**Data preprocessing:** A very important part of data analysis, this involves

1. importing useful libraries such as NumPy, pandas matplotlib, seaborn, and sklearn,

2. importing the dataset under consideration.

3. Data Cleaning which involves removing of duplicates, replacement of missing values or removal of affected rows if any.

4. Data conversion in case of categorical values

5. Scaling in case where values are either too large or far apart. StandardScaler in scikit library will be where necessary to readjust the distribution to have a mean value 0 and standard deviation of 1.

6. Splitting dataset into train and test sets. Selecting at random, the dataset can be divided in ratio usually with test set taking 0.2-0.33 of the datasets while the rest, train set is used in training the model.

the training data is usually much as more data tends to improve modelling.

TASK 1:

Multiple Linear regression (MLR):

a statistical modelling approach that could be used type for predictive analysis, like simple linear regression (SLR) is used to show relationship

between values of a dataset. Unlike SLR, which shows between two variables, MLR tends to use more than

one independent variable hence showing relationship between multiple independent variables against a targeted point

or column of interest.

Fitting MLR to the Training set and Predicting the Test set results:

1. The dataset is divided between test and train in the ratio 1 to 5 to afford more data point for train consideration the data size.

2. varying the random state has little effect on the r2-score, best at around 10, hence it can be said that they are closely related

3. Using scikit package, linear least-squares regression line is being fitted to the training data then used to predict target values for the test set.

Data Analysis:

1. the coefficient shows the relationship of individual independent variable to the target. Only 7 of the independent variables are negatively related to the target i.e., price.

2. An accuracy of 71% was observed

repeating d same analyses without least gives better Result

Result Interpretation:

Suggestion/Conclusion:

prediction

Task2:

Clustering:

Cluster analysis, or clustering, is an unsupervised machine learning approach that tends to group input data based on observed properties or features into natural group.

Clustering has a large no. of applications spread across various domains. Some of the most popular applications of clustering are: Recommendation engines, Market segmentation, social network analysis, Search result grouping, medical imaging, Image segmentation, Anomaly detection etc.

Cluster:

A cluster may then be referred to an area with density of a particular feature or group of properties with data points closer to a central point than those in another cluster

Clustering Algorithms:

clustering algorithms use similarity or distance measures between data points in the feature space

in an attempt to show regions with dense observations; data scaling is often used to ensure data points are spread around a central point with mean of 0 and standard deviation of 1.

For the purpose of this analysis K-means, a popular centroid-based algorithm will be used. it is best used on smaller data sets because it iterates over all of the data points to classify them.

works in these 5 steps:

1. Specify the desired number of clusters K: Let us choose k=2 for these 5 data points in 2-D space.

2. Randomly assign each data point to a cluster: Lets assign three points in cluster 1 shown using red color and two points in cluster 2 shown using grey color.

3. Compute cluster centroids: The centroid of data points in the red cluster is shown using red cross and those in grey cluster using grey cross.

4. Re-assign each point to the closest cluster centroid: Note that only the data point at the bottom is assigned to the red cluster even though its closer to the centroid of grey cluster. Thus, we assign that data point into grey cluster

5. Re-compute cluster centroids: Now, re-computing the centroids for both the clusters.

6. Repeat steps 4 and 5 until no improvements are possible: Similarly, we’ll repeat the 4th and 5th steps until well reach global optima. When there will be no further switching of data points between two clusters for two successive repeats. It will mark the termination of the algorithm if not explicitly mentioned.

Analysis and conclusions:

Ideas to consider when completing this task:

• Can you make any conclusions about the clustering?

• Include as many features as you can. Does the clustering change?

• What advice would you give, in the context of the data, based on the clustering?

Task3:

Clustering:

Cluster analysis, or clustering, is an unsupervised machine learning approach, involving automatically discovering natural grouping in data. Unlike supervised learning, clustering algorithms (like predictive modelling) only interpret the input data and find natural groups or clusters in feature space.

Clustering techniques is used when there is no class to be predicted but rather the given dataset is to be divided into groups based on observed features.

Neural networks:

a deep learning approach which tends to simulating the way human brain works. it can be used for classification or regression.

A typical Neural network comprises of:

input layer: where data is being feed into the model

hidden layer(s): Any layer(s) which can be found between the input and output layers.

output layer: which is the result of our modelling or algorithm computations.

Deep learning: is a technique in which the neural network is left to figure out important features instead of applying feature engineering techniques.

Synapses: take a value from their input, multiply it by a specific weight, then output the value.

Neurons: being more complicated, they add together the outputs of all their synapses, and apply an activation function.

activation functions: they allow neural nets to model complex non-linear patterns, that simpler models may miss.

Using sigmoid activation functions to build in neural net. the sigmoid function limits the output to a range between 0 and 1.

In neural net visuals, circles connote neurons while lines synapses. the process is very similar: initializing with random weights and bias vectors, the model makes a prediction, compare it to the desired output, and adjust the vectors, i.e., weights and bias to predict more accurately the next time.

The process continues until the difference between the prediction and the correct targets is minimal.

Probability functions give you the probability of occurrence for possible outcomes of an event. The only two possible outputs of the dataset are 0 and 1, and the Bernoulli distribution is a distribution that has two possible outcomes as well. The sigmoid function is a good choice if your problem follows the Bernoulli distribution, so that’s why you’re using it in the last layer of your neural network.

Since the function limits the output to a range of 0 to 1, you’ll use it to predict probabilities. If the output is greater than 0.5, then you’ll say the prediction is 1. If it’s below 0.5, then you’ll say the prediction is 0. This is the flow of the computations inside the network you’re building

In the process of training the neural network, you first assess the error and then adjust the weights accordingly. To adjust the weights, gradient descent and backpropagation algorithms could be put to use.

Gradient descent: It involves the use of the derivative or slope to find the direction and the rate to update the parameters.

Data analysis and conclusions:

Confusion matrix: is a matrix (table), usually 2x2 that can be used to measure the performance of a machine learning algorithm.

it shows false positive, true positive, true negative and false negative. It helps to see how well a model predicts the values of the test dataset

Comparison:

Neural network gives best result compare to the rest followed by Logistic Regression using its random state property gives a better result than Naive Bayes

It can be observed that the at some point increasing hidden layers in a neural network hardly improves the system but takes time to complete

The neural network is a really cool approach having to improve itself based on corrections while iterating make the system less rigid on decision.

Ideas to consider when completing this task:

than many, analyses in depth rather than being superficial and repetitive.

� Is there a way of visualizing the model(s)?

� How will you assess the effectiveness of the model(s)?

� Include as many features as you can. Does the model improve? it can be observed that increasing features increases the accuracy of the model till an upper limit is reached.

� How could you make further improvements?

� What can you conclude about your model?

� How strong is the relationship between the predictor and target variables?