Summer Training Project Report

*Submitted in partial fulfillment of requirement for award of degree*

*Of*

**Bachelor of Technology**

In

**Computer Science and Engineering**

**By**

**Priyansh Singhal**

**03815602717**



**Department of Computer Science and Technology**

**Dr Akhilesh Das Gupta Institute of Technology and Management(Affiliated to GGSIPU, Delhi),**

**New Delhi - 110053**

**Candidate’s Declaration**

It is hereby certified that the work being presented in BTech Summer Training Project entitles “IMDB Sentimental Analysis using Neural Network” in partial fulfillment of requirement of award of degree of Bachelor of Technology and submitted in the Department of Computer Science & Engineering of Dr. Akhilesh Das Gupta Institute of Technology and Management, New Delhi(Affiliated to GGSIPU, Delhi) is an authentic record of our own work.

The matter presented in Summer Training Project Report has not been submitted by me for the award of any other degree of this or any institute.

**Priyansh Singhal**

**(03815602717)**

This is to certify that the above statement made by candidate is correct to the best of my knowledge. They are permitted to appear in Summer Training Project Examination.

**Mrs Uma Dr Saurabh Gupta**

**Cse Dept Head, CSE**

**Signature of External Examiner**

**INTRODUCTION**

**Machine learning** is the subfield of Artificial Intelligence that provides computer the ability to learn without being programmed explicitly.

Machine learning algorithms build a mathematical model based on the sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in a wide variety of applications, such as spam detection and face detection, where it is difficult to develop a conventional algorithm for effectively performing the task.

Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning.

There are three types of Machine Learning Algorithm :

1. **Supervised** Machine Learning Algorithm
2. **Unsupervised** Machine Learning Algorithm
3. **Reinforcement** Machine Learning Algorithm

# Supervised learning

**Supervised learning** is the [machine learning](https://en.wikipedia.org/wiki/Machine_learning) task of learning a function that maps an input to an output based on example input-output pairs.It infers a function from *labeled* [*training data*](https://en.wikipedia.org/wiki/Training_set) consisting of a set of *training examples*. In supervised learning, each example is a *pair* consisting of an input object (typically a vector) and a desired output value (also called the *supervisory signal*). A supervised learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples. An optimal scenario will allow for the algorithm to correctly determine the class labels for unseen instances. This requires the learning algorithm to generalize from the training data to unseen situations in a "reasonable" way.

The parallel task in human and animal psychology is often referred to as [concept learning](https://en.wikipedia.org/wiki/Concept_learning).

# Unsupervised learning

**Unsupervised learning** is a type of self-organized [Hebbian learning](https://en.wikipedia.org/wiki/Hebbian_learning) that helps find previously unknown patterns in data set without pre-existing labels. It is also known as [self-organization](https://en.wikipedia.org/wiki/Self-organization) and allows modeling [probability densities](https://en.wikipedia.org/wiki/Probability_density_function) of given inputs. It is one of the main three categories of machine learning, along with [supervised](https://en.wikipedia.org/wiki/Supervised_learning) and [reinforcement learning](https://en.wikipedia.org/wiki/Reinforcement_learning). [Semi-supervised learning](https://en.wikipedia.org/wiki/Semi-supervised_learning) has also been described, and is a hybridization of supervised and unsupervised techniques.

Two of the main methods used in unsupervised learning are [principal component](https://en.wikipedia.org/wiki/Principal_component_analysis) and [cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis). [Cluster analysis](https://en.wikipedia.org/wiki/Cluster_analysis) is used in unsupervised learning to group, or segment, datasets with shared attributes in order to extrapolate algorithmic relationships. Cluster analysis is a branch of [machine learning](https://en.wikipedia.org/wiki/Machine_learning) that groups the data that has not been [labelled](https://en.wikipedia.org/wiki/Labeled_data), classified or categorized. Instead of responding to feedback, cluster analysis identifies commonalities in the data and reacts based on the presence or absence of such commonalities in each new piece of data. This approach helps detect anomalous data points that do not fit into either group.

# Reinforcement learning

**Reinforcement learning** (**RL**) is an area of [machine learning](https://en.wikipedia.org/wiki/Machine_learning) concerned with how [software agents](https://en.wikipedia.org/wiki/Software_agent) ought to take [actions](https://en.wikipedia.org/wiki/Action_selection) in an environment in order to maximize some notion of cumulative reward. Reinforcement learning is one of three basic machine learning paradigms, alongside [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning) and [unsupervised learning](https://en.wikipedia.org/wiki/Unsupervised_learning).

It differs from supervised learning in not needing labelled input/output pairs be presented, and in not needing sub-optimal actions to be explicitly corrected. Instead the focus is on finding a balance between exploration (of uncharted territory) and exploitation (of current knowledge).

# Regression

### 

### Regression is a statistical measurement used in finance, investing, and other disciplines that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).

Regression helps investment and financial managers to value assets and understand the relationships between variables, such as [commodity prices](https://www.investopedia.com/terms/c/commodity.asp) and the stocks of businesses dealing in those commodities.

# Linear Regression

# 

# **Objective of supervised Learning**

- we want to generate a function or hypothesis from training data using some algorithm and then from the function we can predict the output for our test data.

-

### **Objective of linear Regression**

- let the function be a type of linear function i.e for vector x(x1,x2,x3,.....,xn) we have function as

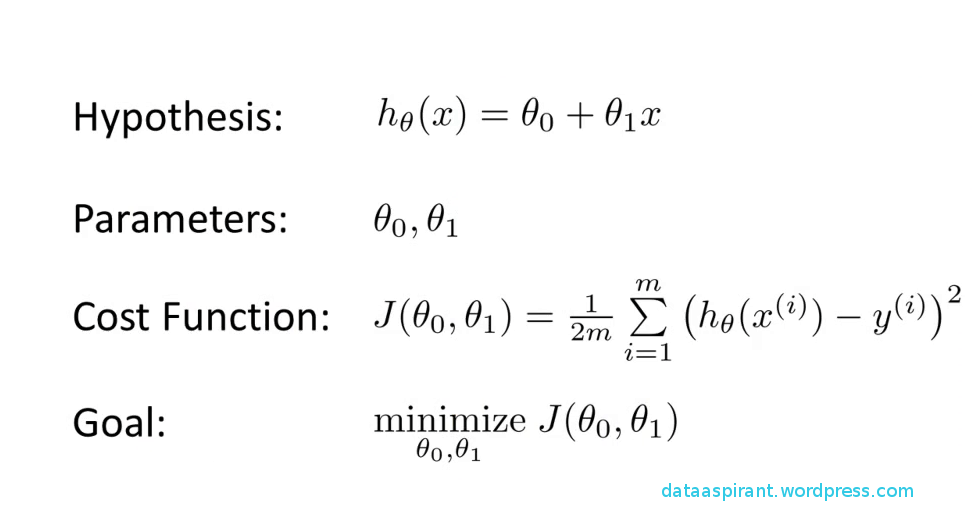
- f(x) = a(x1) + b(x2)+ c(x3)...............+ alpha(xn) + constan

- now we have to chose (a,b,c,d,.......,alpha) in such a way we predict as closely as possible to real outcome

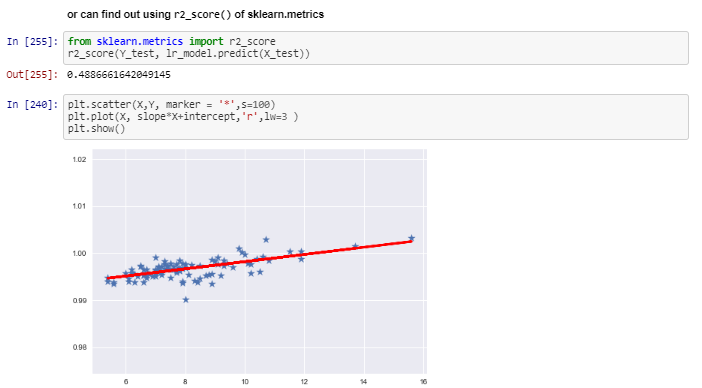
- for that we defined our error/cost function as in figure and our goal is to minimize it : -

- here let data be univariate

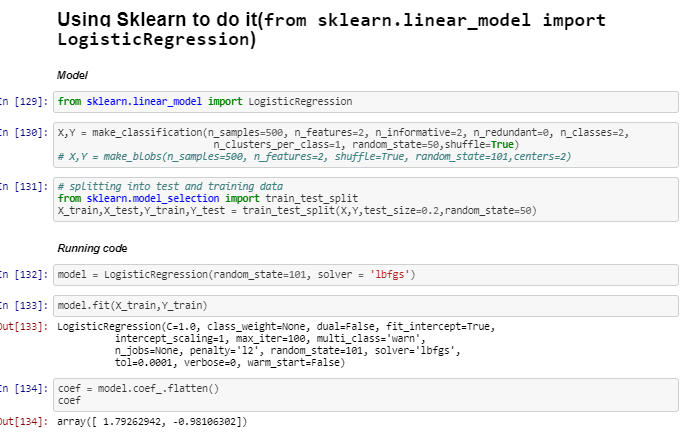
- cost function is convex function so local minima is global minima.

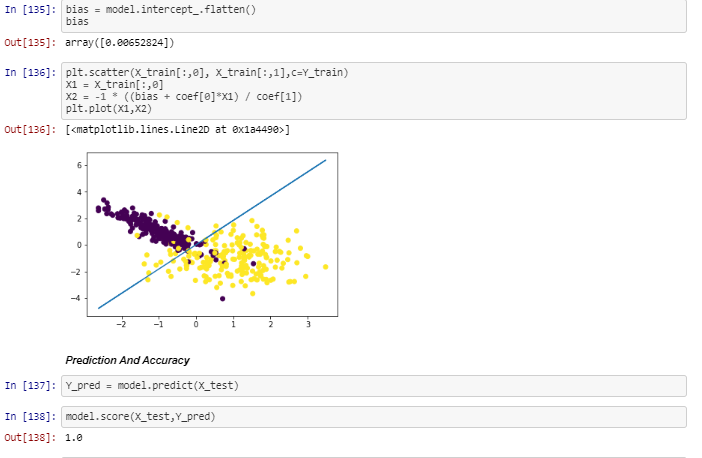






# Logistic Regression

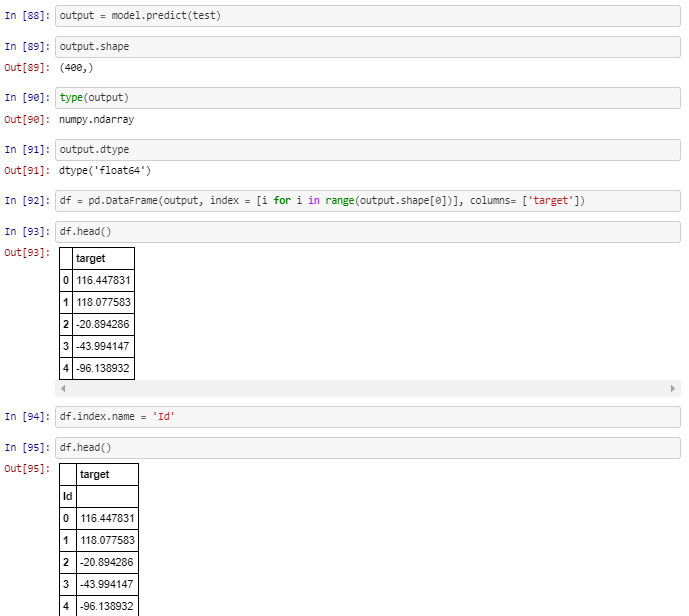
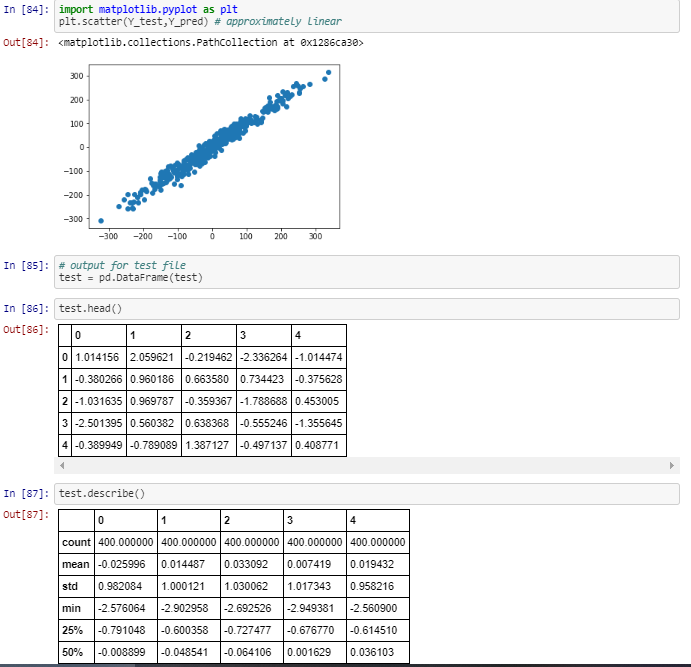
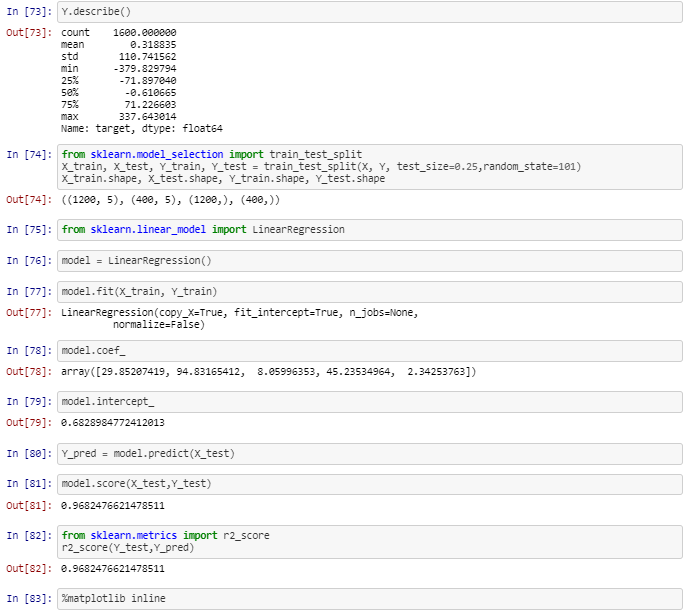
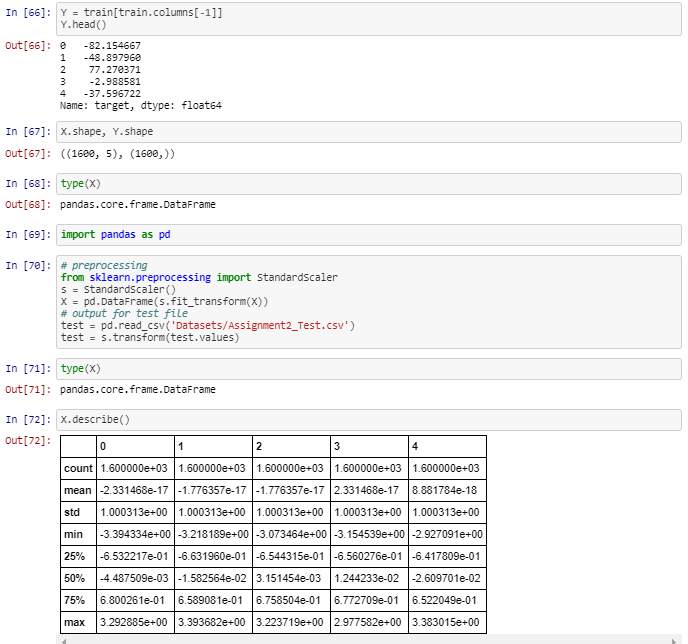
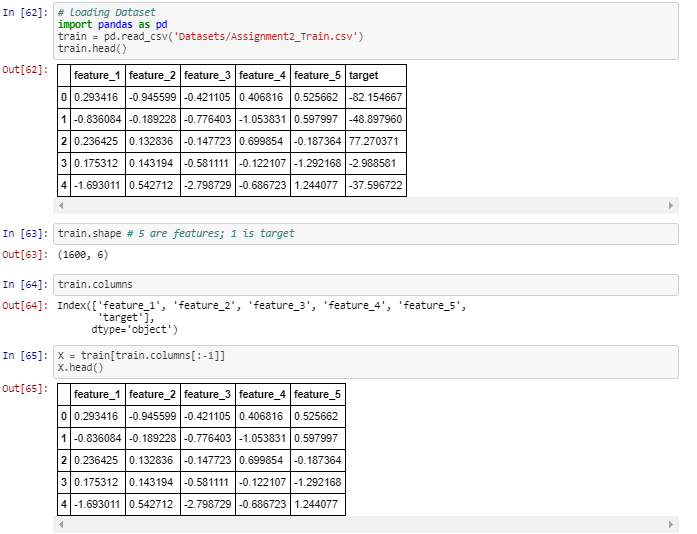




# MultiVariate Regression

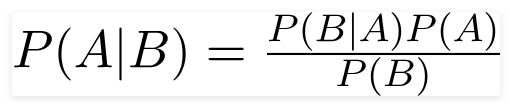
## **Challenge - Air Pollution**

It is winter time in Delhi so Cody decided to go for a walk to the news stand and on reaching was surprised to see the air quality index of Delhi in the newspaper . So he decided to collect the air samples of different locations and then took these samples to his lab where he extracted five features of the air he collected which can be used to predict the air quality index and combined it with the air quality index given in the newspapers. You are provided with the data collected by Cody and your job is to design a machine learning model , which is given by the features extracted by Cody to predict air quality.



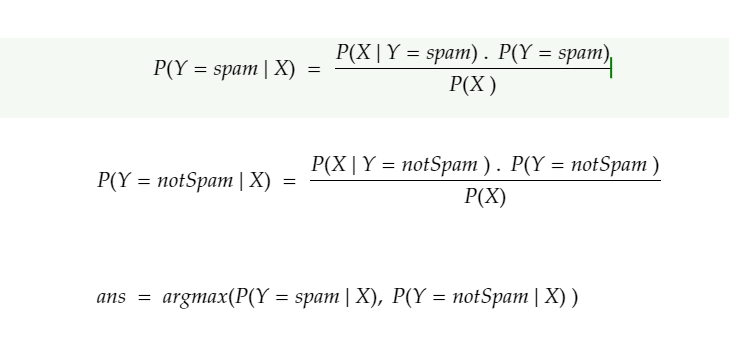
# Naive Bayes Classifier

# **Bayes Theorem**

****

# **Naive Bayes Classifier**

* use to detect whether review is +ve , -ve or neutral(sentiments Analysus)
* use ot classify the category of emails -> promotional,spam,social,inbox
* use to detect whether email is spam or not. using

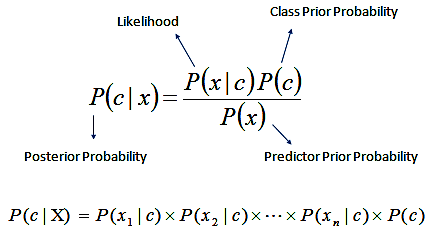


What **is** a classifier?

A classifier **is** a machine learning model that **is** used to discriminate different objects based on certain features.

Principle of Naive Bayes Classifier:

A Naive Bayes classifier **is** a probabilistic machine learning model that’s used **for** classification task.

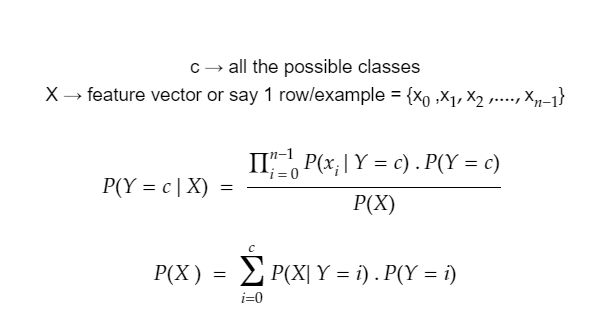


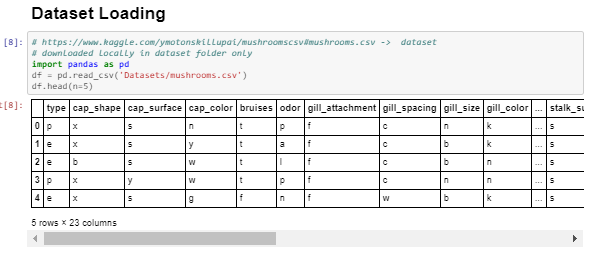
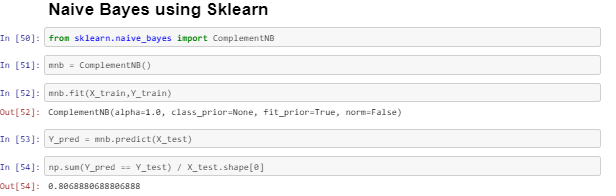
### **Assumption in Naive bayes classifier**

- all the features are independent of each other

- **if** x = <x1,x2,x3,.....,xn>

- then P(X|Y) = P(x1,x2,x3,..,xn|Y) = P(x1|Y)P(x2|Y)P(x3|Y).....P(xn|Y)





# KNN(K-nearest Neighbour Algorithm)

In [ ]:

1) Simplest Machine learning Algorithm you will find **in** Machine Learning

2) Brute Force Approach **is** used **in** it. (Hence Act **as** a ^Baseline^ , **if** u use any other algo then that algo must have accuracy better than KNN)

3) ^ Complexity O( M\*N + (M + klogM) + k ) per Query ^ => **if** dataset **is** Large , this **is** gonna take much time to execute

where M **is** number of samples

**and** N **is** number of features **in** each sample.

The k-nearest neighbors algorithm (k-NN) is a non-parametric method or lazy learners used for both classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small).

If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

# **ALGO**

1) Find distance of query point **from** **all** **other** **point** **in** **dataset** **and** **store** **in** **as** **a** **list** **of** **tuple** (distance,valueORlabel)

2) Sort the list based on distance

3) Take first k-smallest **or** nearest elements **from** **the** **sorted** **list**

4 a) **if** problem **is** classification

- take the majority Vote Label **and** assigned to our Query Point

4 b) **if** problem **is** Regression

- take the mean of values **and** assigned to our Query Point



# KMeans Clustering

# **Algorithm**

The Κ-means clustering algorithm uses iterative refinement to produce a final result. The algorithm inputs are the number of clusters Κ and the data set. The data set is a collection of features for each data point. The algorithms starts with initial estimates for the Κ centroids, which can either be randomly generated or randomly selected from the data set.

The algorithm then iterates between two steps:

1. **Data assigment step:**

Each centroid defines one of the clusters. In this step, each data point is assigned to its nearest centroid, based on the squared Euclidean distance. More formally, if ci is the collection of centroids in set C, then each data point x is assigned to a cluster.

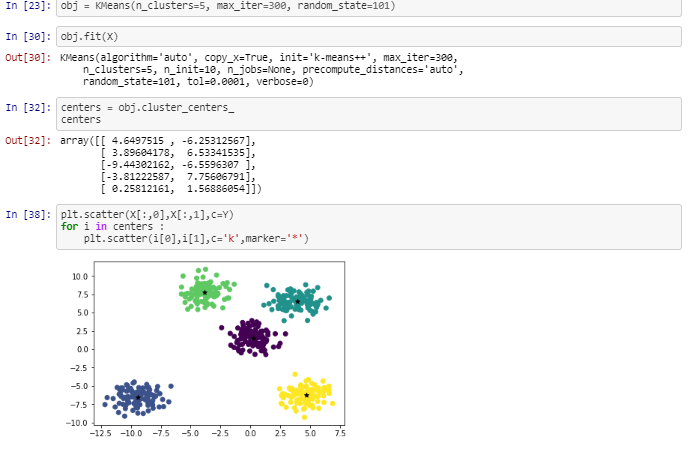
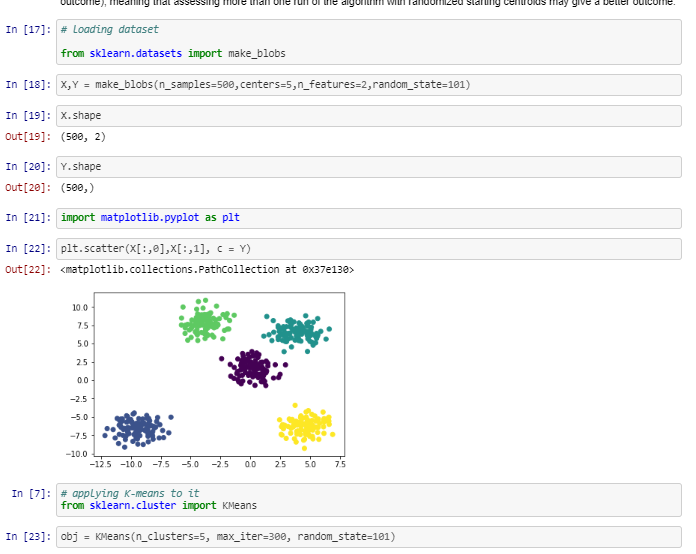
where dist( · ) is the standard (L2) Euclidean distance. Let the set of data point assignments for each ith cluster centroid be Si.

**Centroid update step:**

In this step, the centroids are recomputed. This is done by taking the mean of all data points assigned to that centroid's cluster.

The algorithm iterates between steps one and two until a stopping criteria is met (i.e., no data points change clusters, the sum of the distances is minimized, or some maximum number of iterations is reached).

This algorithm is guaranteed to converge to a result. The result may be a local optimum (i.e. not necessarily the best possible outcome), meaning that assessing more than one run of the algorithm with randomized starting centroids may give a better outcome.



**Conclusion and Further Enhancement**

* The analysis of review using Neural Network on IMDB dataset is done succesfully with reviews classified as “Positive” or “Negative”.
* The further enhancement includes avoiding even slightest of overfitting and errors.
* The approach of this project is to extend and scrap data from any review providing website and perform sentimental analysis on that data.