dataset)

**Machine learning models :**

* Logistic regression
* Decision Trees
* Random forest
* Convolutional Neural Network

**Logistic Regression**

Logistic regression is a statistical model used for binary classification problems. It predicts the probability of an event occurring based on a set of predictor variables. The model uses a logistic function to map the linear combination of predictor variables to a probability between 0 and 1.

**Random Forest**

Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. Each decision tree is trained on a random subset of the data and features, and the final prediction is made by aggregating the predictions from all trees. This approach helps to reduce overfitting and improve generalization performance.

**Decision Trees**

Decision Trees are a machine learning algorithm used for both classification and regression tasks. They create a tree-like model where each internal node represents a test on a feature, each branch represents a possible outcome of the test, and each leaf node represents a prediction. The algorithm recursively splits the data based on the features that best separate the classes or predict the target variable.

**Convolutional Neural network**

Convolutional Neural Networks (CNNs) are a type of deep learning architecture specifically designed for processing and analyzing image data. They are inspired by the human visual cortex and use filters to extract features from images, making them highly effective for tasks like image classification, object detection, and medical image analysis.

**Processes :**

Exploratory data analysis

Data preprocessing - Normalization and removing missing values.

Splitting data to test and train.

Training the model using training data.

Testing the model using test data.

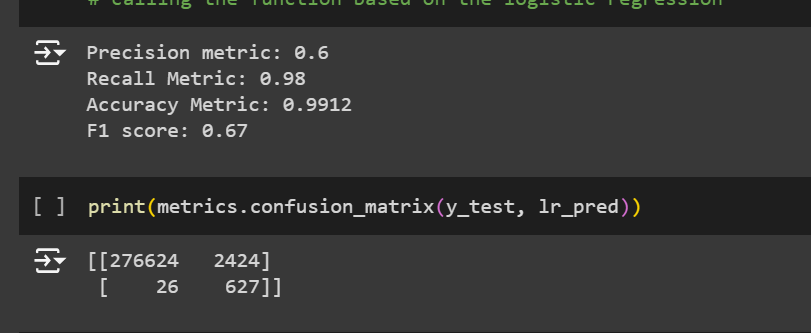
Performance evaluation - Accuracy, f1 score, precision, recall.

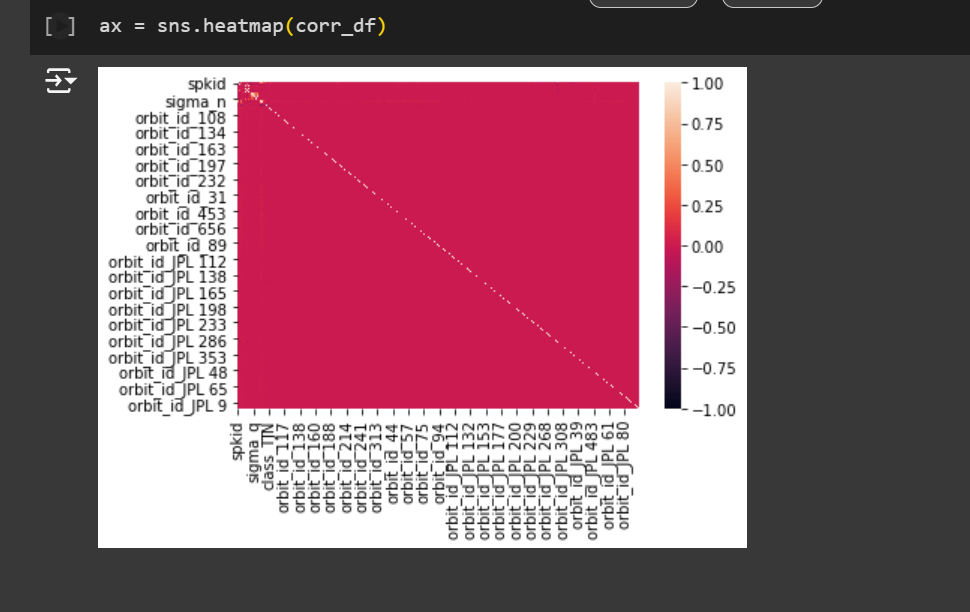
**Workflow :**



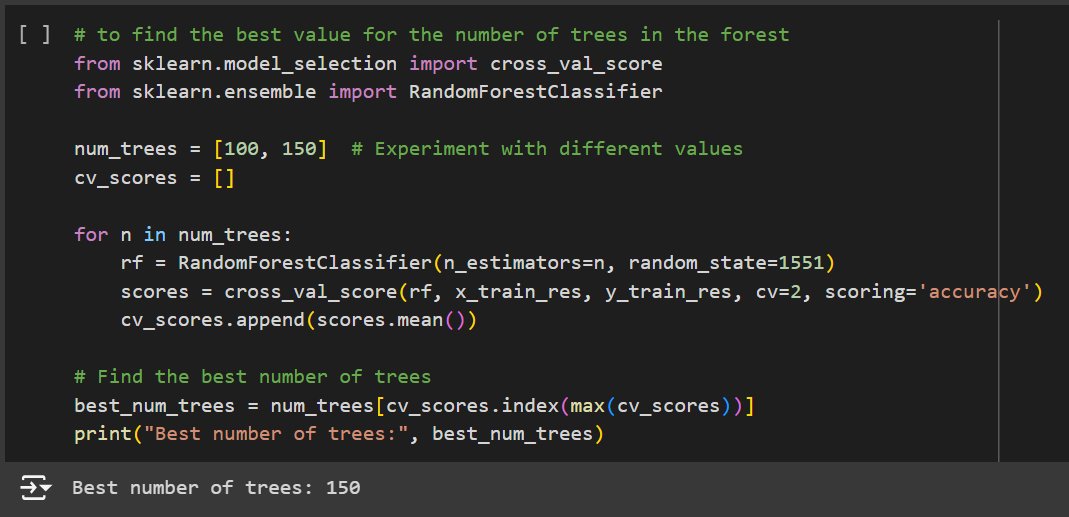
**Results :**

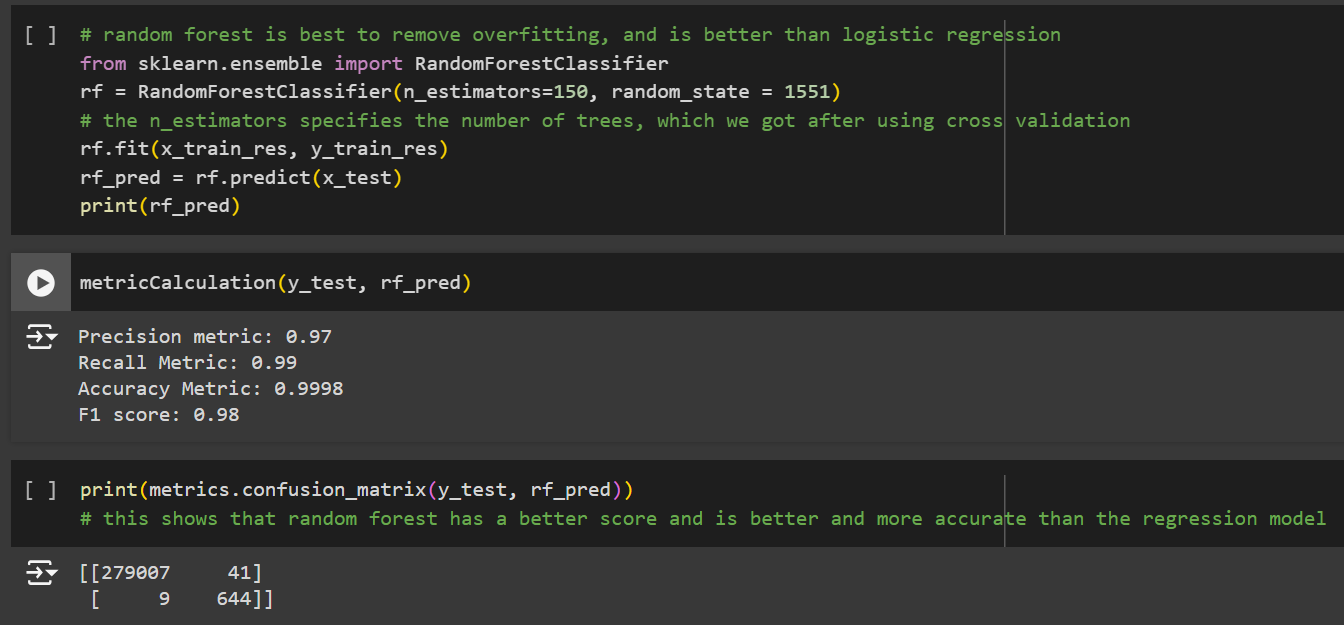
**Logistic regression results :**





**Random forest results:**





**Observations and Additional Results :**

* **Initial class imbalance:** The dataset was heavily skewed towards the 'Y' class, with significantly fewer samples in the 'N' class.

Before OverSampling, counts of label 'N': 651221

Before OverSampling, counts of label 'Y': 1413

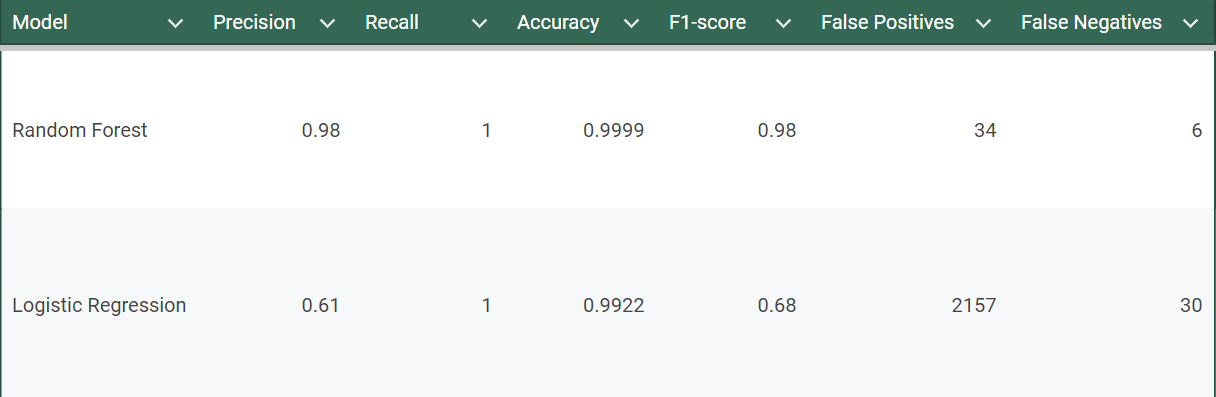
* **Successful oversampling**: SMOTE effectively balanced the class distribution by increasing the number of samples in the minority class ('N').
* **Balanced dataset:** After oversampling, the dataset had an equal number of samples for both classes.

After OverSampling, counts of label 'N': 651221

After OverSampling, counts of label 'Y': 651221

* **Potential improvement in model performance:** The balanced dataset may lead to improved performance on the minority class, reducing the risk of bias towards the majority class.
* **Convolutional layers:** Effectively capture spatial patterns in asteroid data, making them suitable for analyzing the geometric and structural characteristics of asteroids.
* **Pooling layers:** Downsize feature maps, reducing computational cost and preventing overfitting, while preserving essential information about the asteroid's features.
* **Dense layers:** Combine the extracted features into a single output, allowing the model to make informed predictions about the likelihood of an asteroid being potentially hazardous.
* **Activation functions:** Introduce non-linearity, enabling the model to learn complex relationships between features that cannot be represented by linear models.
* **Output layer:** The sigmoid activation function in the output layer represents the likelihood of an asteroid being classified as potentially hazardous.

**Model Performance Comparison :**



## **Learning Outcomes**

**Skills Used:**

* **Machine Learning:**
  + Model selection and training (logistic regression, random forest, deep learning)
  + Hyperparameter tuning
  + Model evaluation and interpretation
* **Data Analysis:**
  + Data cleaning and preprocessing
  + Feature engineering
  + SMOTE
  + Exploratory data analysis
* **Programming:**
  + Python programming proficiency
  + Use of libraries like pandas, NumPy, Scikit-learn, TensorFlow/Keras
* **Problem-Solving:**
  + Ability to identify and address challenges in the dataset and model development

**Tools Used:**

* Python
* Jupyter Notebook
* Pandas
* NumPy
* Scikit-learn
* TensorFlow/Keras (optional, depending on the deep learning model)

**Dataset Used:**

<https://www.kaggle.com/datasets/sakhawat18/asteroid-dataset>

**Topics Learned:**

* Asteroid detection and classification
* Machine learning algorithms for binary classification
* Overcoming class imbalance issues (using SMOTE)
* Deep learning for image or tabular data (depending on your data format)
* Model evaluation and interpretation
* Data preprocessing and feature engineering techniques

**GitHublink:** <https://github.com/myserendipityinsolitude/Machine-Learning-End-Sem>