FLIP ROBO

Rating Prediction Model

FLIP ROBO TECHNOLOGIES
Authored by: Seep Bansal

FLIP ROBO

ACKNOWLEDGEMENT

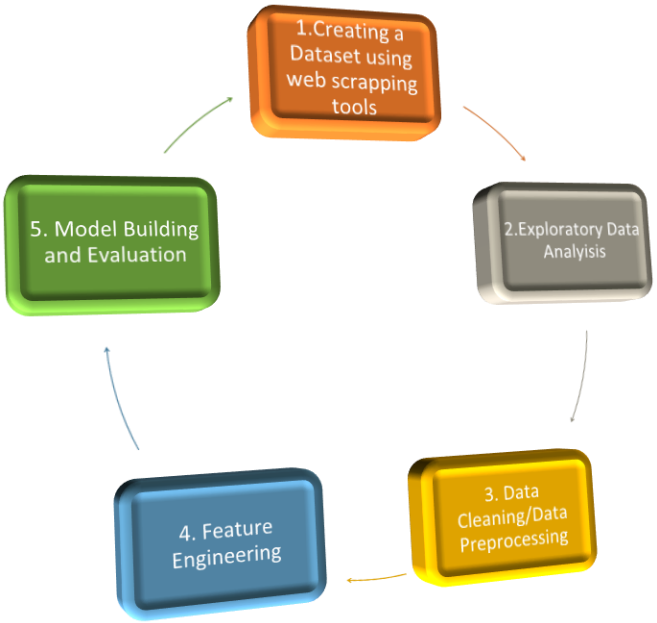
I would like to express my special thanks of gratitude to the team Data Trained and my Mentor MS Deepika Sharma for their exemplary guidance, monitoring and constant encouragement thought the journey of learning Data science and Machine learning techniques. I would also like to express my heartly gratitude to the support team of data trained for their constant support. Last but not the least, I would also like to thank the team of Flip Robo technologies for giving me this opportunity to work on this project and the mentors in Flip Robo Technologies who are constantly guiding me to enhance my knowledge and work. This project helped me not only to learn how to do proper research but also helped me in learning many new things.

INTRODUCTION

E-commerce these days plays a vital role in our life and one of the most important components of e-commerce is Online shopping. Online shopping took a significant segment of retail market during first decade of the 21st century. Interestingly, India is expected to be the third largest Internet market in the world in the next five years. The various online platforms are available where one can shop online are amazon, myntra, flipkart and many more. One can avail various products like electronics, mobile phones, home decors and clothes and many more. These online shopping websites also provides us the facility where a buyers can add their own reviews about the product and the ratings for the same.

PROJECT SUMMARY

Though these online platforms provide us various options to shop online like clothes, electronics, Smart TV and various other appliances reviews and rating for the same plays very important role as it helps the user to make purchase easy. The focus of this project is to develop a machine learning model that can accurately predict the rating of a product based on reviews given by the user. Various machine learning algorithms like Logistic Regression, Support Vector Mechanism, Decision Tree Classifier, Random Forest Classifier, Ada Boost Classifier are implemented and evaluated to predict the rating of the product. The best results are given by Decision Tree Classifier.



MOTIVATION

Deciding upon the rating of the product based on the reviews entered by the user plays a very crucial role as it only helps the user make a perfect purchase but also helps the website to increase the loyal buyer as well as increases its business. Though it a very challenging task to decide upon the rating based on the reviews but on the other hand it is very important too.

ANALYTICAL PROBLEM FRAMING

DATASET

For this project the data is collected from one of the leading platform called www.flipkart.com. The rating, rating summary and the reviews have been collected for the various products like iphone, Smart TV, Smart watches etc. The data is scrapped from the website using the web scrapping tools like selenium and beautiful Soup using Python language.

	Rating	Rating Summary	Review
0	1	Horrible	There is no stand in this pack.
1	1	Worthless	Next day power IC not working
2	1	Unsatisfactory	PLEASE DO NOTE THAT: THE TV DOES NOT COME WITH A TABLE TOP STAND. TECHNICIAN FROM JEEVES PROVIDED B
3	1	Useless product	This model does not come with a table mount.we will have to call samsung customer care to get this. Its been more th
4	1	Absolute rubbish!	Picture quality is very bad
5	1	Absolute rubbish!	Worst flipkart shopping,
6	1	Worthless	Please make sure that if you want Bluetooth facility then don't buy this tv, it don't have Bluetooth connectivity. I am re
7	1	Horrible	Stand missing
8	1	Worthless	I bought this product 11 months ago now itâ€™s not working. I bought this product because itâ€™s a branded one but
9	1	Worthless	Worst experience from flipkart, I will never ever buy products from flipkart hereafter...
10	1	Hated it!	table stand not avaialable
11	1	Worthless	Very bad quality iam both this after 4months tv board fallt

Fig 1: Snapshot of dataset

Feature Description:

Rating: The column contains 5 values i.e 1,2,3,4,5. 1 being the worst rating and 5 being the best.
Rating Summary: It contains a small summary for the rating.
Reviews: It contains the reviews that are added by the customers.

DATA PREPROCESSING/CLEANING

Preprocessing is one of the important steps in building a model. In this phase we usually deal with missing values if any or if there are any unrealistic values. In case of any irrelevant value, we will remove that data. In case if the data loss is huge then removing/dropping of data is not a good practice. We will try to improve the quality of data in this phase so that we can develop a model with high accuracy score. For the dataset, we will first of all fill the missing values if any, then we have most of the data in categorial form we will convert the data in numerical form so that we are able to fed the data into classification algorithms. We can also check if there exist any multicollinearities through VIF FACTOR calculation.

EDA concluding Remarks:

- ✓ There are null values in the dataset.
- ✓ Calculation of VIF Factor to check the presence of multicollinearities among the difference variables.
- ✓ We will be performing standard scaling technique to bring all the columns on the same scale.
- ✓ Skewness of the model was checked.
- ✓ As the dataset was imbalanced the SMOTE technique was followed to balance the dataset.
- ✓ As the dataset is in the categorical form the data label encoding was done to convert the data into numeric form.

HARDWARE AND SOFTWARE REQUIREMENTS

Hardware Requirements: Hardware Requirements followed while developing this model:

- Intel core i5
- 11^{th} generation
- 16 GB RAM
- Windows 10

Software Requirements: Software required are:

- Anaconda Navigator (64-bit Graphical Installer)
- Juypter Notebook
- Microsoft Edge
- Knowledge of Python Language and Machine learning Algorithms

LIBRARIES USED:

```
import pandas as pd  #importing libraries
import numpy as np
import warnings
warnings.filterwarnings("ignore")
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_curve,roc_auc_score
from sklearn.model_selection import GridSearchCV
import statsmodels.api as sm
from scipy import stats
from statsmodels.stats.outliers_influence import variance_inflation_factor
```

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score,classification_report,confusion_matrix
from sklearn.tree import DecisionTreeClassifier
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import roc_curve,roc_auc_score
from scipy.stats import zscore
from sklearn.ensemble import RandomForestClassifier,AdaBoostClassifier
```

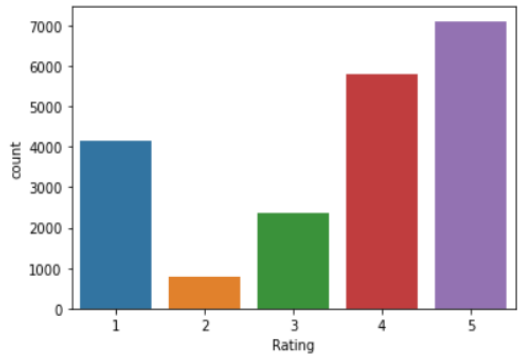
Fig:2 Libraries used

- Import and export of data takes place with the help of pandas
- All the numerical operations are carried out with the help of numpy library.
- Matplotlib.pyplot and seaborn libraries helps in graphical representation of data.
- Warnings library is used to ignore the unwanted warnings
- Sklearn library helps in importing all the machine learning algorithms and evaluation matrix that are required.
- Scipy library helps VIF factor calculation and zscore which is used for removal of outliers.

MODEL DEVELOPMENT/EVALUATION

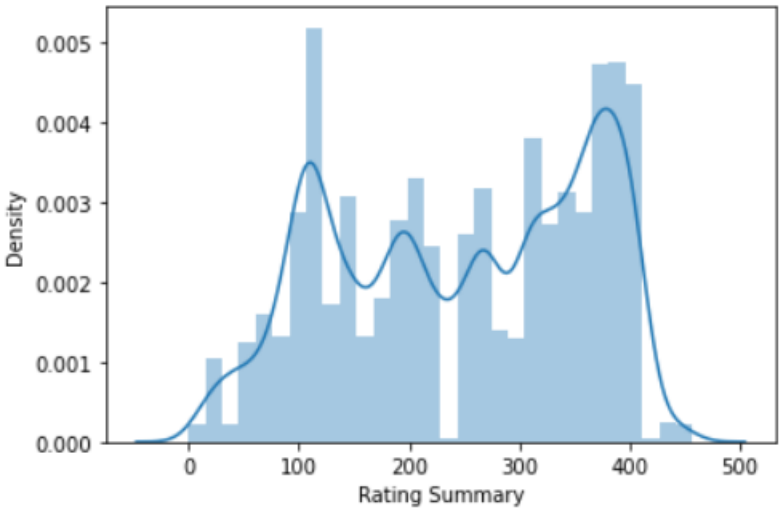
VISUALIZATION

Following are some of the visualizations/Graphical Representations created to interpret the data:

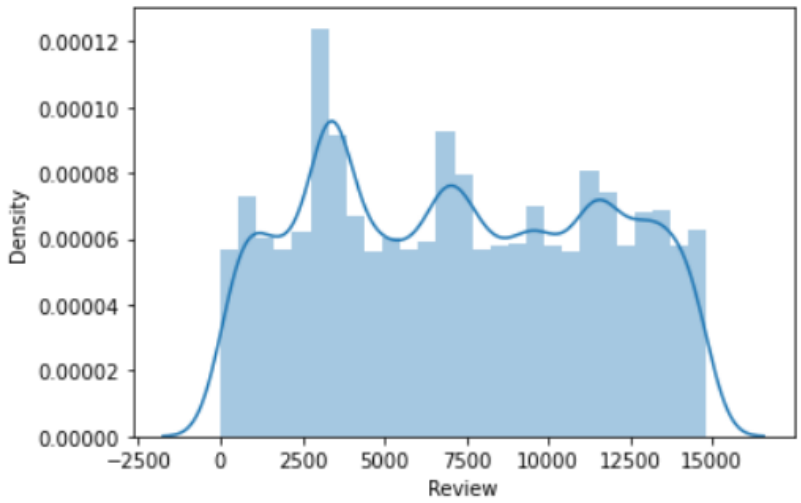


As we could see here that we we have 5 categories in rating and the dataset is imbalalnced we will be performing SMOTE to balance our dataset.

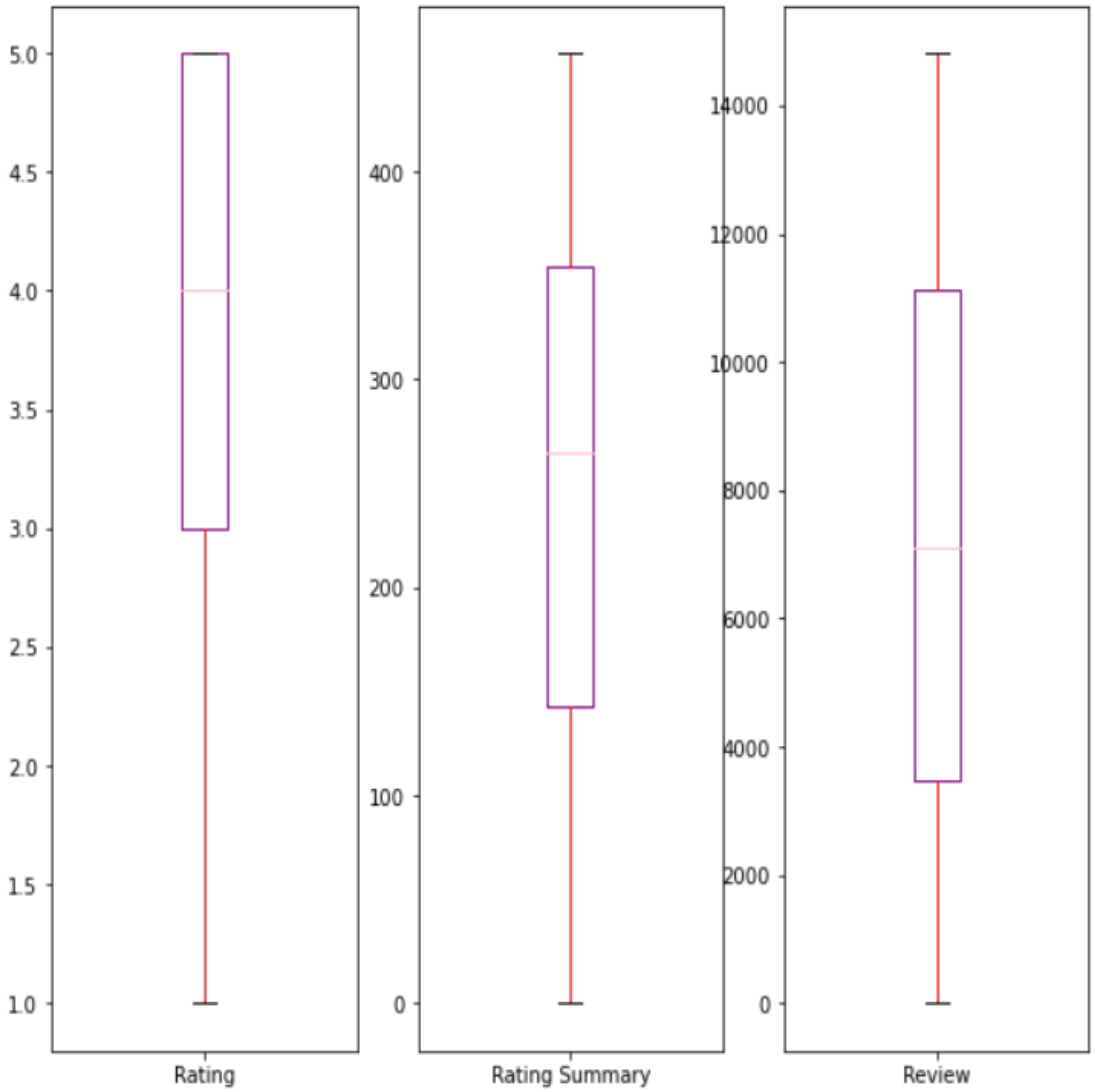
From the above heatmap we can see that rating summary and review are positively correlated to each other.



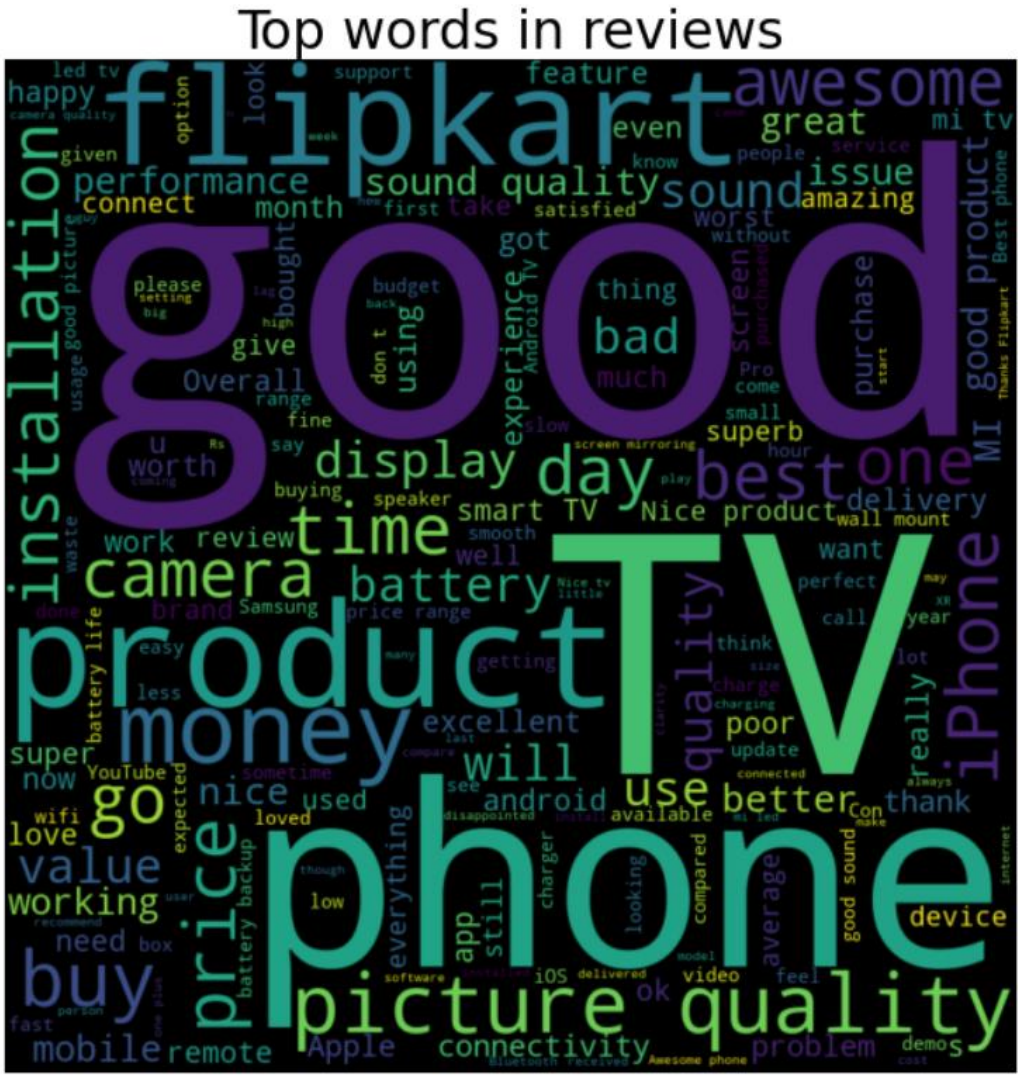
After coverting the textual data into the numeric we could see that data is normarlly distributed.



After coverting the textual data into the numeric we could see that data is normarlly distributed.



From the abaove plot we can say that there are no outliers present in the dataset.

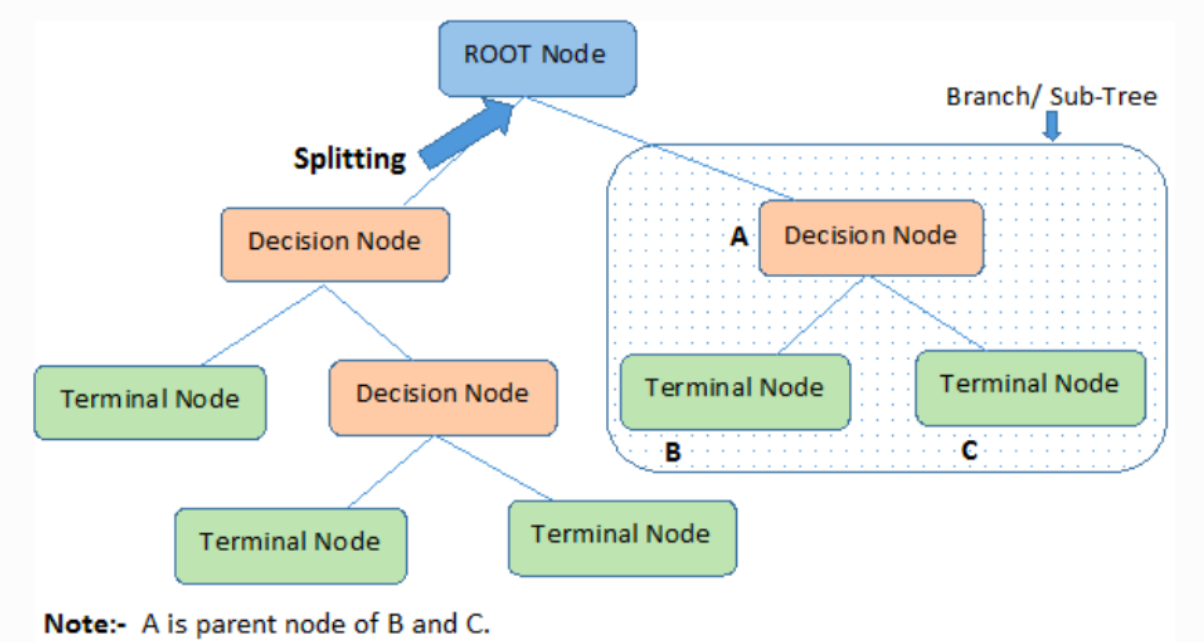


ALGORITHMS USED For Training and Testing the Model:

- Decision Tree Classifier
- Random Tree Forest Classifier
- Support Vector Classifier
- Logistic Regression
- Ada Boost Classifier

CLASSIFICATION EXPERIMENTS

We have performed various classification algorithms like Logistic Algorithm, Decision Tree Algorithm, Ada Boost Classifier, Random Tree Classifier to check the accuracy. The main algorithm which has predicted the best results is the Decision Tree Classifier. Here, the Decision tree classifier is giving us the best results with the accuracy score of 99%.



EVALUATION MATRIX

Evaluation matrix plays a very important role in evaluating the performance of any classification model. The metrices that are evaluated here are the accuracy score, ROC_AUC score, classification matrix which includes F1, recall, Precision and the confusion matrix. Time taken to test the model on dataset plays a very crucial role. Here, for Decision Tree Classifier the accuracy score we are getting is 99% and the classification matrix i.e precision, F1, recall also scores to 99% and the confusion matrix shows that some values lie in the error type 1 and 2 as the model is predicting with 99% accuracy.

```
Accuracy Score 98.14450379533315
              precision    recall  f1-score   support

           1       0.99      0.99      0.99      2168
           2       0.99      0.99      0.99      2123
           3       0.99      0.99      0.99      2142
           4       0.96      0.98      0.97      2107
           5       0.97      0.96      0.96      2131

    accuracy                           0.98     10671
   macro avg       0.98      0.98      0.98     10671
weighted avg       0.98      0.98      0.98     10671

[[2144    6    6    5    7]
 [   7 2098    1    0   17]
 [   5    3 2123    5    6]
 [   5    3    4 2068   27]
 [   4    7    9   71 2040]]
```

Fig 3: Summary of Alogorithm

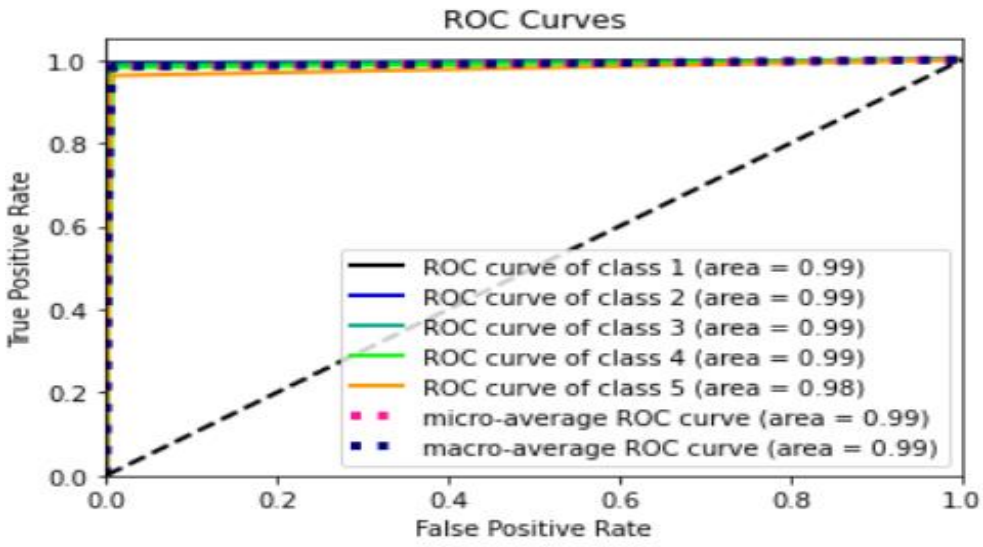


Fig: 4 AUC ROC curve for the Decision Tree Algorithm

CONCLUSION

Deciding upon the rating of the product based on the reviews entered by the user plays a very crucial role as it only helps the user make a perfect purchase but also helps the website to increase the loyal buyer as well as increases its business. Though it a very challenging task to decide upon the rating based on the reviews but on the other hand it is very important too. The focus of this project is to develop a machine learning model that can accurately predict the rating of a product based on reviews given by the user. Various machine learning algorithms like Logistic Regression, Support Vector Mechanism, Decision Tree Classifier, Random Forest Classifier, Ada Boost Classifier are implemented and evaluated to predict the rating of the product. Here, for Decision Tree Classifier the accuracy score we are getting is 99% and the classification matrix i.e precision, F1, recall also scores to 99% and the confusion matrix shows that some values lie in the error type 1 and 2 as the model is predicting with 99% accuracy.