WOC PROJECT REPORT

-BY HARSHIT JAIN

(23JE0394)

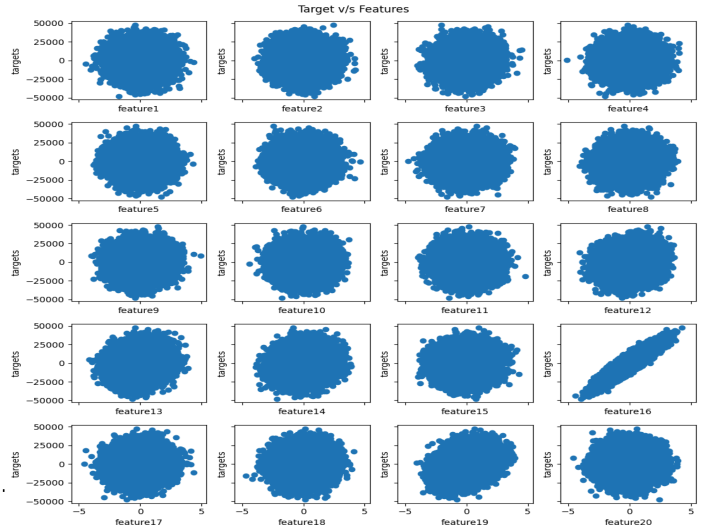
Linear Regression :

Linear regression is a supervised learning algorithm used to model a linear relation between a dependent variable and one or more independent variables.

Hypothesis Function : $h(x) = a_0 + a_1.x_1 + a_2.x_2 \ldots\ldots + a_n.x_n$

Where a_i are learned parameters , and n is the number of independent variables.

First the dataset was shuffled and then split into training(80%), cross-validation(10%) and test set(10%) and then features were Z-score normalized.



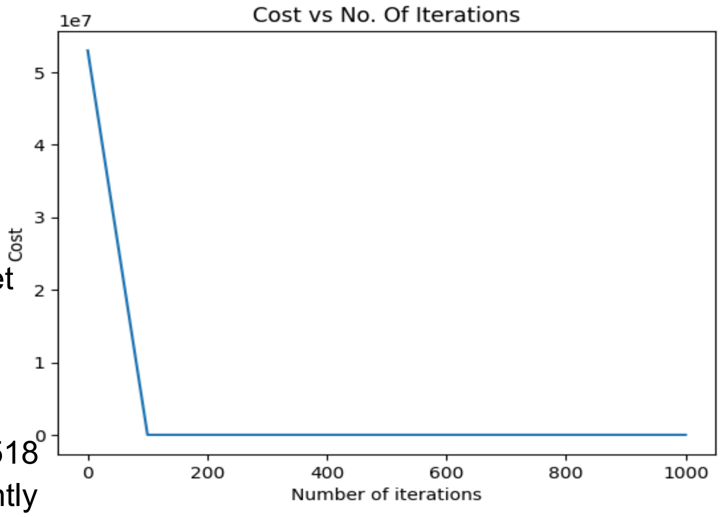
Hyperparameters :

- Learning Rate : 0.1
- Number of iterations : 1000

Results :

- Cost on training set = 0.005074040043648406
- R2-Score on training set = 0.9999999999224729
- Cost on Cross-validation set = 0.0050032835166230
- R2-Score on Cross-validation set = 0.9999999999235262
- Cost on test set = 0.004919975960777203
- R2-Score on test set = 0.9999999999256693

- Convergence was achieved within 400 iterations



Issues faced :

Initially there was inconsistent normalization across set leading to higher cost on cross-validation and test set.

Cost due to inconsistent normalization :

- Cost on test set = 25088.729358504872.
- Cost on cross_validation set = 18050.194176418518

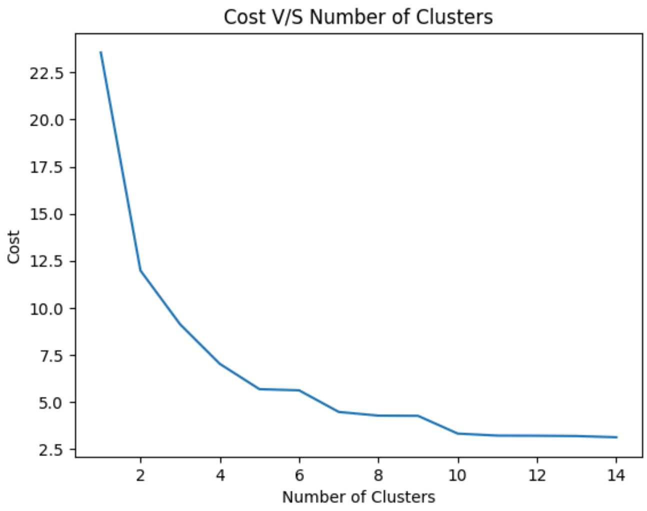
Modifications : All datasets were normalized consistently

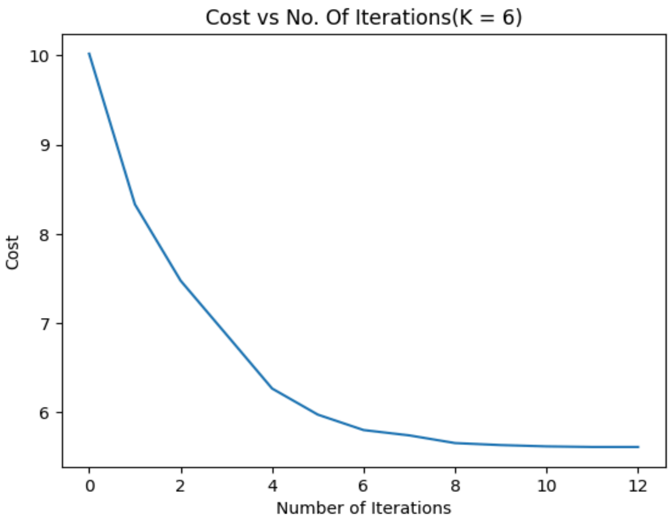
K-Means Clustering :

K-means clustering is a type of unsupervised machine learning algorithm used to group data points into clusters based on their similarities.

Hyperparameters :

- Number of Clusters :
  Initial Value = 7
  A graph between cost versus number of clusters(K) was plotted and “K” was chosen using the elbow method.
  Optimal value of Number of clusters = 6





- Number of iterations :
  Training of model stops when there is no change in position of centroids , hence there is no fixed number of iteration but maximum number of iterations is limited t0 150.In case of six clusters , total iterations were 13.

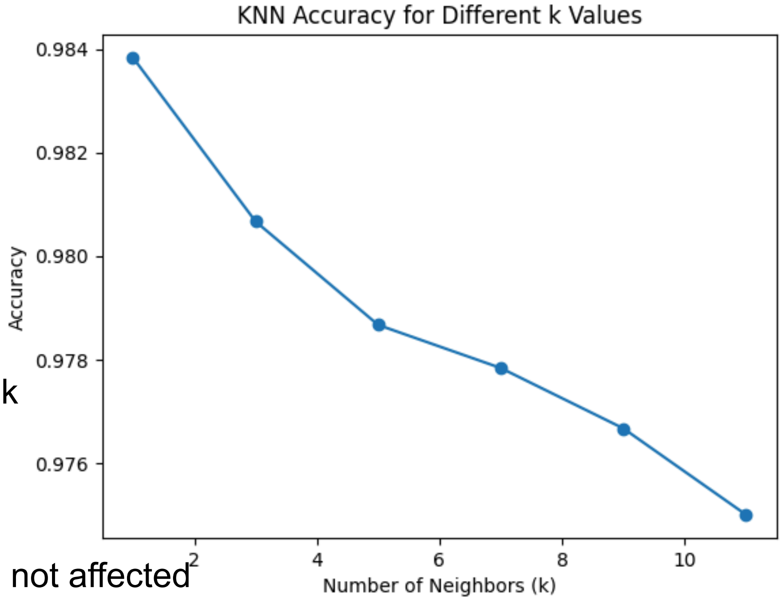
K-Nearest Neighbours :

It is the simplest supervised machine learning algorithm for classification.It classifies a data point based on how its neighbors are classified.Nearest neighbors are decided on the basis of distance metric.

First the dataset was shuffled and then split into training(80%) and test set(20%) , then features were scaled by dividing by the maximum value of them.
Z- score score is not preferred as it will lead to negative data for some pixels which does not have any interpretation in image data.

Hyperparameter :

- Number of nearest neighbors (K) :
  Initial Value = 3
  Tested with different values of 'K' (1 ,3 ,5, 7, 9 ,11) and chose the 'K' resulting in maximum accuracy.
  Optimal Value of 'K' = 1



Results :

- F1-score : 0.985123966941487
- Accuracy : 0.9838333333333333

Issues faced :

Major issue that i faced was the code took lot of time to run (around 11 minutes)

Modifications :
For i calculated squared distances , it did not affected Much but runtime was reduced to approx.9 minutes)

Polynomial Regression :

In Polynomial regression, the original features are converted into Polynomial features of required degree (2,3,..,n) and then modeled just like linear regression model

First the dataset was shuffled and then split into training(80%), cross-validation(10%) and test set(10%) and then features were Z-score normalized.

Hyperparameters :

- Degree of polynomial : 6

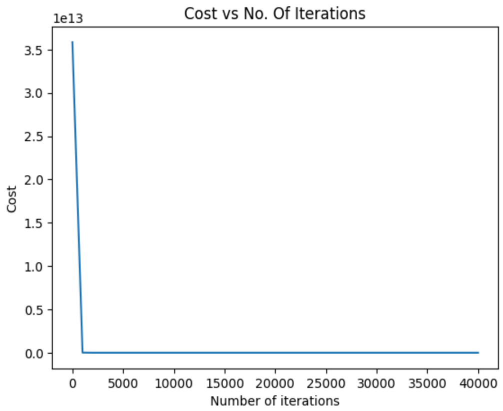
Tried experimenting with different degrees and best results were for six.

Degree of poly.	Cost on training set	R2-score on training set
2	42857832257278.62	0.39730851804146117
3	41697316592292.35	0.41362835665971454
4	4330210785160.58	0.9418800563254649
5	2638000737162.2534	0.9529118339879462
6	1.17420600708820613 x 10^{-15}	1.0

- Learning Rate : 0.1
  Started with an initial value of 0.01 cost was decreasing continuously then increased it to 0.1 then further increase in learning rate resulted in divergence.
- Number of Iterations : 40000
  Initial value : 10000
  Then i kept on increasing till cost was negligible (as runtime was short , there were no issue in increasing to larger number of iterations)

Results :

- Cost on training set = 1.17420600708820613 x 10^{-15}
- R2-Score on training set = 1.0
- Cost on Cross-validation set =7.420624551148468 x 10^{-16}
- R2-Score on Cross-validation set = 1.0
- Cost on test set = 1.3043303141224729 x 10^{-15}
- R2-Score on test set = 1.0



Logistic Regression :

It is a supervised machine learning algorithm used for binary classification.It uses a logistic function, also known as a sigmoid function that takes input as independent variables and produces a probability value between 0 and 1.

First the dataset was shuffled and then split into training(80%) , cross-validation(20%), test set(20%) , then features were scaled by dividing by the maximum value of them.

Z- score score is not preferred as it will lead to negative data for some pixels which does not have any interpretation in image data.

Then given targets were one hot encoded.

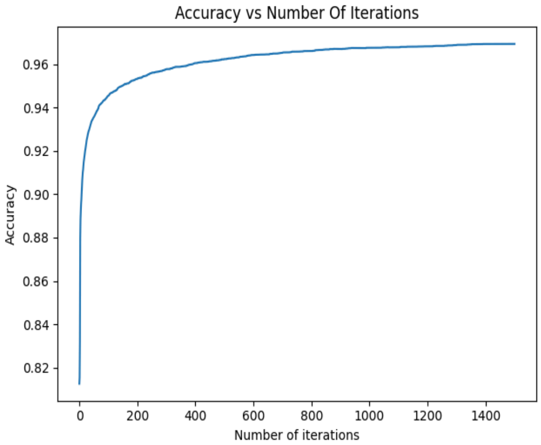
Model Type : one vs all

Hyperparameters :

- Learning Rate : 1

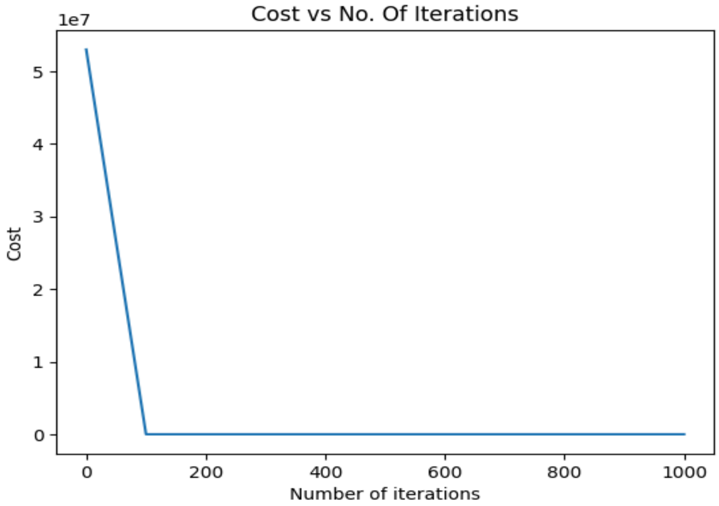
- Number of iterations : 1500

(As runtime was high , the number of iterations were limited.)



Results :

- Cost on training set = 0.04335458249246991
- F1-Score on training set = 0.9875052061640982
- Accuracy on training set = 0.9692916666666667
- Cost on Cross-validation set = 0.013911263085621604
- F1-Score on Cross-validation set = 0.9851239669421487
- Accuracy on Cross-validation set = 0.9593333333333334
- Cost on test set = 0.026011124894460263
- F1-Score on test set = 0.984873949579832
- Accuracy on test set = 0.9636666666666667



Neural Network :

A neural network is a supervised machine learning model designed for various tasks, including binary or multiclass classification. It Inspired by the human brain, it comprises interconnected neurons arranged in layers.

First the dataset was shuffled and then split into training(80%) , cross-validation(20%), test set(20%) , then features were scaled by dividing by the maximum value of them.
Z- score score is not preferred as it will lead to negative data for some pixels which does not have any interpretation in image data.
Then given targets were one hot encoded

Model :

Architecture :

- Input layer
- Hidden layers : 256(reLu) , 128(reLu) , 64(reLu)
- Output layer(softmax)

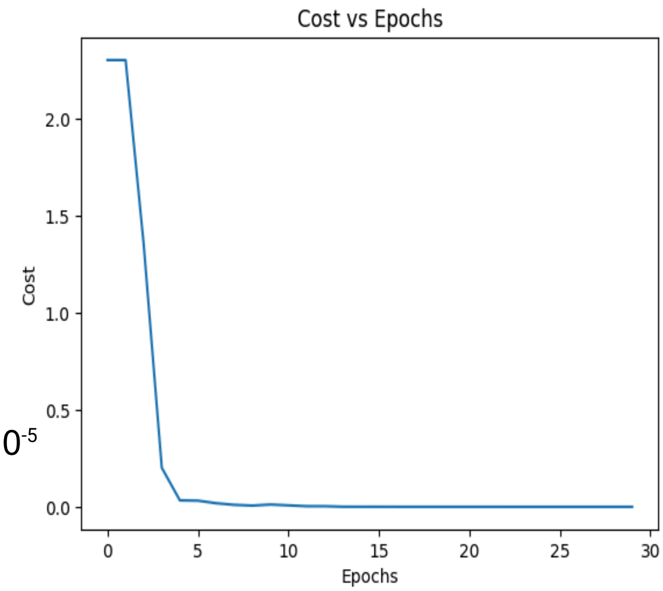
Epochs : 30

Mini-batch size : 32

Learning Rate : 0.1

Results :

- Accuracy on training set : 1.0
- Accuracy on test set : 0.9826666666667
- Accuracy on Cross-validation set : 0.986
- Cost on training test : 3.5149029668815986 x 10^{-5}



Issues faced :

I used reshaping the feature array instead of transposing which resulted in low accuray (around 0.09) and cost was not decreasing at all **,** after correction above results were achieved.