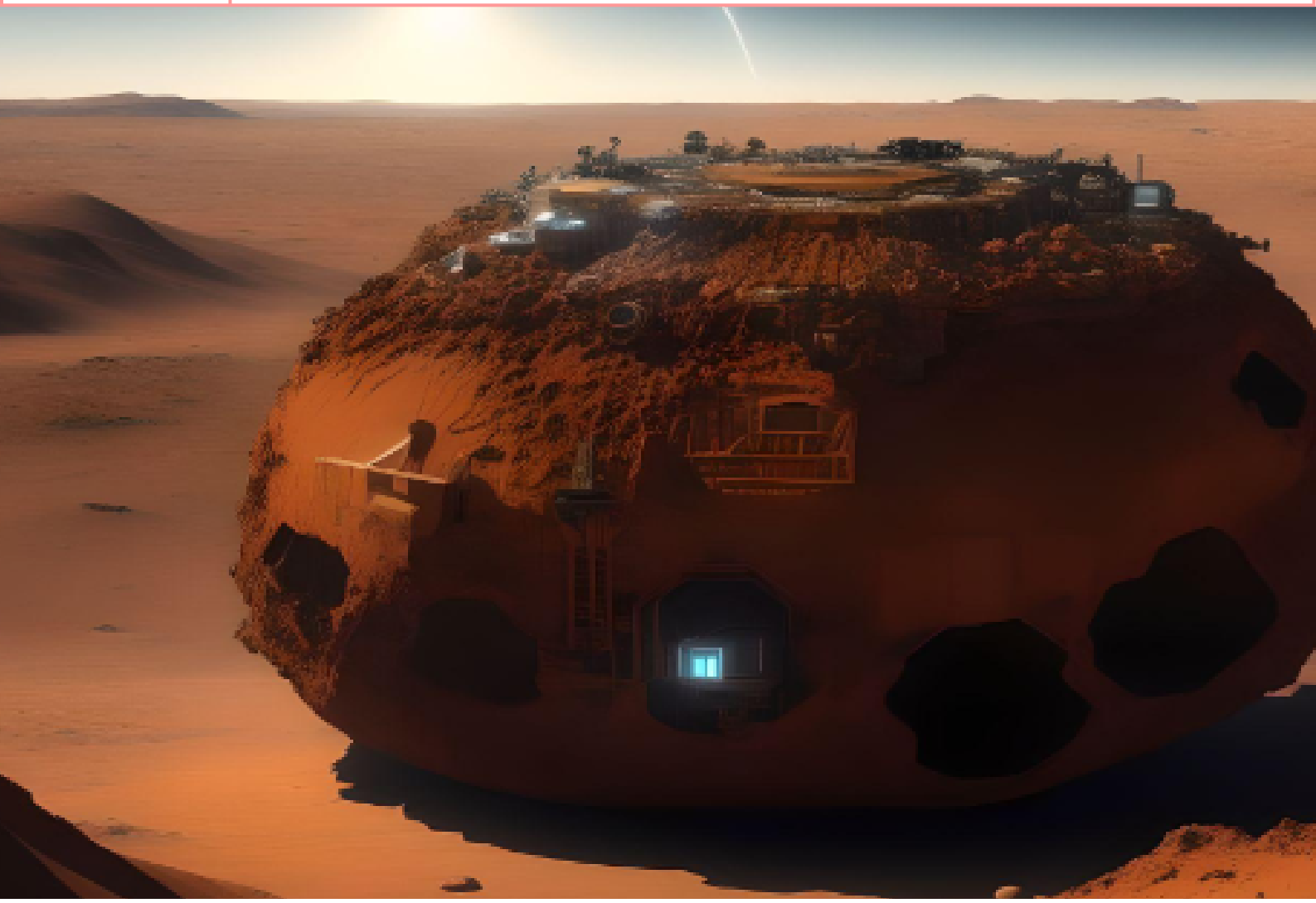
COSMOCOPE REPORT

Team :

Sl.no	NAME
1	JUNAID
2	D.MAHESH
3	A.V.S.SREENIVASU
4	K.GANESH KUMAR



ASTRO ANALYTICS

TABLE OF CONTENTS

INTRODUCTION

- In order to complete this challenge, we must develop a model that will automatically analyse data from GCMS (gas chromatography-mass spectrometry) in order to find nine families of chemical substances that are important to understanding Mars' potential for previous habitability.
- SVM (supporting vector machine) is used to solve the problem; from the outputs of the training model, we created output for var features.Ploynomial regression is carried out for mass and intensity at a particular time .We used the train test split approach and the SVC (Support Vector Classifier) to predict the results from the given train features output.Log loss is also calculated to know the efficiency of result obtained.
- Libraries Used:

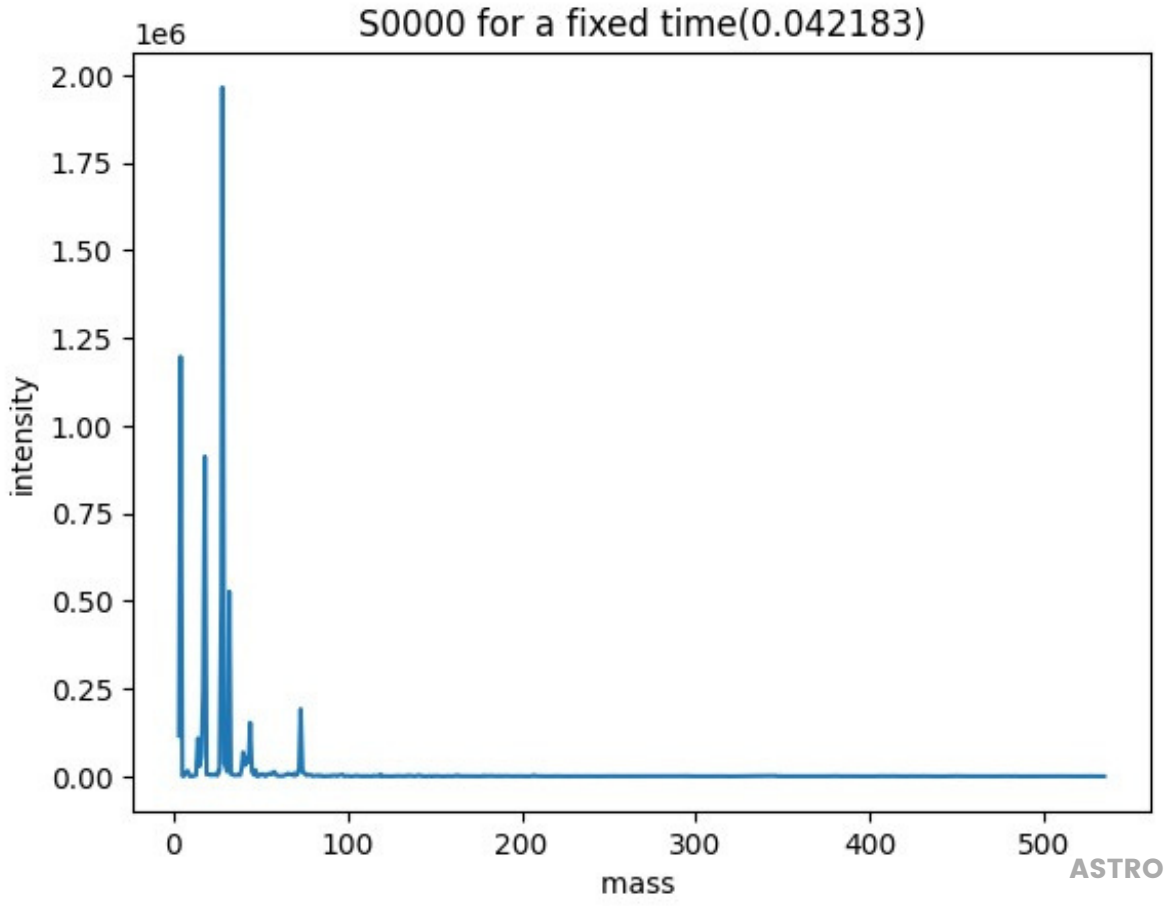
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import math
import random
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import log_loss
from sklearn.metrics import r2_score
```

DATA PROCESSING

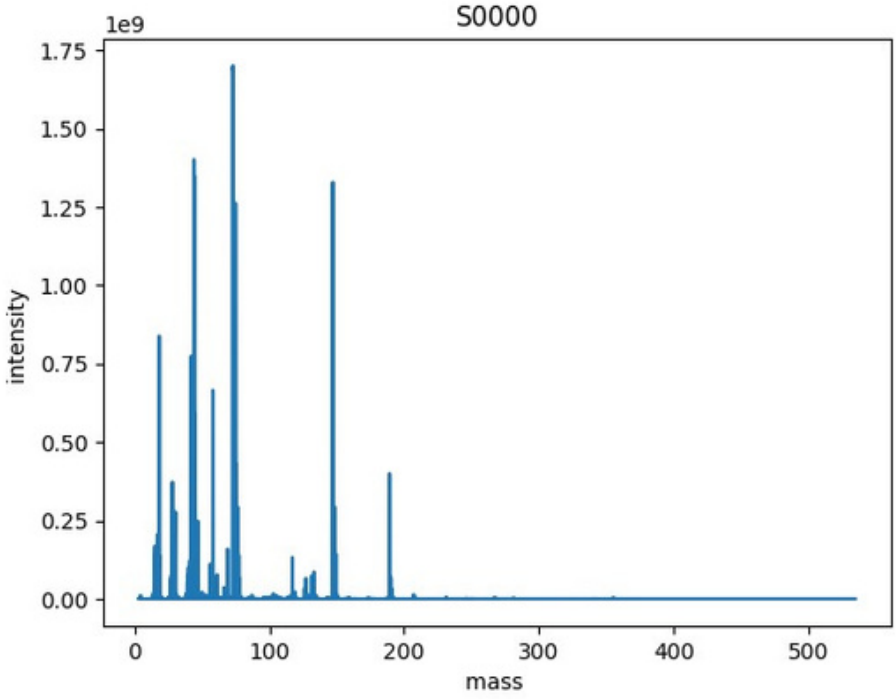
- We normalised the intensity and mass values because the raw data's high standard deviation makes it difficult for the model to understand the characteristics of the data.

- We used ploynomial regression for intensity and mass records at every fixed time to obtain co-efficient parameters and an intercept for a sample with intensity, mass, and time .

- Graph:

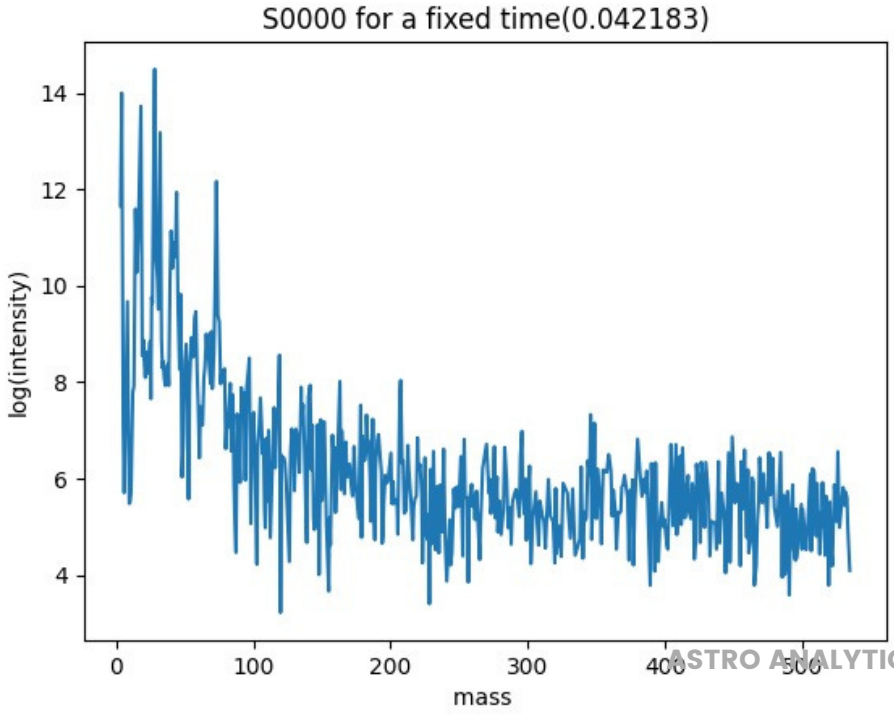


- From various fixed times, we deduced a set of parameters, and we mean those parameters to provide the final (coefficient, intercept) parameters for a sample of data.
- For more accuracy we applied regression to log(intensity) and mass.
- Graphs:



Graph for one train feature sample data

Graph with log(intensity) and mass for a sample.



- The process is continued for all train feature samples and parameters stored in an list.Same procedure is done for var features to obtained list of parameters of each sample in it.

- From the train_labels.csv with sample id's and parameters list that obtained from samples in train features ,we made a train test split and applied svc method to predict result for parameters list of samples in var features. this is process done in model1.

- model2 is the updated version of model in which mean of all fixed time mass,intensity is taken from function idea2 in model2.py.

- model3 is the updated version of model2 in which we oversampled the data to get good results.

RESULT

- With svc method we fit the trained data and predicted for parameters list of samples in var features.

- From the result we created a csv file which contains information about the presence of various gases in each sample in var features.

- To check the efficeincy, we applied log loss and we obtained value as 2.8923919386328585.