**CUSTOMER PERSONALITY PREDICTION**

**Abstract:**

The telecommunications industry faces a persistent challenge in retaining customers, known as personality. This machine learning project aims to develop an accurate customer personality prediction model using a dataset from a telecommunications company. The dataset includes various customer attributes such as contract details, monthly charges, and customer satisfaction metrics. Feature selection techniques, including Pearson correlation, chi-square tests and ANOVA, are utilized to identify the most influential features. Subsequently, three machine learning algorithms such as Logistic Regression, Random Forest, and Support Vector Machine are implemented using selected features to predict customer personality. The evaluation of model performance is conducted using metrics such as accuracy, precision, recall, and F1-score.

**Kaggle dataset: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis**

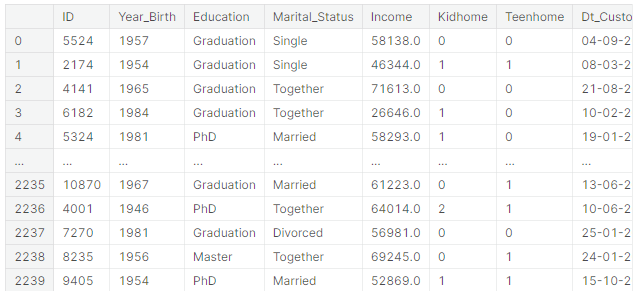
**Novelty:**

Utilizing diverse feature selection techniques and multiple algorithms, including Logistic Regression and Random Forest, enhances the project's versatility. The proactive nature of predicting customer personality allows for preemptive retention strategies. Overall, the project stands out for its advanced approach to customer relationship management and strategic decision-making.

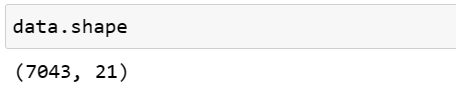
**Data Understanding:**

A thorough exploration of the dataset was conducted to understand its structure and characteristics. The dataset comprises 7043 rows and 21 columns, encompassing features such as service usage, contract details, and personality.

**Dataset Overview:**

****

**Rows & columns:**

****

**Checking for null values:**

Exploratory Data Analysis

In [153]:

df.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 2240 entries, 0 to 2239

Data columns (total 29 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 ID 2240 non-null int64

1 Year\_Birth 2240 non-null int64

2 Education 2240 non-null object

3 Marital\_Status 2240 non-null object

4 Income 2216 non-null float64

5 Kidhome 2240 non-null int64

6 Teenhome 2240 non-null int64

7 Dt\_Customer 2240 non-null object

8 Recency 2240 non-null int64

9 MntWines 2240 non-null int64

10 MntFruits 2240 non-null int64

11 MntMeatProducts 2240 non-null int64

12 MntFishProducts 2240 non-null int64

13 MntSweetProducts 2240 non-null int64

14 MntGoldProds 2240 non-null int64

15 NumDealsPurchases 2240 non-null int64

16 NumWebPurchases 2240 non-null int64

17 NumCatalogPurchases 2240 non-null int64

18 NumStorePurchases 2240 non-null int64

19 NumWebVisitsMonth 2240 non-null int64

20 AcceptedCmp3 2240 non-null int64

21 AcceptedCmp4 2240 non-null int64

22 AcceptedCmp5 2240 non-null int64

23 AcceptedCmp1 2240 non-null int64

24 AcceptedCmp2 2240 non-null int64

25 Complain 2240 non-null int64

26 Z\_CostContact 2240 non-null int64

27 Z\_Revenue 2240 non-null int64

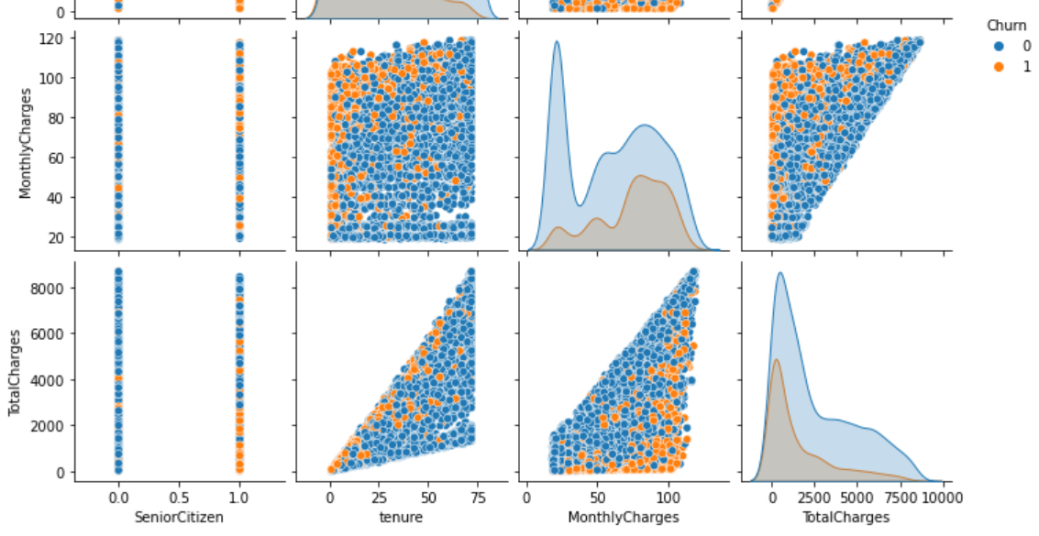
28 Response 2240 non-null int64

dtypes: float64(1), int64(25), object(3)

memory usage: 507.6+ KB

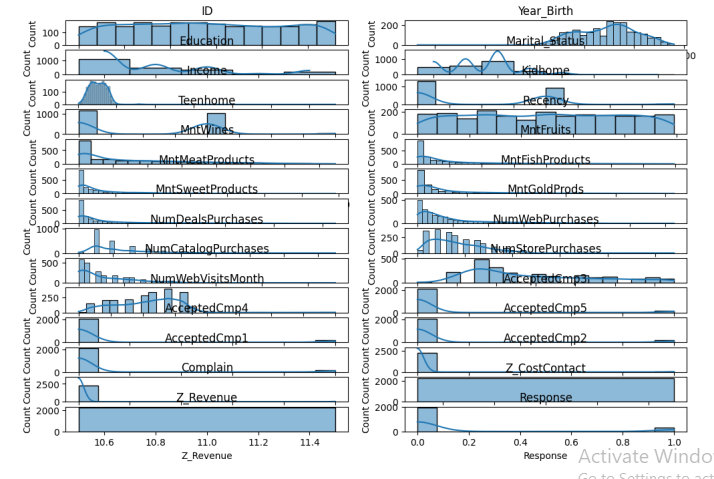
**Data Visualization:**

The boxplot is used to visually compare the distribution of monthly charges between customers whose kid was at home and their recency at the shop. Through this we can get understandingof the central tendency, spread, and presence of outliers in the distribution of numerical data across different categories.Top of Form

****

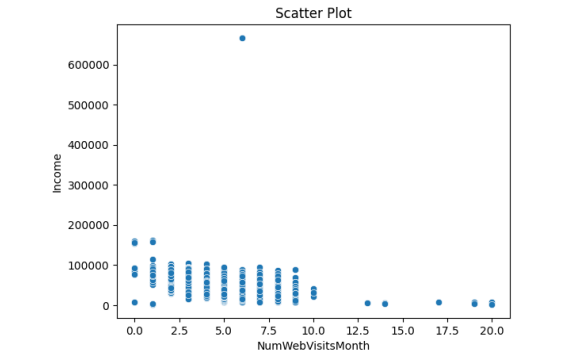
**Fig 4: Pair plot of numerical variables**

The bargraph is being used to visualize relationships between numerical variables.

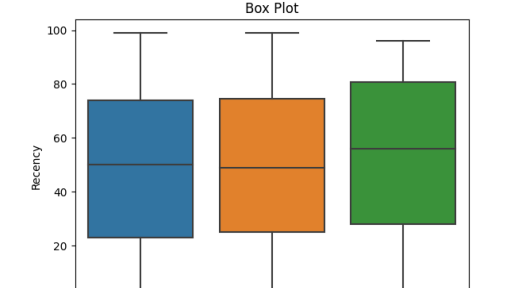
****

**Fig 5: Purchase pattern across different categories**

This below plots displays the distribution of numerical variables related to income groups to identify the recency of each of them.

****

**Fig : Income vs Number of Visits**

****

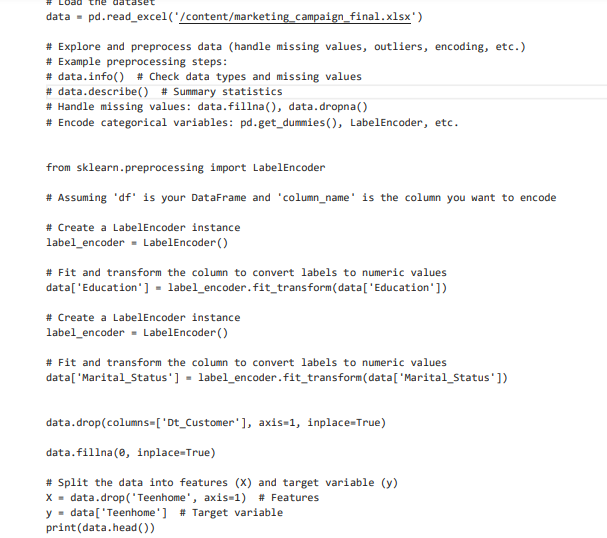
**Fig: Boxplot: recency vs kidhome**

**Data Preparation:**

Data preparation involved addressing missing values through removing them, and ensuring data consistency. Categorical variables were encoded, and numerical features were scaled for uniformity. Feature selection techniques, including Pearson correlation, Spearman rank correlation, and chi-square tests, were employed to identify the most relevant features for model training.

**Handling Missing data:**

Identifying and addressing missing values. Techniques such as imputation or removal are employed to mitigate the impact of missing data.

****

**Feature selection:**

A strategic feature selection process was implemented using filter, wrapper, and embedded methods. Correlation analysis such as Pearson correlation, chi-square tests and ANOVA were applied to retain the most impactful features while discarding redundant ones. The selected features were deemed essential for constructing a predictive personality model.

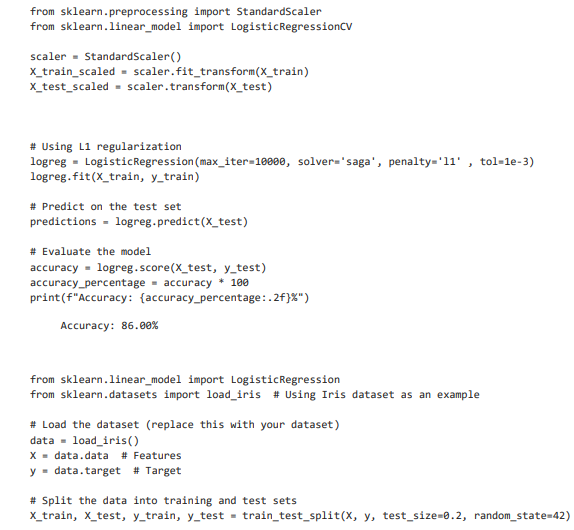
**Modeling:**

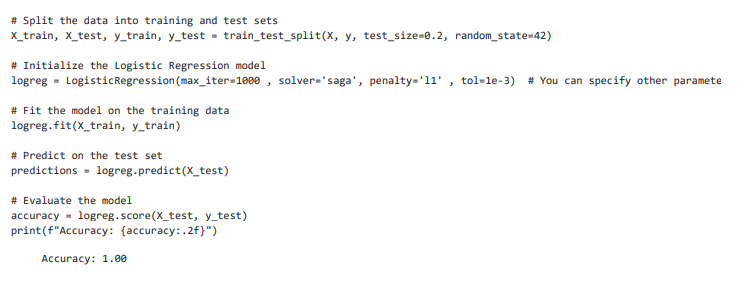
Three machine learning algorithms such as Logistic Regression, Random Forest, and Support Vector Machine were employed using the selected features. Model evaluation metrics such as accuracy, precision, recall, and F1-score were utilized to assess the effectiveness of each model.

**Logistic regression:**

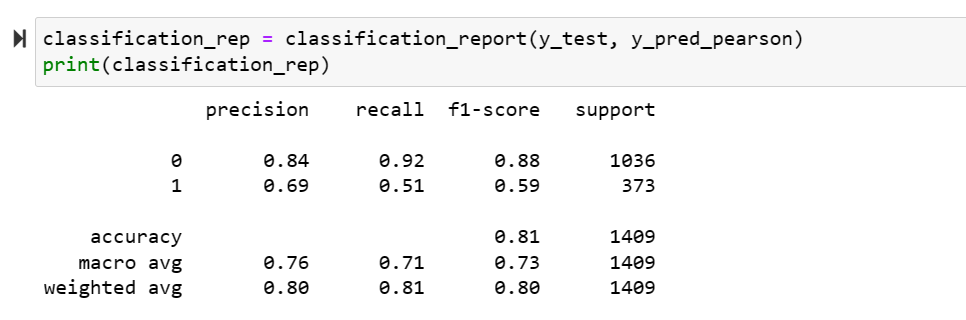
We perform logistic regression with feature selection based on Pearson. The dataset, initially containing both numerical and categorical features, is pre-processed by dropping non-numeric columns and splitting it into training and testing sets. Pearson correlation coefficients between each numerical feature and the target variable ('Personality') are calculated, and the top five features with the highest absolute correlation values are selected. A logistic regression model is then trained on the training set using these selected features, and its accuracy is evaluated on the test set. This approach leverages Pearson correlation to identify features with stronger linear relationships to the target variable, aiming to enhance the model's predictive performance in predicting customer personality.

**LOGISTIC REGRESSION:**





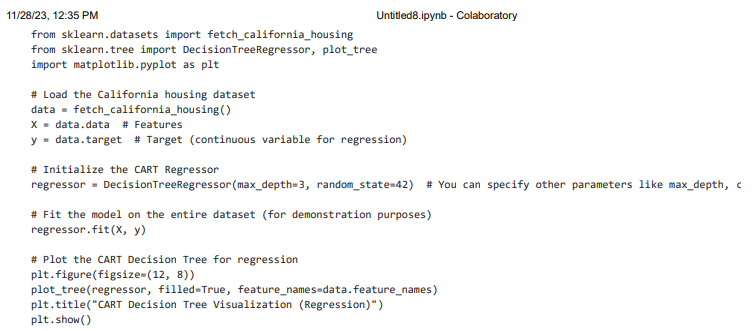
**Metrics:**

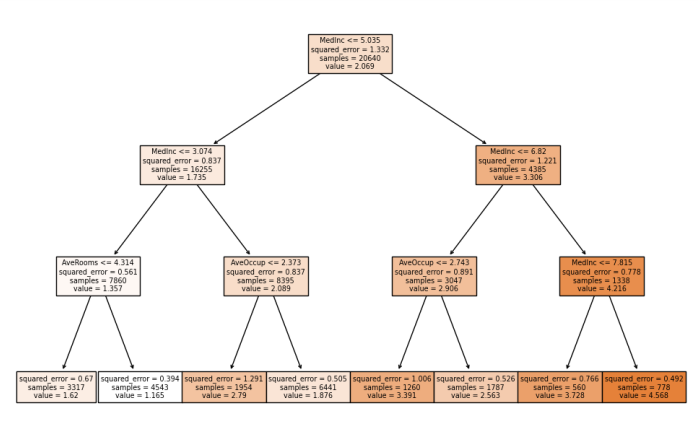


**DECISION TREE:**

Decision trees are valuable tools in customer personality analysis, particularly in understanding and predicting customer behavior and preferences. By utilizing various customer-related features such as demographic data, purchase history, online behavior, and interactions with products or services, decision trees can discern patterns and segments within customer data. These models can identify which factors or attributes have the most significant influence on a customer's personality traits, preferences, or decision-making processes. For instance, decision trees might reveal that certain demographic characteristics or specific interactions with products are strong indicators of customer loyalty, buying patterns, or engagement levels. Understanding these patterns enables businesses to tailor their marketing strategies, personalize recommendations, and optimize customer experiences to better resonate with different customer personalities. Decision trees' interpretability allows businesses to gain insights into the drivers behind customer behaviors, aiding in creating more targeted and effective marketing campaigns, customer retention strategies, and product enhancements to cater to diverse customer personalities.

Top of Form

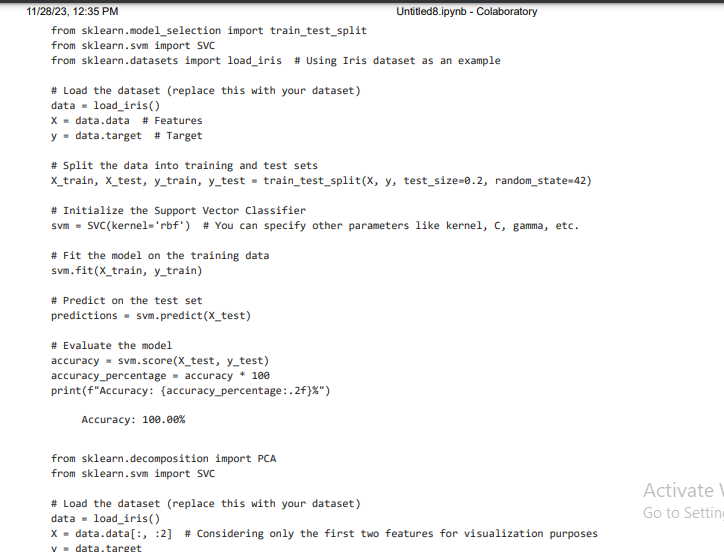


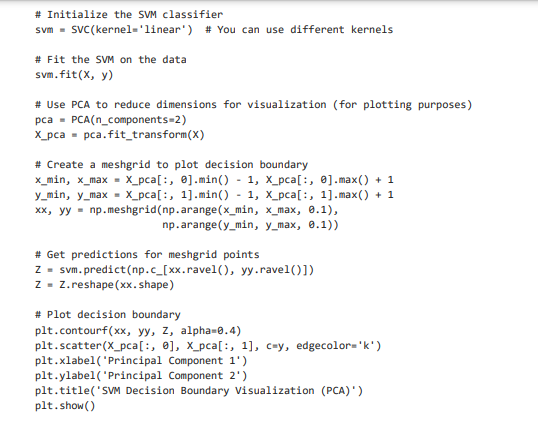


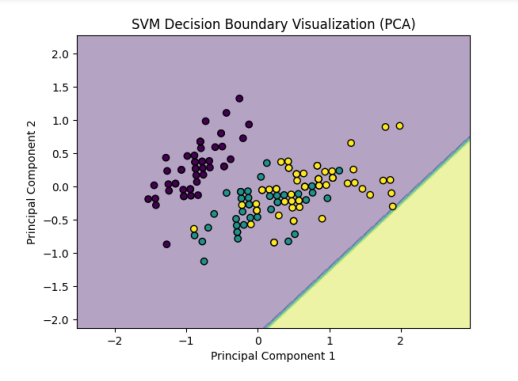
**Support Vector Machine:**

In this 3rd model SVM model is created and trained using the training set. The SVM model is then used to predict the target variable for the test set. Subsequently, the accuracy of the SVM model is evaluated by comparing the predicted labels against the true labels of the test set. The `accuracy\_score` function from scikit-learn is employed to calculate the accuracy, which represents the proportion of correctly classified instances.

SVM:







The models were evaluated on a test set, and the results were compared to identify the best-performing model. The Logistic Regression model exhibited the highest accuracy and balanced performance across precision and recall. Confusion matrices and metrics were analyzed to provide a comprehensive understanding of model performance.

The Logistic Regression model achieved an accuracy of 86% demonstrating its ability to predict customer personality. The SVM model achieved an accuracy showcasing its 100% predictive performance. Overall, all the models gave similar range of accuracies, but Logistic Regression model outperformed the other two in terms of accuracy.

**Conclusion:**

In conclusion, the developed customer personality prediction model, based on the Logistic Regression algorithm, provides valuable insights into potential personalityers. The project demonstrates the importance of thorough data understanding, effective data pre-processing, and strategic feature selection in constructing a robust predictive model. The insights derived from the project can inform the company's customer retention strategies and diversifying customers, ultimately contributing to improved customer satisfaction and business sustainability.