**Project Report**

**CS 412 - Introduction to Machine Learning**

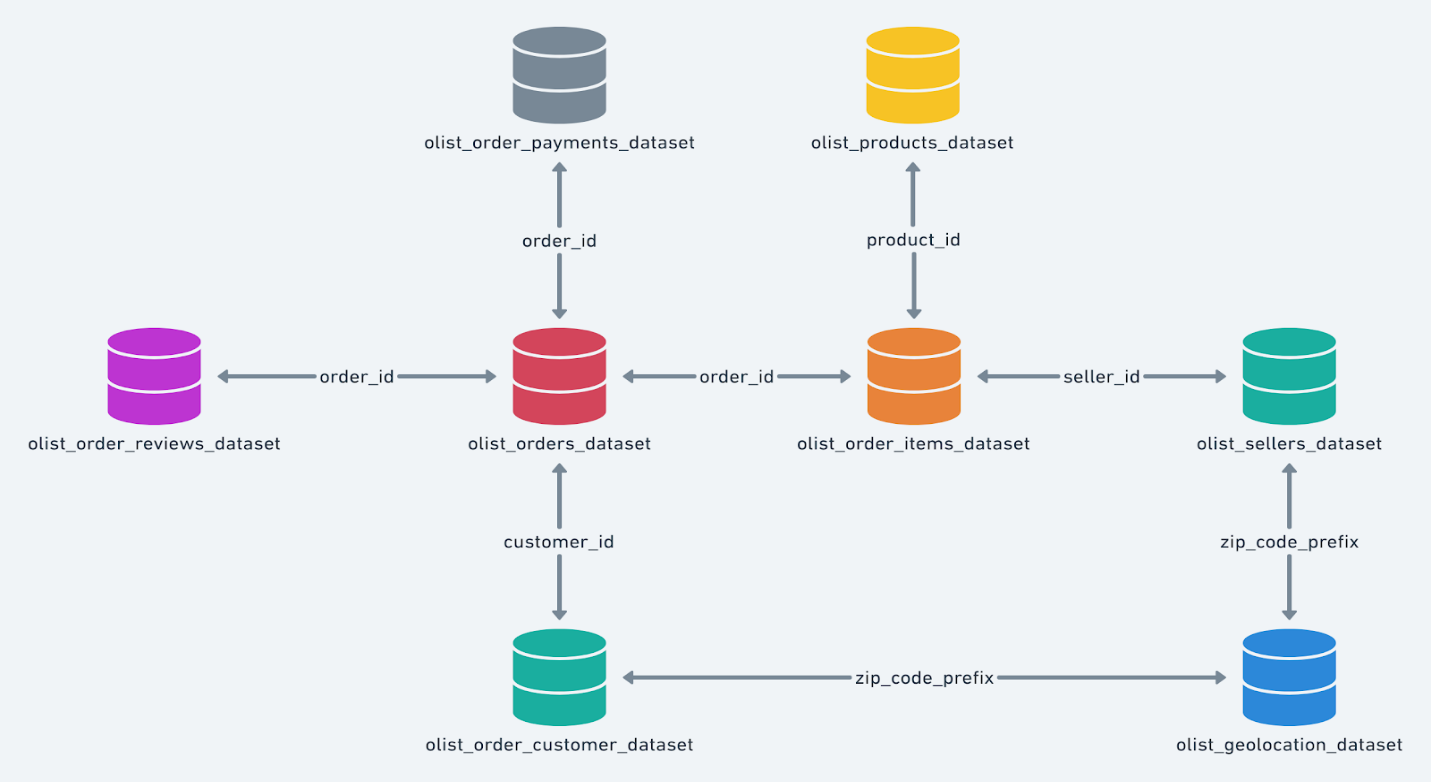
**Machine Learning Tasks on Brazilian E-Commerce Dataset**

**Summary:**

**Dataset:**

The dataset that we used for our machine learning tasks is Brazilian E-commerce public dataset by Olist[1]. Olist is an online store in Brazil using which merchants are able to sell their products and ship them directly to the customers using Olist’s  logistic partners.

The dataset has transactional details of 990,000 orders placed through the website. The dataset has details about customers, sellers, products, order delivery, order payments and reviews of the products by the customers in Portuguese. There are 9 csv files in total, each representing a specific table. The data schema is shown as below:



**Regression:**

**Task** : Predict sales price for a product category in a particular state in a given month.

**Pre-processing** :

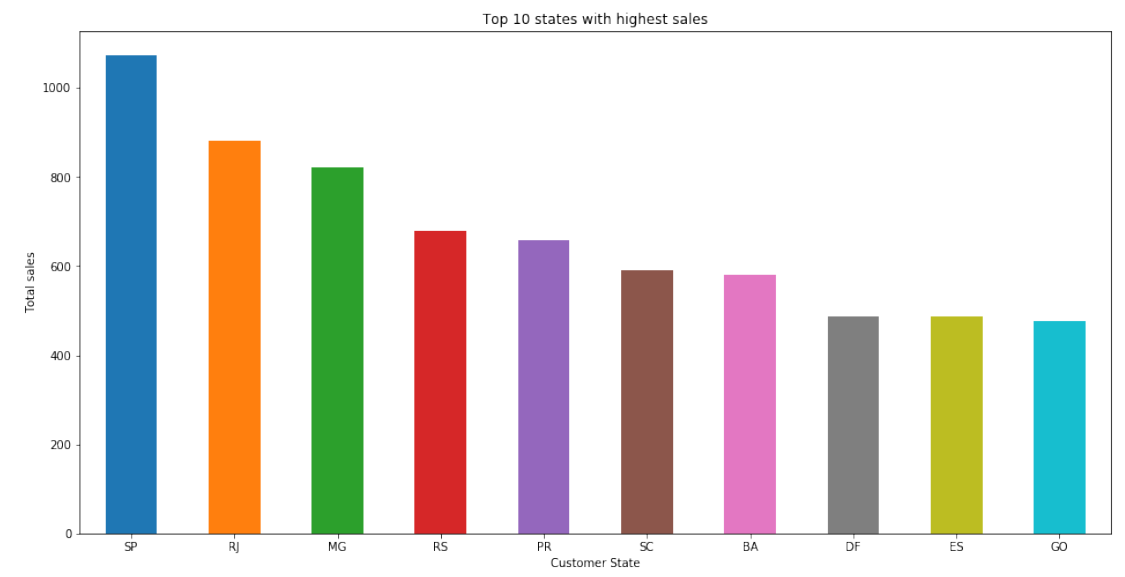
To perform the regression task, we needed overall sales price of previous transaction records based on product category for each state filtered by the month. To get the data in the required format, we had to combine the records from multiple tables. Below is the list of tables from which data was extracted for specific attributes  as part of the pre-processing step:

* Orders - order id, customer id and order purchase time
* Products - product id and product category
* Customer - customer id and customer state
* Order Items - order id, product id and price
* Product Category Translation - product category in english

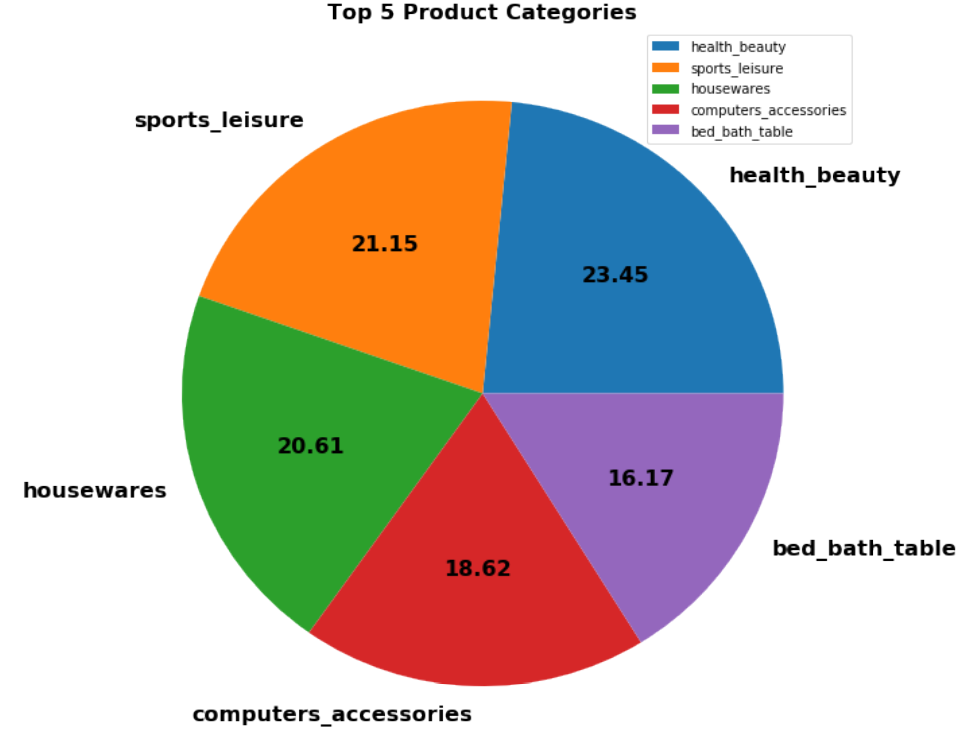
Groupby is done using year, month, state, product category on price and label encoding is done on the columns state and product category.

**Data Visualization** :

To find out states with the highest sales -



To find out top product categories :



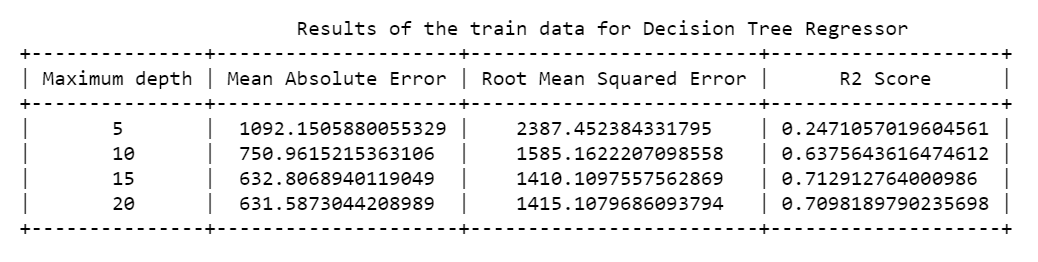
**Models used for the machine learning task**:

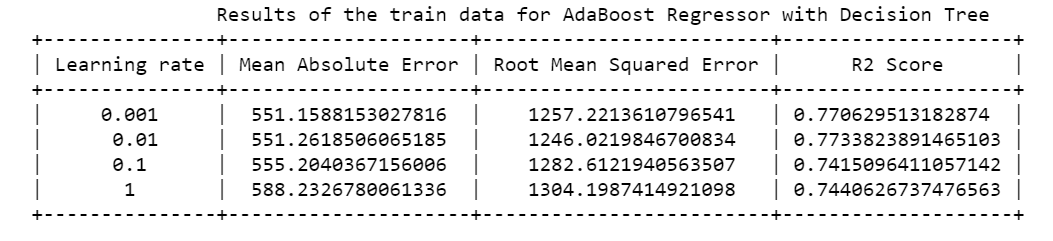
* Decision Tree
* AdaBoost with Decision Tree
* Random Forest

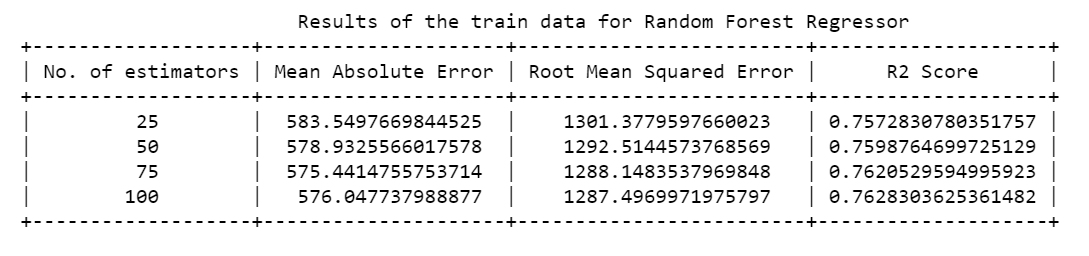
**Approach** :

* 30-Fold cross-validation was performed to tune the hyperparameters of the different classifiers.
* The preprocessed dataset was divided into 80% training set and 20% test set.
* The hyperparameters of the classifiers were tuned based on their performance on the validation set.
* The classifiers were again trained using the entire 80 % training set with the best hyperparameters found.

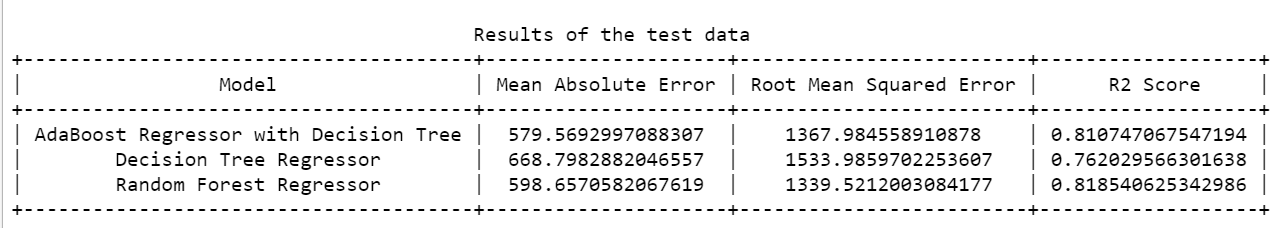
**Evaluation metrics and hyperparameters used (Hyperparameter tuning) :**







**Resulting performances of the 3 classifiers :**



**Challenges faced :**

* Data was skewed. Very high sales in some states and very low in the others.
* The variance in sales was very high ranging from $4 to $40000  leading to higher values while evaluating the mean squared error metric.
* Hence, metrics  mean absolute error and R2 (for evaluating variance) were also used for evaluation.