**Fresco Problem Analysis - Part-B**

In this project, we have used data analysis and EDA to understand various features.

- Fresco market being one of the biggest business markets in the UK wishes to identify trends and patterns in a sample of weekly data collected for a number of their loyalty cardholders during a 26-week period. The data includes information on the customers’ gender, age, shopping frequency per week and shopping basket price. Fresco operates two different types of stores (convenience stores and superstores) but they also sell to customers via an online shopping platform. The collected data are from all three different types of stores. Finally, the data provide information on the consistency of the customer’s shopping basket regarding the type of products purchased. These can vary from value products to brands as well as the supermarket’s own high-quality product series Fresco Top.

- Now The Fresco marketing management team is only interested in identifying whether the spending of the potential customer will fall into one of three possible groups including:

- Low spender (shopping basket value of £25 or less)

- Medium spender (shopping basket value between £25.01 and £70) and

- High spenders (shopping basket greater than £70)

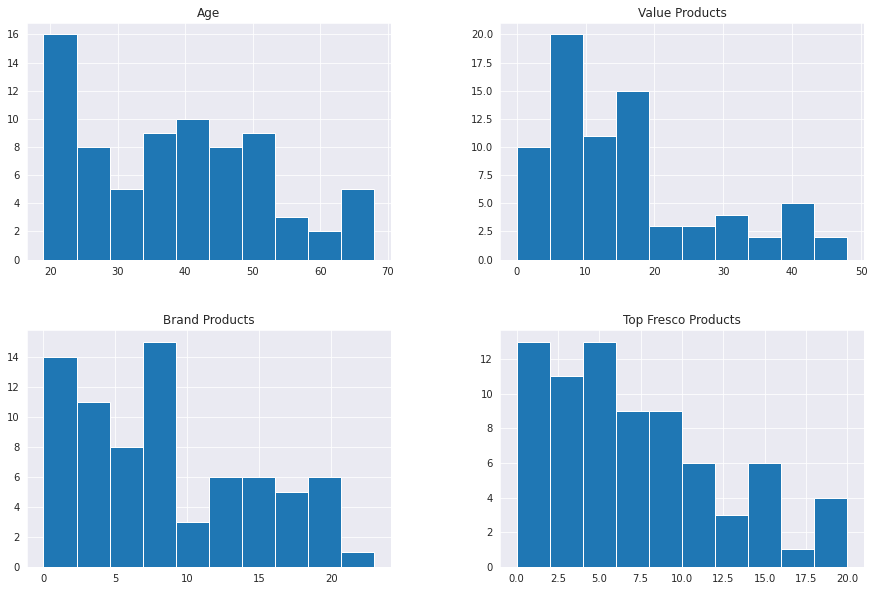
**Data Analysis and EDA**

- We observed that the data has about 75 rows and 8 columns

- Our target column should be a categorical problem according to the problem statement but a continuous value is found in the shopping basket column so binning is performed using cut() function to convert the continuous values into 3 categories

- columns with completely unique values are removed from data as they would lead to overfitting of the model

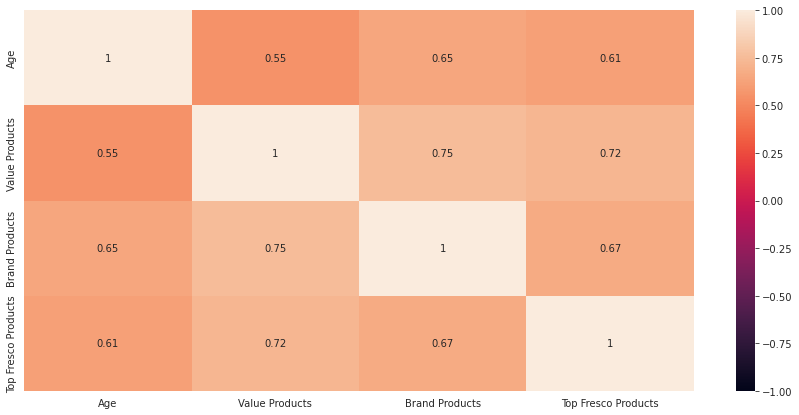
- We can observe from the histogram of ages that most of the data is normally distributed and the number of young people are more in the dataset



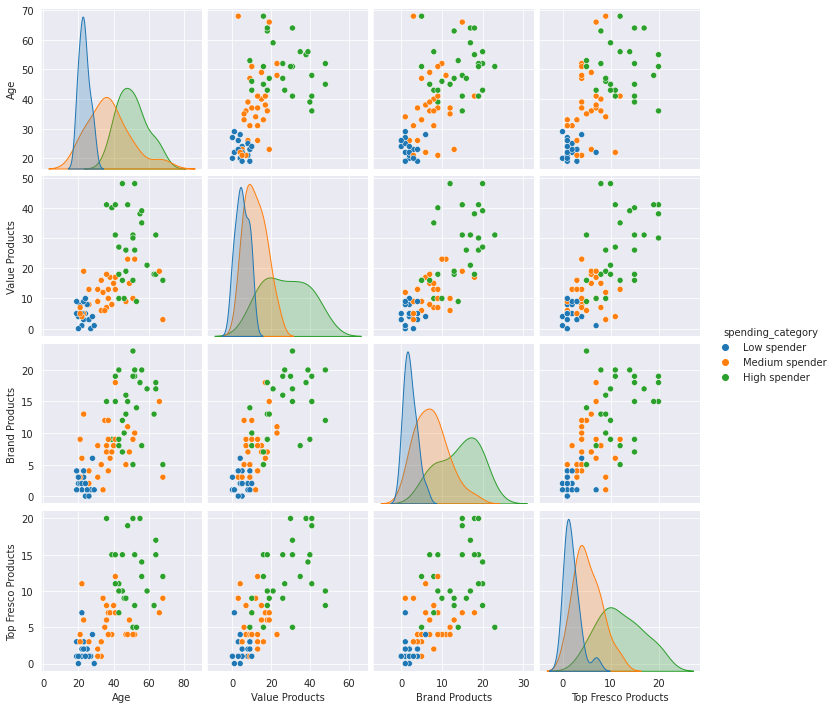
- We can also observe from the countplot that most of the purchases are done using superstore rather than convenient stores or online.

- We can also observe from scatter plots that most brand value products are purchased by high age people than younger ones

- We can also observe high correlation between features from the heatmap so we cannot simply use logistic regression, we would also require some tree based model to validate the performances.



- We can also observe that brand products and value products have a skewed distribution as observed from pairplot



- We have simply encoded the target features using a dictionary and map() function as it is the best method to encode it and provide a rank high to high spend people and low to low spend people

- We have also used dummy encoding to gender and store type features.

**Model Building and evaluation**

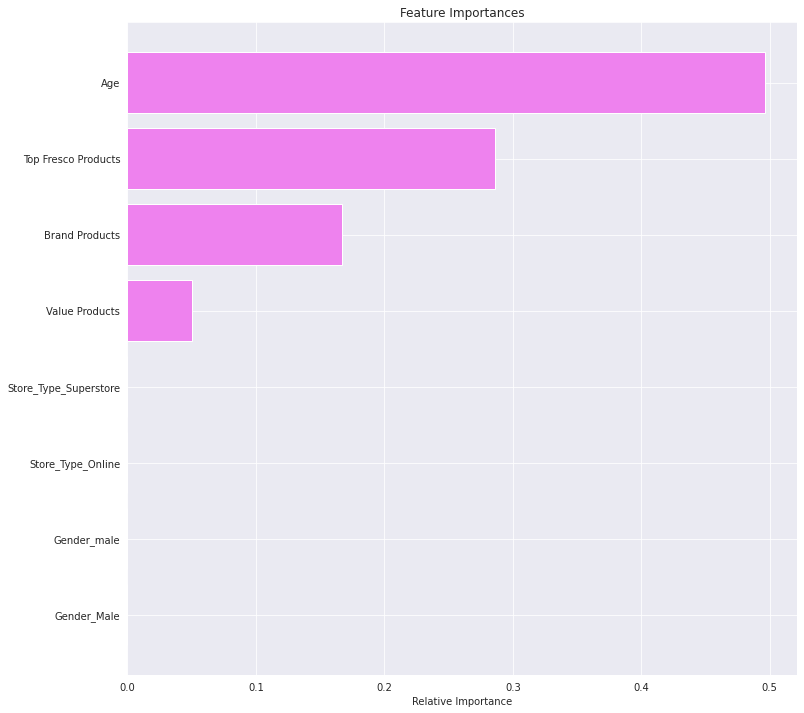
- We built two machine learning models - the first one being logistic regression

- Logistic regression has given about 69% accuracy on validation data but the variance inflation factor was high as there was a high multicollinearity among different features in the data.

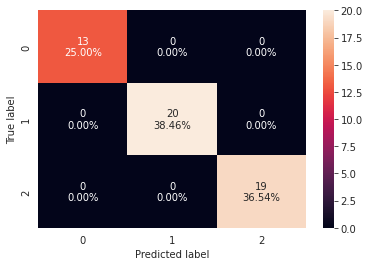
- So there was an idea to use tree based models and a decision tree model was used. decision tree model overfitted on the data so in order to tune the model, we have two approaches pre pruning and post-pruning

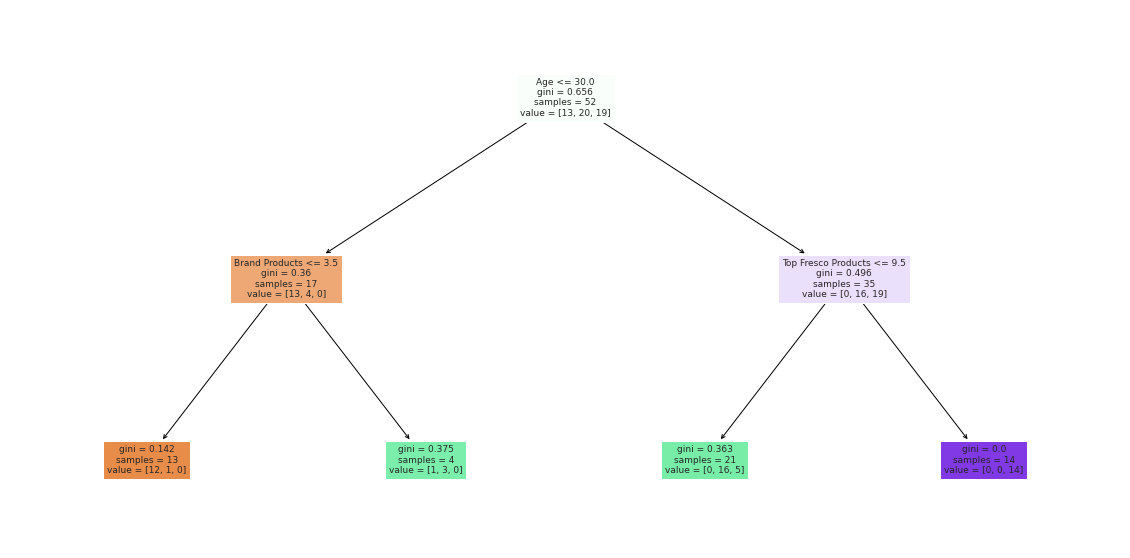
- Both the methods were used and it is observed that the post pruning method has given the best accuracy, precision and recall\_values both on the training and test data.

- Further we used gini impurity to check the impurity of the nodes and the feature importance is also obtained. We observed that top\_fresco\_products, Age and brand\_products have the highest feature importance.



- Even the confusion matrix and classification report proved the same.





**Conclusion**

- Finally we can observe that the model is about to classify the 3 classes correctly with an accuracy and precision above 75%. So the categories of high spend can be foreseen and more offers can be given for high spend people as they are likely to spend more