**Methodology**

* **The classifiers used**

Hotel\_loc, Hotel\_Name features values are considered as classifiers. Predict using logistic regression and decision tree.

* **Ensemble pipeline**

**Data Preprocessing** -> Dropped duplicates, Categorize the hotel\_address, Change the range of values between -1 and 1 for the columns have score values. Also used MinMaxScaler() for data normalization.

**Feature Selection** -> Used Pearson correlation and spearman correlation techniques to idenfy the highly correlated features and negatively correlated features. By using heatmap, understood which features are in top ranked.

**Feature Created** -> New features are created by calculating based on the given ratings and review scores of an hotel

**Data visualization** -> Plot the graph for No.of reviews and Scoring, Positive review word count and score, Negative review word count and score, top hotels and reviews, countries and No. of reviews.

**Model Selection** -> Tried Logistic regression to predict the result since the input features are numerical.

* **Other models considered**

Tried Decision tree model to predict the results and compare the same with Logistic regression model.

* **Hyper-parameter tuning**

Hyper parameter tuning is not used in this methodology.

**Overview**

* **Objective**

Build a hotel recommendation system to the user based on the customer reviews and ratings who belongs to the same category.

* **Methodology**

After data preprocessing and EDA, use logistic regression and decision tree models to predict the result. Conclude best among them based on the results.

**Dataset**

* **How many features** – 17 features in the given dataset
* **Size of the dataset** – 515738 rows

Created one more sub dataset from the given dataset contains 1493 rows and only has hotel names and corresponding addresses, avg\_score, total\_reviews and additional\_score.

* **Multiple files** – One file
* **What kind of data** – Both numerical and categorical
* **Balanced or imbalanced** – Balanced
* **Distribution of Training set, validation set, testing set**

70% Training set, 30% Validation set

* **Missing data and Preprocessing challenges**

Years and lng both are having missing data. We are ignoring the null values and not filtering the records since these features are not mandatory for our prediction.

**Results**

* Table for the evaluation metric for each ML technique used

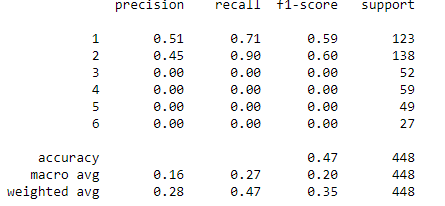
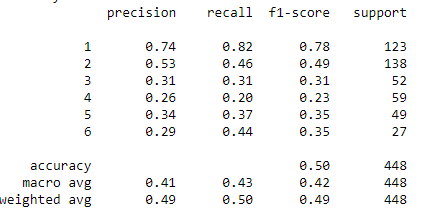
**Logistic Regression:**

Accuracy: 0.5736607142857143

**Decision Tree**:

Accuracy: 0.4955357142857143

Logistic Regression Decision Tree

* **Conclusion**

When comparing the Logistic regression and decision tree models, Logistic regression has good accuracy rate hence it is better than the Decision tree. Logistic regression is more suitable for numerical features with categorical class feature.

**Feature Engineering Techniques**

* **Features removed**

Review\_Date,days\_since\_review,lat,lng,Tags,Total\_Number\_of\_Reviews\_Reviewer\_Has\_Given,Positive\_Review,Negative\_Review,Reviewer\_Score

* **Feature creation**

Hotel\_Loc feature which has country code in the range of 1-6

* **Feature ranking**

Used Pearson correlation and spearman correlation techniques to identify the highly correlated features and negatively correlated features. By using heatmap, understood which features are in top ranked.

* **Class imbalance treatment**

Not applicable for the given dataset. Enough data already exists in the dataset.