# 

**CSE523 : Machine Learning**

Winter 2021 - 2022

**Weekly Report - 5**

Dt : 16-03-2022

**Group Name : Discover Decipher**

**Group Members**

Nimisha Patel - AU1940146 - Btech Computer Science and Engineering.

Sakshi Shah - AU1940213 - Btech Computer Science and Engineering.

Astha Patel - AU1940312 - Btech Computer Science and Engineering.

Kareena Matwani - AU1940314 - Btech Computer Science and Engineering.

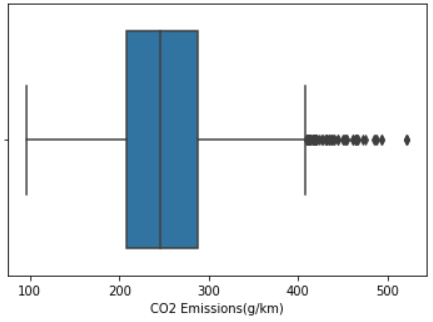
**Task Performed this week**

**Prepared the classes for the prediction of CO2 emission using classification.**

The accuracy from linear regression and multivariate linear regression was around 0.89 which shows non accurate prediction of CO2 emission value by regression. Thus, before implementation of other algorithms like simple vector machines and random forest. We would like to predict and analyze the likelihood that the data would fall into the predetermined categories.

The data obtained from source have a real integer value not the categorical classification, thus, firstly the classes (lower-upper limit) were created in a way that is equal and balanced distribution among each category class.

Know the CO2 Emission distribution, analyze the maps to develop the classification categories.



Very few data points below 100 and above 500(g/Km) CO2 Emission.

CO2 emission (in g/km)

Data count : 7385

Mean : 250.584699

Standard deviation : 58.522679

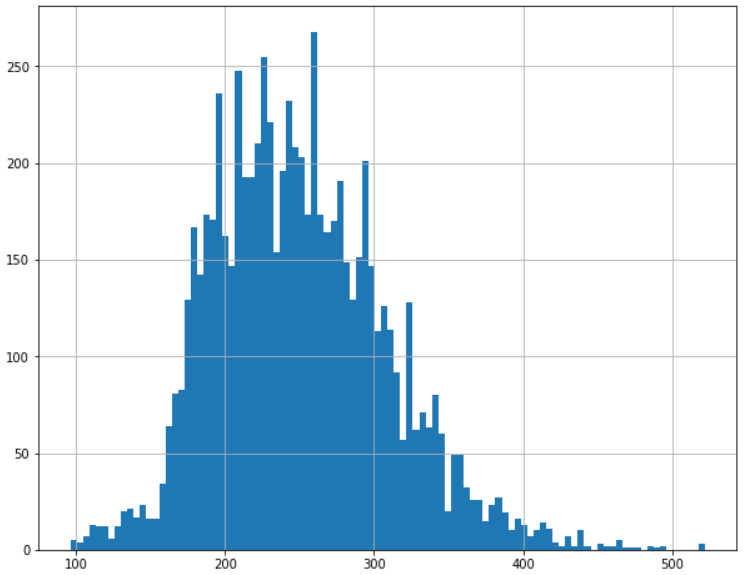
Minimum value : 96

Maximum value : 522

Maximum - Minimum = 522-96 = 426

Number of Bins : 4

Range approximately =



Range : 100

Class 0 : <= 100 : Low Emission Range

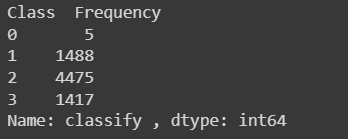
Class 1 : 100 - 200 : Permissible Emission Range

Class 2 : 200-300 : Medium Emission Range

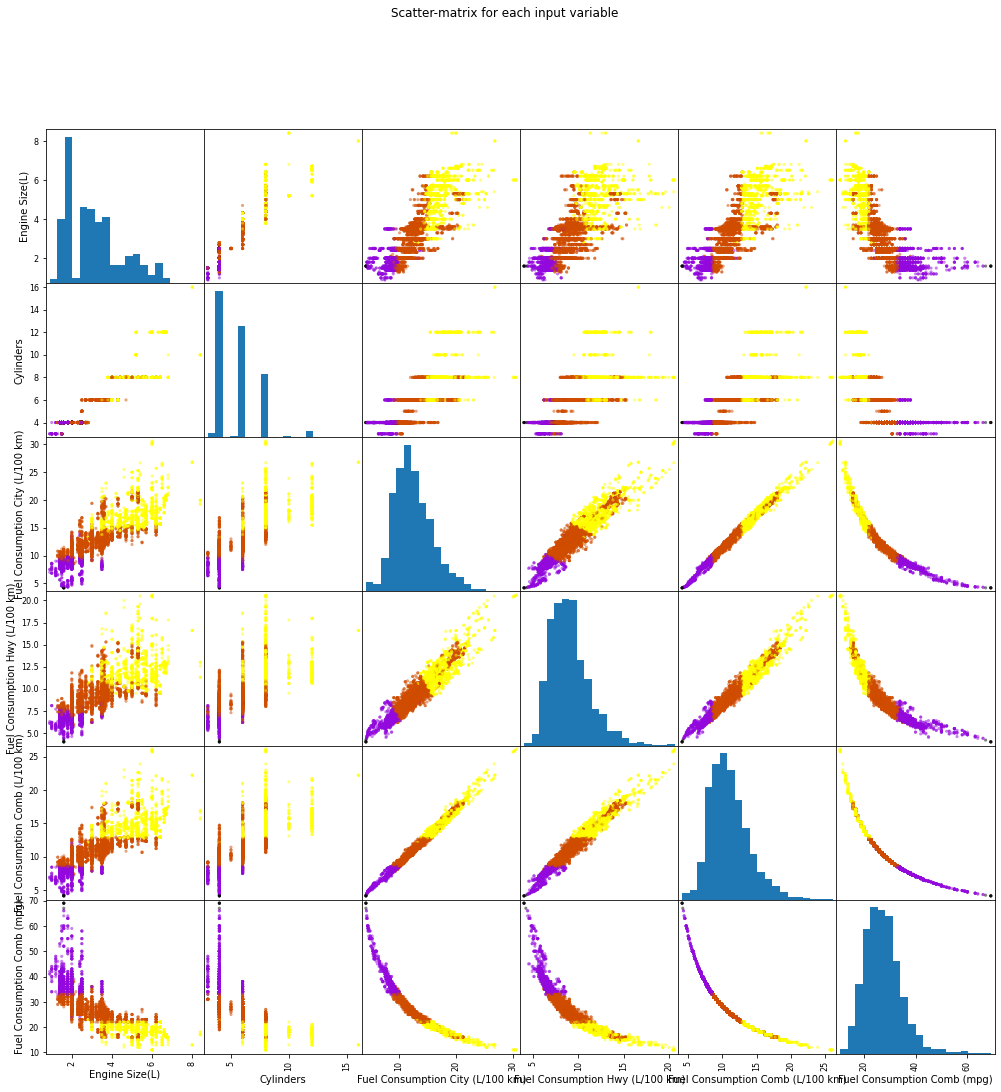
Class 3 : >=300 : High Emission Range (Not permissible)

The classes are based on the calculation as mentioned above using maximum and minimum values along with researched value which is <100 g/km is considered as low emissions and >= 300 g/km is considered as very high emission.[[1]](https://www.continental-automotive.com/getattachment/8f2dedad-b510-4672-a005-3156f77d1f85/EMISSIONBOOKLET%202019.pdf)

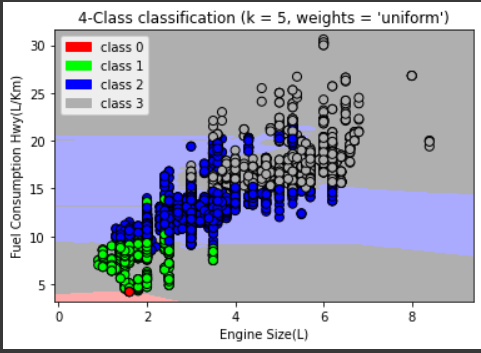
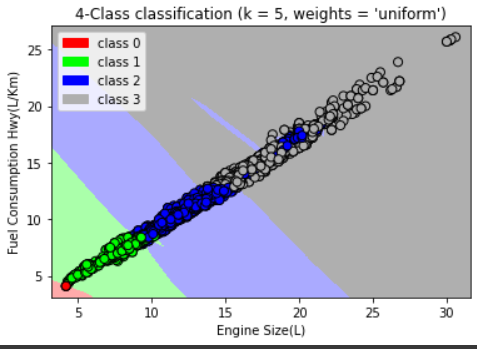
Frequency Distribution :



**Scatter Matrix for each feature to know the distribution.**



**Decision Boundary for KNN classifier for different variables**



**Why Polynomial Regression ?**

* As we saw in the last report that in single variable linear regression the value of r square is 0.59, r square basically depends on how the variable affects the model, so here we brought the concept of higher order regression.
* The Polynomial regression for degree 2 (Quadratic regression) and degree 3 was implemented in continuation to linear regression and multivariate linear regression.

**Quadratic regression:**

With the help of the Scikit-learn library, we took an independent variable - Engine size and second degree polynomial generates coefficients for degree = 0, degree = 1, degree = 2.

**Goal :** To predict CO2 emissions given engine size using polynomial regression.

**Formula :** y = b + θ1x + θ2x^2

Where θ1 and θ2 are coefficients. And b is the intercept.

We need to find the value of b, θ1 and θ2 such that the distance between each point (xi, yi) and the quadratic curve will be minimal.

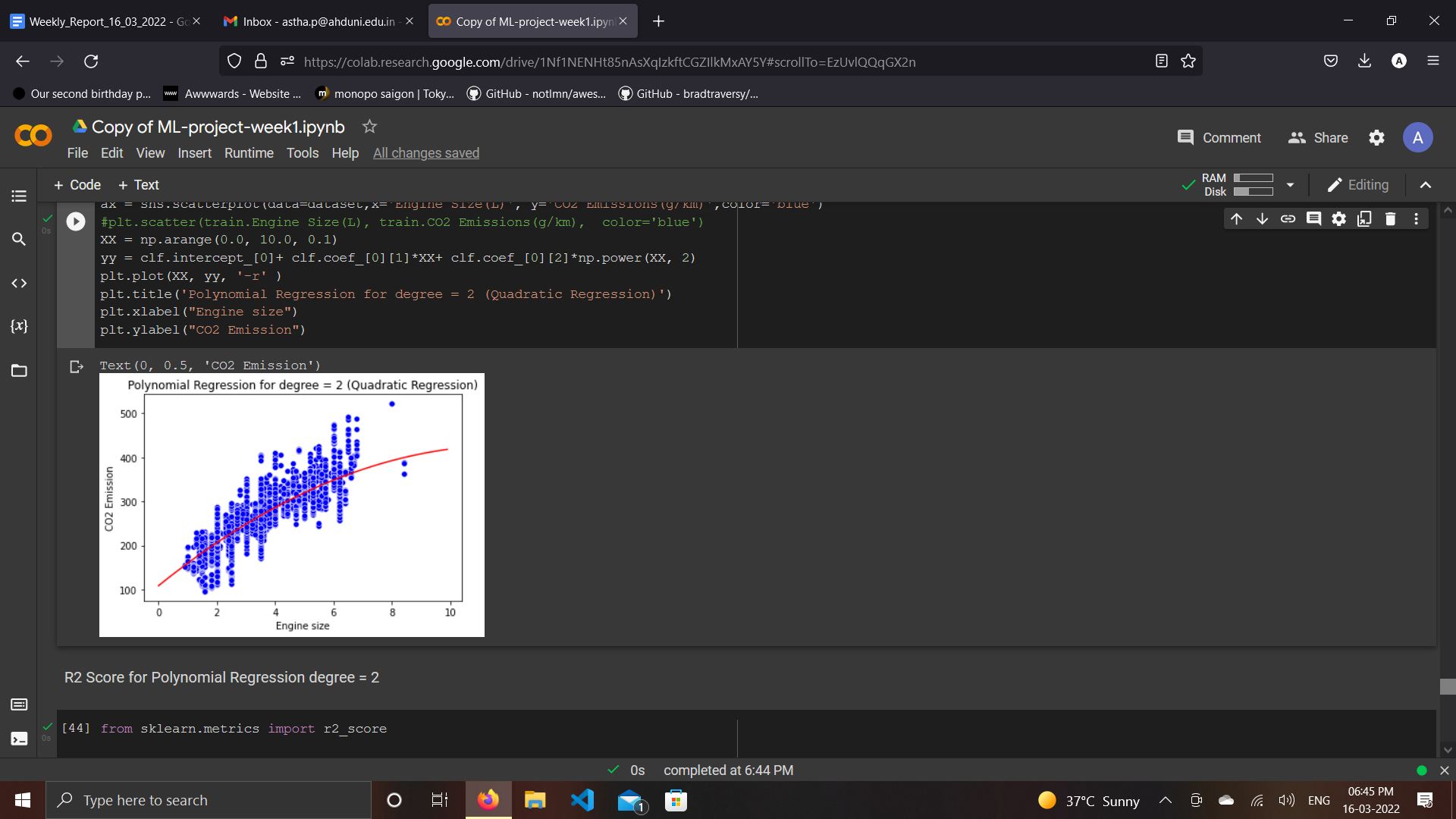
**Output:**

* We found the value of b, θ1 and θ2 using the function of the Sklearn library.

Here we got the value of θ1=53.33020824, the value of θ2=-2.22799823 and the value of b=108.69385134.



* Also we found the value of Mean absolute error: 22.26, Residual sum of squares (MSE): 876.34 and R2-score: 0.65 with the help of above coefficients and intercept value and then we plot the curve.



**Plot for Degree = 2**

**Polynomial regression For Degree 3 :**

With the help of the Scikit-learn library, we took the independent vector as Engine size and the third degree polynomial generates coefficients of degree = 0, degree = 1, degree = 2, degree=3.

**Goal :** To predict CO2 emissions given engine size using polynomial regression.

**Formula :** y = b + θ1x + θ2x^2 + θ3x^3

Where θ1, θ2 and θ3 are coefficients. And b is the intercept.

We need to find the value of b, θ1, θ2 and θ3 such that the distance between each point (xi, yi) and the 3 degree curve will be minimal.

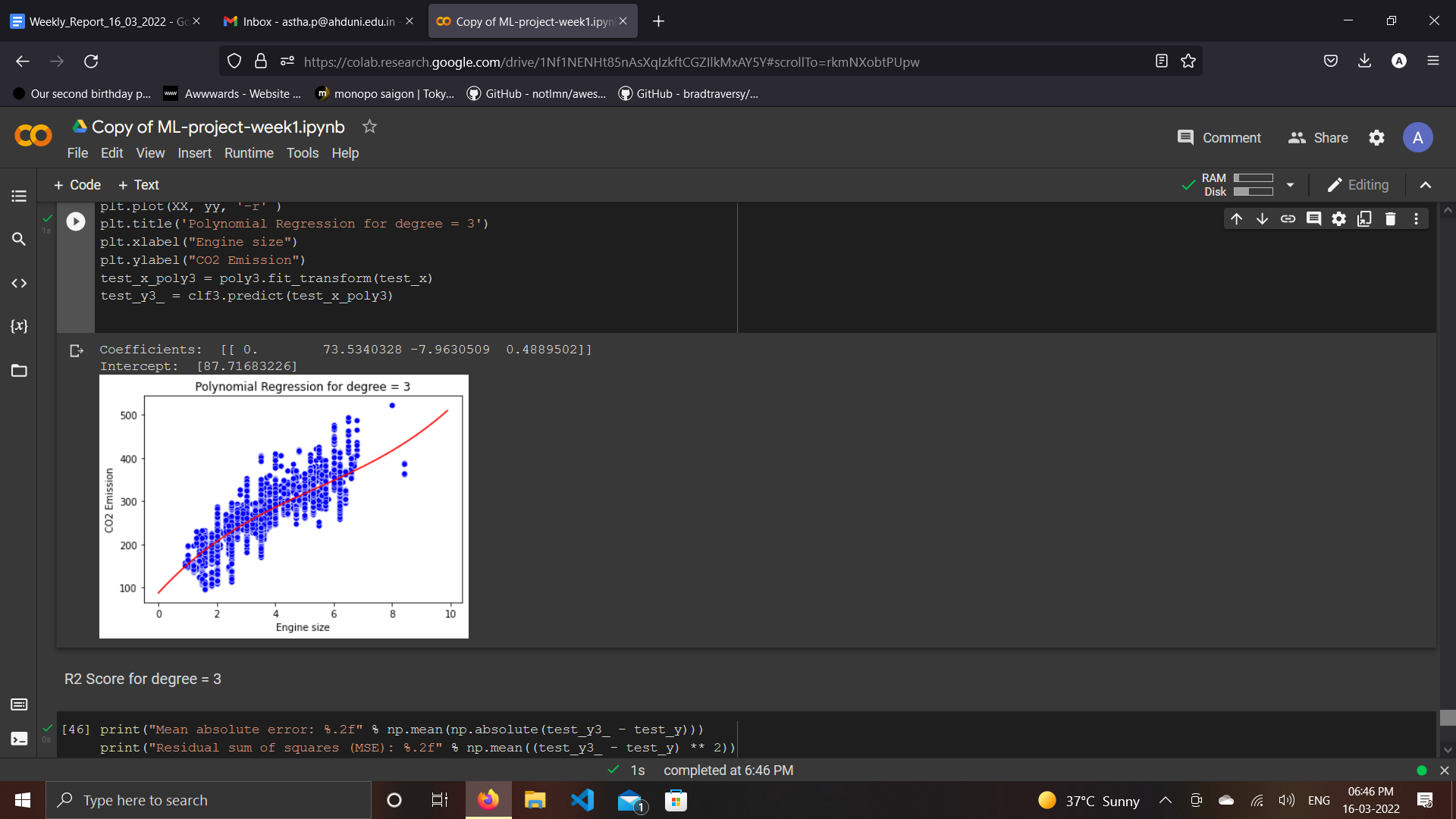
**Output:**

* We found the value of b, θ1, θ2 and θ3 using the function of the Sklearn library.

Here we got the value of θ1=73.5340328, the value of θ2=-7.9630509, the value of θ3=0.4889502 and the value of b=87.71683226.



* Also we found the value of Mean absolute error: 22.14 Residual sum of squares (MSE): 871.76 R2-score: 0.65 with the help of above coefficients and intercept value and then we plot the curve.



**Plot for Degree = 3**

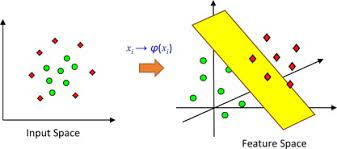
**Conclusion :**

* The R squared value in linear regression is 0.59 where in quadratic regression and polynomial regression for degree 3, it is 0.65 and in multivariable linear regression is 0.87.
* R squared value depends on the how a variable affects the modeland as the accuracy (R2 score) of the multivariable regression is the highest (87%) , it can be considered as best fitted model than linear , quadratic and polynomial regression.
* Many other algorithms / ML techniques can be used (we would be using in future) in order to increase the accuracy of the model.

**How will SVM (Support Vector Machine) help us in approaching the results?**

SVM can be used in CO2 prediction model by considering 4-5 variables having varied properties of existence and then the prediction at the risk of high or low emissions from varied types of cars can be taken into consideration, the final selection of features can be based on the discriminative classification analysis , based on data after we achieve clusters during classification process to have optimal output the data points can be led down to multidimensional space which can be segregated through a hyperplane for discrimnation in binary classes also.

Evaluations in test data sets and k-fold cross validation in training data can also be done to have optimed output predictions.



[Example](https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.sciencedirect.com%2Ftopics%2Fengineering%2Fhyperplanes&psig=AOvVaw3oO--ezZpAvjLWTPIilB8X&ust=1647525263576000&source=images&cd=vfe&ved=2ahUKEwja1IKC5Mr2AhWIRmwGHaARDM0Qr4kDegUIARDAAQ)

**Tasks to be Performed Next Week**

Implementation of SVM model and Decision tree model.