**Article on Avocado Project**

**Problem Definition:-**

In the Avocado project the retail data has been taken from the retailers based on the actual retail sales of Hass Avocados. So in this project we have to predict two things “region” from where the observations were taken and the “Average Price” means the average price of the single Avocado. So we have to build two models for the above two target variables.

**Data Analysis:-**

First of all we have read the csv file using the read\_csv from pandas. After that we have checked the shape of the dataset means how many rows and columns present in the dataset. Then we describe the data using the describe() function. It will give lots of information about the dataset but it will show only those columns which are of int and float data types. It does not the show the columns which are of object data type until you mention it through the include parameter.While describing the dataset we have analyzed that all the values are present in the columns by seeing the count of every column, the mean will show the average value of each column, the std shows the standard deviation in every column, the min shows the minimum value of every column and through this we can analyze that which column has the 0 value since it shows the minimum value of every column and later we can deal the zero values in the dataset if necessary. Then 25%,50%,75% will show the data value upto the 25%,50%,75% of the dataset and in the last the max will show the maximum value of each column.

After describing the data we check for the null values whether they present or not in the columns by using the function .isna().sum(). The “.isna()” will check the for the null values and the “sum()” function will count how many null values present in the current column . We analyzed that no null values present in the data set. After that we check for the 0 values present in the data set and we analyze that many column are having the null values we can deal with them later. After that we check for the Data type of each column. We analyze that Out of 14 , 3 columns are of object data type, 2 are of integer and the rest columns are of float data type. After that we have seen the column named “Unnamed 0” which is of no use so we have dropped that column by using the drop function. After that we have dealt with the columns which have the 0 values by replacing them with tha mean of that column.

Then the Visualization of the dataset come into the picture and for that we have used the count plot for the categorical columns as it clearly shows the number of values lies in the particular category and also the graph is also so descriptive. For the numerical columns we have used the distribution plot as it shows the distribution of the data clearly and we can also observe whether the skewness present in the columns or not and we can see that the skewness present in many columns which we can deal later. Then the encoding of the dataset id done with the help of the Label encoder. After that we check the correlation with the target variable i.e. “region” and we observe that the columns['Date','AveragePrice','type','year'] have less relation with the target variable so we have dropped them from the dataset. After that we checked for the outliers using the box plot and we found that max of the data are the outliers so we didn’t remove the outliers as we lose lots of data from the dataset. After that we checked for the skewness and remove it with the help of the power transform . Then we checked for the multicollinearity using the Variance inflation factor and we found it so we have removed the columns which have correlation with each other ['Total Bags','Small Bags'].

**EDA Concluding Remarks:-**

After doing the EDA we have conclude that there are no null values present in the data set that we have checked with the isna() function which checks whether the null values are present in the dataset or not. After that we check for the 0 values present in the data set and we analyze that many column are having the null values and we replaced them with the mean of that column. After that we check for the Data type of each column. We analyze that Out of 14 , 3 columns are of object data type, 2 are of integer and the rest columns are of float data type. After that we have seen the column named “Unnamed 0” which is of no use so we have dropped that column by using the drop function. Then the encoding of the dataset id done with the help of the Label encoder. After that we check the correlation with the target variable i.e. “region” and we observe that the columns['Date','AveragePrice','type','year'] have less relation with the target variable so we have dropped them from the dataset. After that we checked for the outliers using the box plot and we found that max of the data are the outliers so we didn’t remove the outliers as we lose lots of data from the dataset. After that we checked for the skewness and remove it with the help of the power transform . Then we checked for the multicollinearity using the Variance inflation factor and we found it so we have removed the columns which have correlation with each other ['Total Bags','Small Bags'].

Then in the Visualization of the dataset we have used the count plot for the categorical columns as it clearly shows the number of values lies in the particular category and also the graph is also so descriptive. For the numerical columns we have used the distribution plot as it shows the distribution of the data clearly. By seeing the count plot of the categorical column we have seen that the type of the Avocados are almost equal means 9126 avocados are od conventional type and 9123 are of organic type. And for the region column that there are many regions from where the avocados were taken. And while visualizing the numerical data we have seen lots of skewness in the columns which we remove with the help of the power transform .

**Pre-processing Pipeline:-**

Pre-processing pipeline is a predominant step in machine learning to yield highly accurate and insightful results. Greater the quality of data, the greater is the reliability of the produced results. **Incomplete, noisy, and inconsistent data** are the inherent nature of real-world datasets. Data preprocessing helps in increasing the quality of data by filling in missing incomplete data, smoothing noise, and resolving inconsistencies.

* **Incomplete data** can occur due to many reasons. Appropriate data may not be persisted due to a misunderstanding, or because of instrument defects and misfunctions.
* **Noisy data** can occur for a number of reasons (having incorrect feature values). The instruments used for the data collection might be faulty. Data entry may contain human or instrument errors. Data transmission errors might occur as well.
* There are many stages involved in data preprocessing,

1. **Data Cleaning**
2. **Data Integration**
3. **Data Transformation**
4. **Data Reduction**

**Data cleaning** attempts to impute missing values, smooth out noise, resolve inconsistencies, removing outliers in the data.

**Data integration** integrates data from a multitude of sources into a single data warehouse.

**Data transformations**, such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements.

**Data reduction** can reduce the data size by dropping out redundant features. Feature selection and feature extraction techniques can be used.

For preprocessing of the data we have seen the column named “Unnamed 0” which is of no use so we have dropped that column by using the drop function.After that we check for the null values and we observe that no null values are present in the data set and after that we check for the 0 values in the columns and we dealt them by replacing them with the mean of that column. Then the encoding of the dataset id done with the help of the Label encoder. After that we check the correlation with the target variable i.e. “region” and we observe that the columns['Date','AveragePrice','type','year'] have less relation with the target variable so we have dropped them from the dataset. After that we checked for the outliers using the box plot and we found that max of the data are the outliers so we didn’t remove the outliers as we lose lots of data from the dataset. After that we checked for the skewness and remove it with the help of the power transform . Then we checked for the multicollinearity using the Variance inflation factor and we found it so we have removed the columns which have correlation with each other ['Total Bags','Small Bags'].

**Building Machine Learning models:-**

The machine learning models which we have used for the target variable(“Average Price ”) are Linear Regression, Lasso Regression, Ridge Regression, Random Forest Regressor(Ensemble Technique).

* + **Linear Regression:-** Linear regression is a **linear model**, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y).It is giving the r2 score of 98.18 % which is very good accuracy and the cross validation score for that is 94.25%.
  + **Lasso Regression:- Lasso regression** is a type of [**linear regression**](https://www.statisticshowto.com/probability-and-statistics/regression-analysis/find-a-linear-regression-equation/)that uses [shrinkage](https://www.statisticshowto.com/shrinkage-estimator/). Shrinkage is where data values are shrunk towards a central point, like the [mean](https://www.statisticshowto.com/mean/). The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). It is giving the r2 score of 98.18 % which is very good accuracy and the cross validation score for that is 94.31%.
  + **Ridge Regression:-** Ridge regression is **a model tuning method that is used to analyze any data that suffers from multicollinearity**. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values. It is giving the r2 score of 98.18 % which is very good accuracy and the cross validation score for that is 94.25%.
  + **Random Forest Regressor:**- A random forest regressor is used for averaging to improve the predictive accuracy and control over-fitting. It is giving the r2 score of 99.99 % which is very good accuracy and the cross validation score for that is 99.99%.

The machine learning models which we have used for the target variable(“Region”) are Logistic regression,Decision tree classifier,Random forest, SVC.

* + **Logistic Regression:-** Logistic regression is a useful analysis method for classification problems, where you are trying to determine if a new sample fits best into a category. It is giving theaccuracy of 58.09 % which is very good accuracy and the cross validation score for that is 43.93%.
  + **Decision Tree Classifier:-DecisionTreeClassifier** · The function to measure the quality of a split.The strategy used to choose the split at each node. It is giving theaccuracy of 100 % which shows that there is some problem in Decision tree and the cross validation score for that is 100%.
  + **Random Forest:-** A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. It is giving the accuracy of 99.47 % which extremely good accuracy and the cross validation score for that is 96.59%.
  + **SVC:-** . It is giving the accuracy of 92.85 % which extremely good accuracy and the cross validation score for that is 82.61%.

**Conclusion remarks:-**

We conclude that for the target variable “Region” Random Forest is Performing better among all so we will choose Random forest classifier as our final model. And for the target variable ”AveragePrice” the best accuracy is given by Random forest regressor so we choose it our final model.