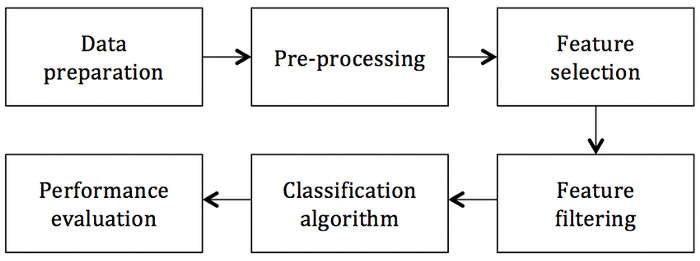
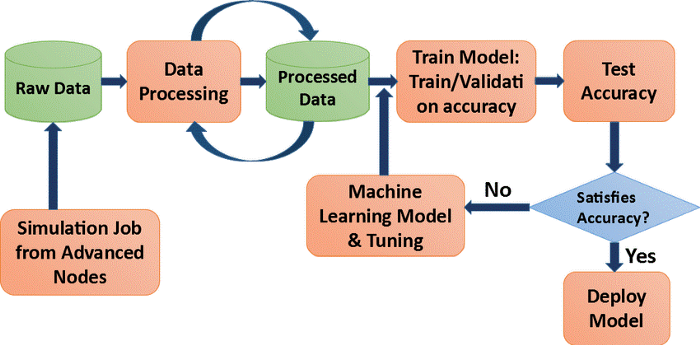


This data was downloaded from the Hass Avocado Board website in May of 2018 & compiled into a single CSV. Here’s how the Hass Avocado Board describes the data on their website:

The table below represents weekly 2018 retail scan data for National retail volume (units) and price. Retail scan data comes directly from retailers’ cash registers based on actual retail sales of Hass avocados. Starting in 2013, the table below reflects an expanded, multi-outlet retail data set. Multi-outlet reporting includes an aggregation of the following channels: grocery, mass, club, drug, dollar and military. The Average Price (of avocados) in the table reflects a per unit (per avocado) cost, even when multiple units (avocados) are sold in bags. The Product Lookup codes (PLU’s) in the table are only for Hass avocados. Other varieties of avocados (e.g. green skins) are not included in this table.





**Fields/Columns:**

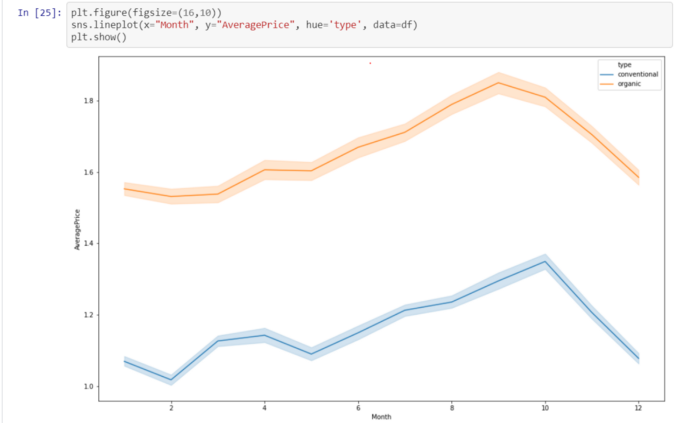
* Date — The date of the observation
* Average Price — the average price of a single avocado
* type — conventional or organic
* year — the year
* Region — the city or region of the observation
* Total Volume — Total number of avocados sold
* 4046 — Total number of avocados with PLU 4046 sold
* 4225 — Total number of avocados with PLU 4225 sold
* 4770 — Total number of avocados with PLU 4770 sold

I am importing the all library which I required for EDA, visualization, prediction and finding all matrices. The reason of doing this is that it become easier to use all the import statement at one go and we do not require to import the statement again at each point. We could find all the importing statement at one place without finding it on whole notebook and can update also.



From above we came to know that:

* Year 2017 is that year where the price is maximum as compared to other year, and there is less difference among rest of the year.
* September and October are the month where max no of average price is there, but the thing is almost for whole year the price is almost same for the avocado this prove that there is so much craze of avocado rather than India.



From above graph:

* There is hike between month 8–10 of both type of avocado both for conventional and organic type of avocado.
* Also, the conventional type of avocado is varying in term of price as seen in line plot because in starting the price is high but then it get decrease and so on.

**Plotting Histogram:**

* A **histogram** shows the frequency on the vertical axis and the horizontal axis is another dimension. Usually it has bins, where every bin has a minimum and maximum value. Each bin also has a frequency between x and infinite
* So, in this we can also check whether the graph is right skewed, left skew or the graph is normally distributed graph



From plotting this histogram, I used the bin size as 30, we can take any bin size (suited as per as data).

* Average price column is normally distributing over the histogram.
* Rest of the data are not much varying in term of numbers, so they are almost left skewed data
* To make the column as normal distributed we can use different methods, but I am using numPy log to make the skew values as normal distributed.

In prediction:

* I had done this prediction by taking Average price as an output variable which is continuity in nature so that why I’m using the regression technique
* While calculating the best random state the 80 is best state which providing the highest R2 score value for this model.
* After using the GridSeachCV, I can find the best param and then I used these param for that model.
* After using the best param I can get the best R2 score and the model is KNeighboursRegressor.
* There are following matrices which I find, and which are providing the best score.
* I also plot the scatter plot graph and we can see that the actual value and predicted values are very close to each other, so the line is best fit line.

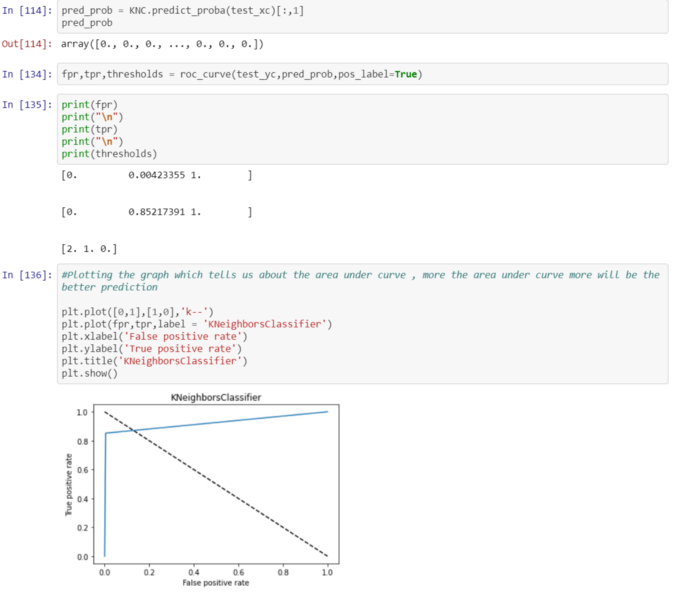
Now I am finding the score by taking region as an y value, I am using classification method because the region data is categorical in nature, so I am importing the classification model and their matrices.

**Roc Curve:**

A useful tool when predicting the probability of a binary outcome is the [Receiver Operating Characteristic curve](https://en.wikipedia.org/wiki/Receiver_operating_characteristic), or ROC curve.

It is a plot of the false positive rate (x-axis) versus the true positive rate (y-axis) for a number of different candidate threshold values between 0.0 and 1.0. Put another way, it plots the false alarm rate versus the hit rate.

The true positive rate is calculated as the number of true positives divided by the sum of the number of true positives and the number of false negatives. It describes how good the model is at predicting the positive class when the actual outcome is positive.

**Observation:**

* Taking price as y variable is predicting well for this model as compare to region
* Also, I used the Label Encoder to make the categorical data into numeric data i.e. Region and Sex
* Also, R2 score value is also greater then accuracy score
* Average price, total bags and total volume is well normally distributed data among all other column
* There are no outliers in the data set after replacing it through mean value
* As year is most negative co related column among all columns
* In between August to October the price of avocado is much higher as compared to other months
* Date 28,29 and 30 the price of avocado is high
* Hartford Springfield, San Francisco and New York are having more average price as compared to another region
* There is hike between month 8-10 of both type of avocado
* As organic type of avocado is having the more price per unit then conventional
* I had done prediction using region and price but using price the prediction score is high as compared to region
* So in this data set I am using both regression and classification technique for making this model