**A Project Report**

**on**

**“Online News Popularity Prediction”**

**submitted for the partial fulfilment of the award of**

**the degree of**

**BACHELOR OF TECHNOLOGY**

**in**

**Artificial Intelligence and machine Learning**

**(4th Semester)**

****

**Submitted to: Submitted by:**

**Dr. Amit Choudhary       Name: Yajat Dev Prajapati**

**Asst. Prof., USAR Batch: AIML B1 - A**

**Roll No: 02519011621**

**University School of Automation and Robotics**

**GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY**

**(EAST DELHI CAMPUS)**

**SURAJMAL VIHAR, NEW DELHI-110032**

Online News Popularity Prediction using Machine Learning

**YAJAT DEV PRAJAPATI 02519011621 AIML - B1**

**Abstract**

This project aims to develop a classification model using various machine learning algorithms, including AdaBoost Classifier, Logistic Regression, Random Forest, GaussianNB, Support Vector Classifier, KNeighborsClassifier, and Decision Tree. The goal is to predict the popularity of online news articles using the UCI Online News Popularity dataset.

By analyzing features such as the number of words, social media shares, article category, sentiment analysis scores, and others, the model provides valuable insights for publishers and content creators to optimize their content strategy.

The dataset is preprocessed, relevant features are extracted, and the classification models are trained and evaluated. The results and discussions showcase the effectiveness of the proposed methodology in accurately classifying the popularity of online news articles.

**Keywords**

Online news, popularity classification, machine learning, AdaBoost, Logistic Regression, Random Forest, GaussianNB, Support Vector Classifier, KNeighborsClassifier, Decision Tree.

1. **Introduction**

The UCI Online News Popularity dataset is a collection of articles published by Mashable, a popular online news website. It consists of a wide range of features related to the articles, such as the number of words, shares on social media platforms, and various content-related attributes.

The dataset contains a total of 61 attributes, including both numerical and categorical variables. Some of the key attributes include the article title, author, number of images/videos, number of links, number of keywords, and the target variable 'popularity', which represents the number of shares the article received.

The dataset aims to explore the factors influencing the popularity of online news articles. It provides researchers and data scientists with an opportunity to investigate the relationships between different features and the level of engagement an article generates on social media.

By analysing this dataset, one can gain insights into the characteristics of highly shared articles and potentially identify patterns that contribute to their popularity. This information can be valuable for content creators, marketers, and media organizations in understanding audience preferences and optimizing their content strategies to increase online engagement and reach.

1. **Proposed Methodology**
   1. *Dataset*: The UCI Online News Popularity dataset, obtained from Kaggle, serves as the basis for this project. It includes attributes such as the number of words, social media shares, article category, sentiment analysis scores, and others. The dataset is split into training and testing sets for model development and evaluation.
   2. *Preprocessing*: (This step involves removing irrelevant columns, handling missing values, and normalizing or scaling the data as necessary.)

The correlation matrix helps in identifying patterns, relationships, and dependencies among the variables in the dataset. It also provides insights into which variables are strongly/inversely inversely correlated, which can be useful for feature selection, (identifying multi collinearity) or understanding the relationships between different attributes in the dataset.



In data preprocessing initially the drop\_duplicates() was used and there was no change in the shape of the dataset which implied that there were no duplicate data.

Then isna().sum() was used to check if there were any duplicates and it turned out to be true that the dataset also doesn't contain any duplicates.

Then white space characters were also eliminated from the feature names.

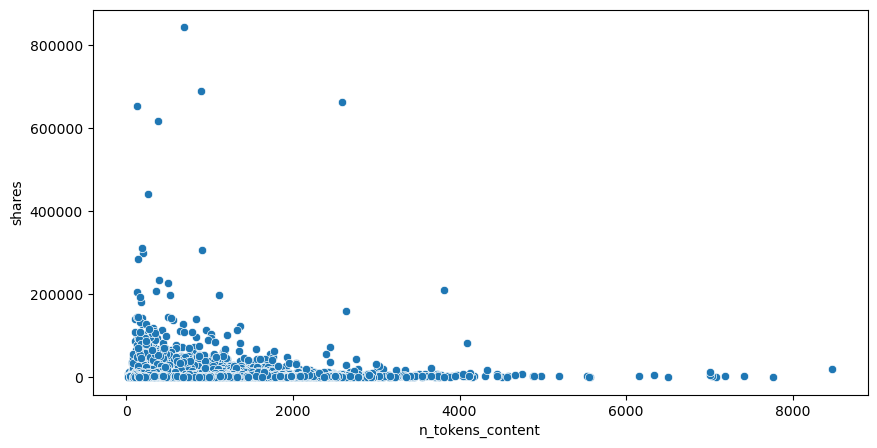
Subsequently, the articles/items without any words were also dropped because they would be of no use in the model training and testing phases.

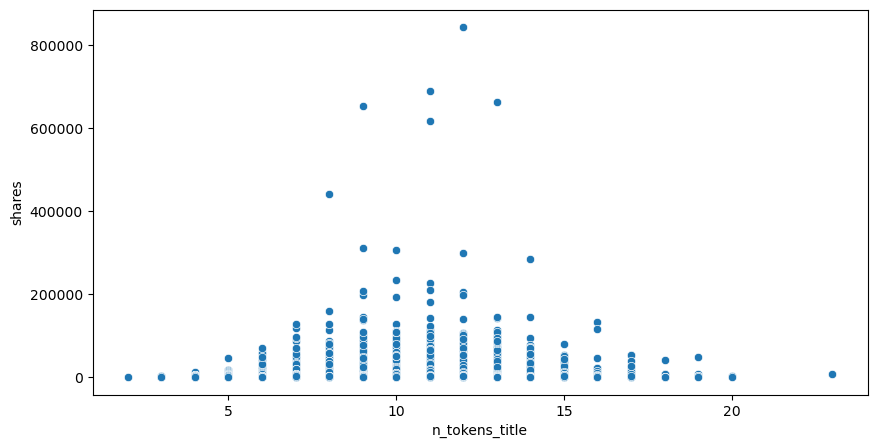
Also, URL being a non numeric attribute was dropped as it won't add any value to the dataset, time-lapse is a non predictive attribute and hence, can be dropped.

n\_non\_stop\_unique\_tokens, n\_non\_stop\_words & kw\_avg\_min are highly correlated and are therefore being dropped.

By using the number of median shares, a new target variable was made: popularity, with articles having below 1400 median shares classified as unpopular otherwise popular.

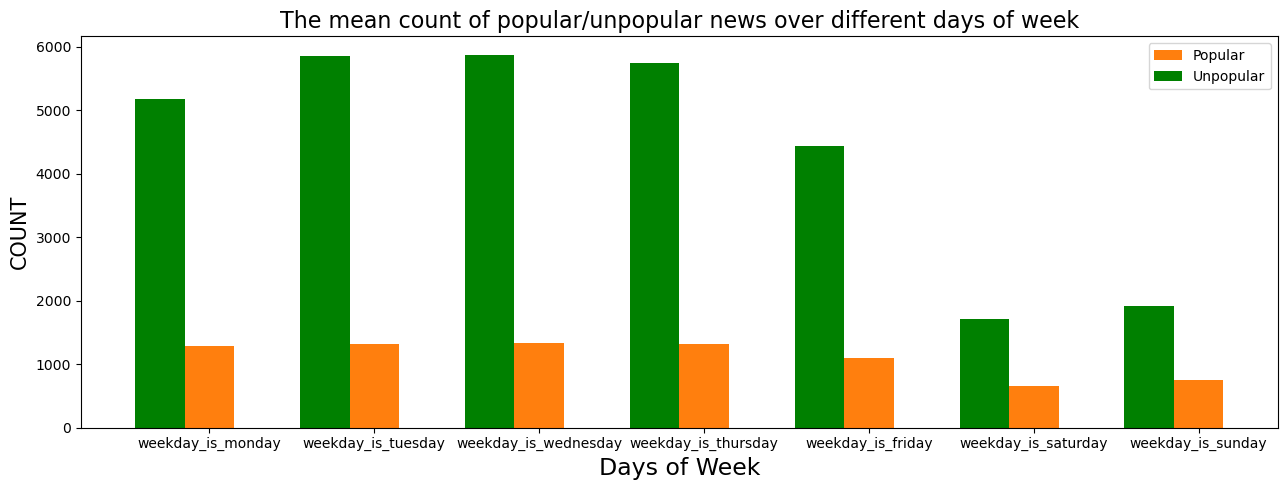
Two scatter plots were generated one between n\_tokens\_content & shares and the other being n\_tokens\_title & shares. This gave an understanding on how the number of shares of an article is affected by the number of words in the context and the number of words in the title.

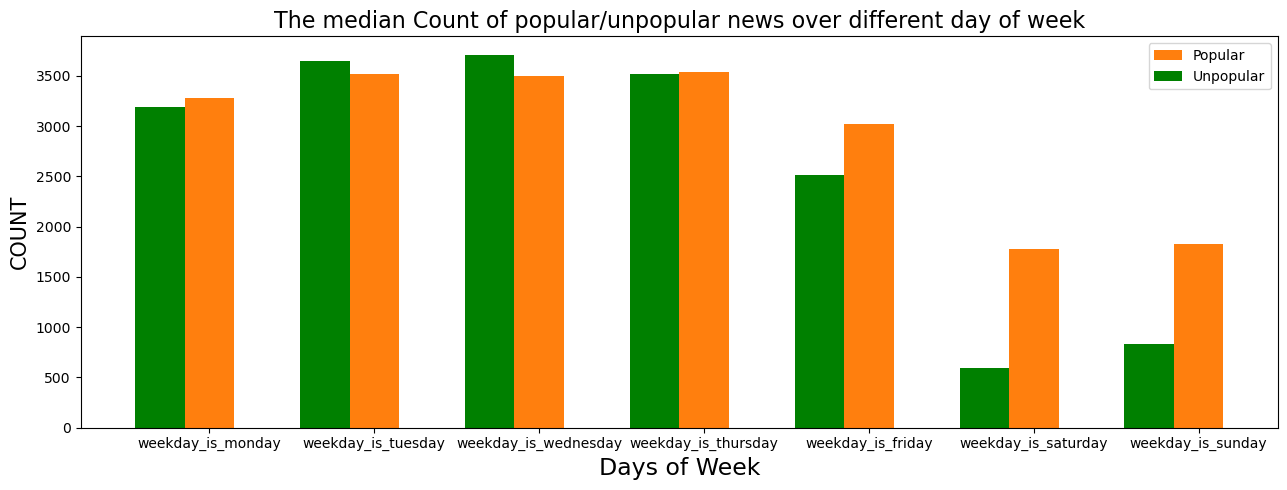


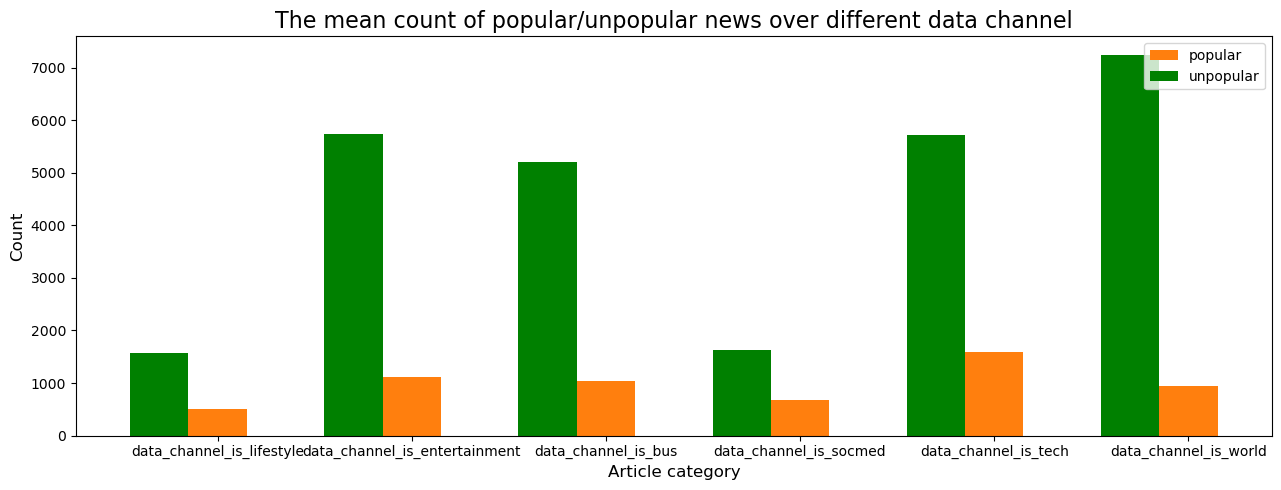


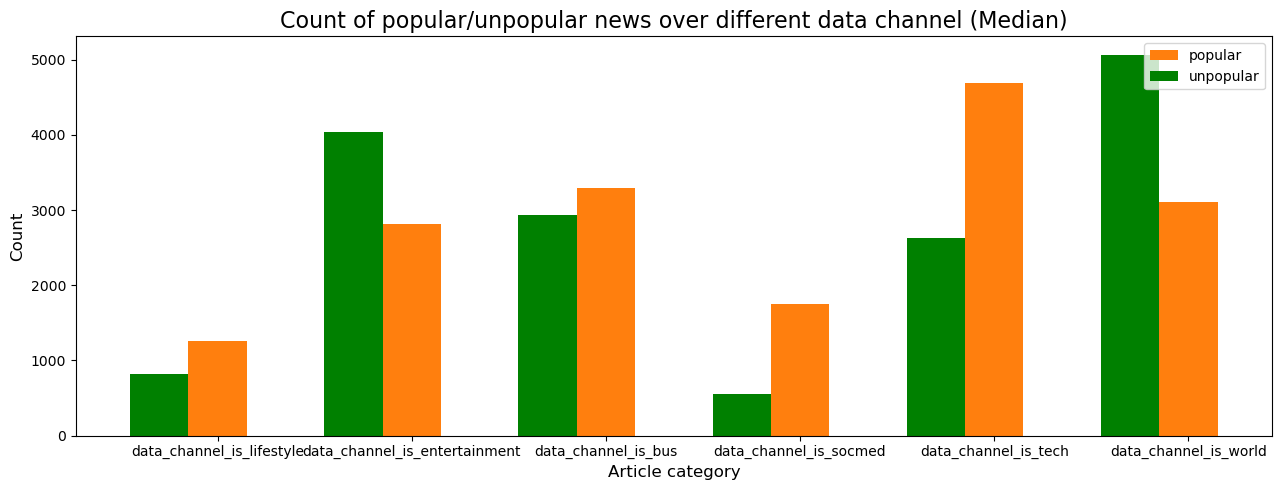
The distribution of popular and unpopular news articles based on the number of shares over different days of the week is analyzed by calculating the sum of shares of popular and unpopular news articles on each day of the week and visualized using a bar plot. This is done based on both the mean and median number of shares of articles.

The distribution of popular and unpopular news articles on the basis of different data channels/categories is also analyzed by calculating the sum of shares of popular and unpopular news articles in each data channel and visualized using a bar plot. Here also it's based on the mean & median number of shares of articles.









Box plots were generated for each attribute in the dataset so as to determine the outliers present visually Then using the Inter-Quartile Range (IQR) method for each attribute the number of outliers, lower limit, upper limit and the IQR was calculated.

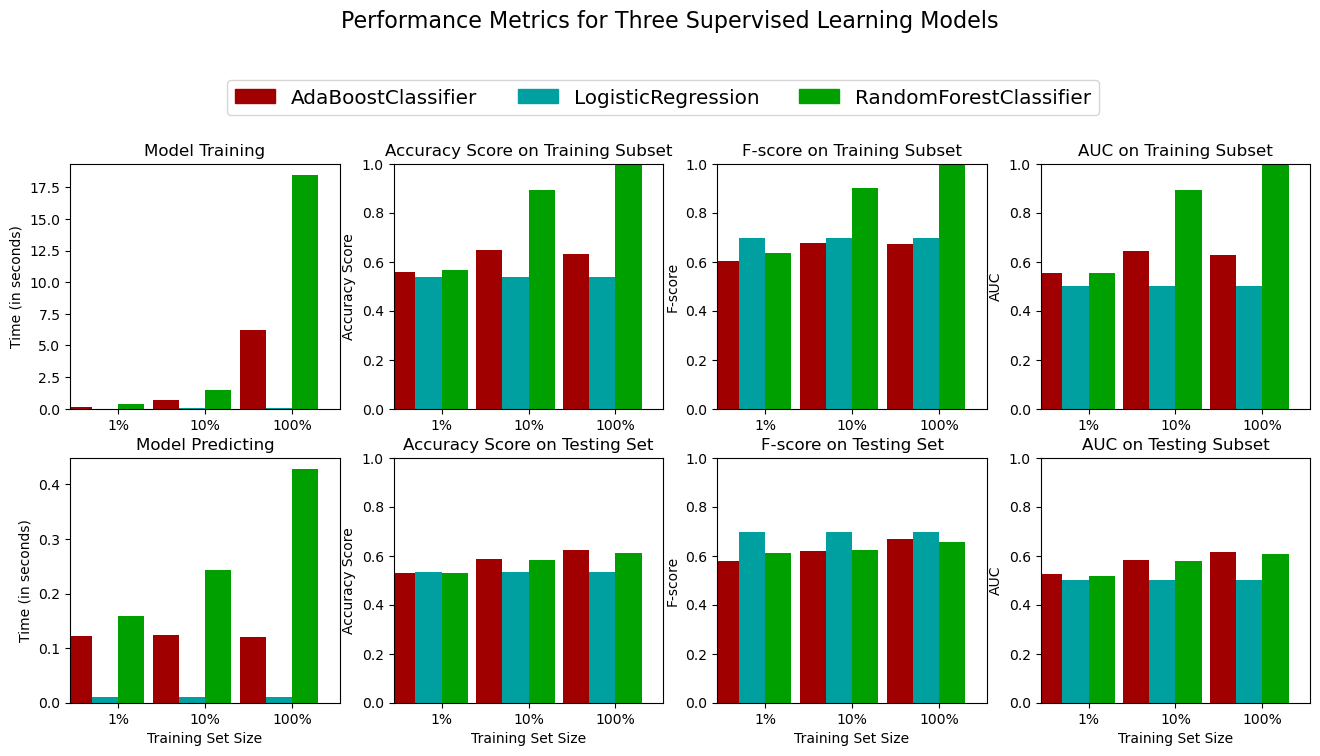
The dataset was then split into two separate data frames one that has numerical features and the other that have categorical features. It allows for specific handling and transformation of numerical and categorical data as needed.

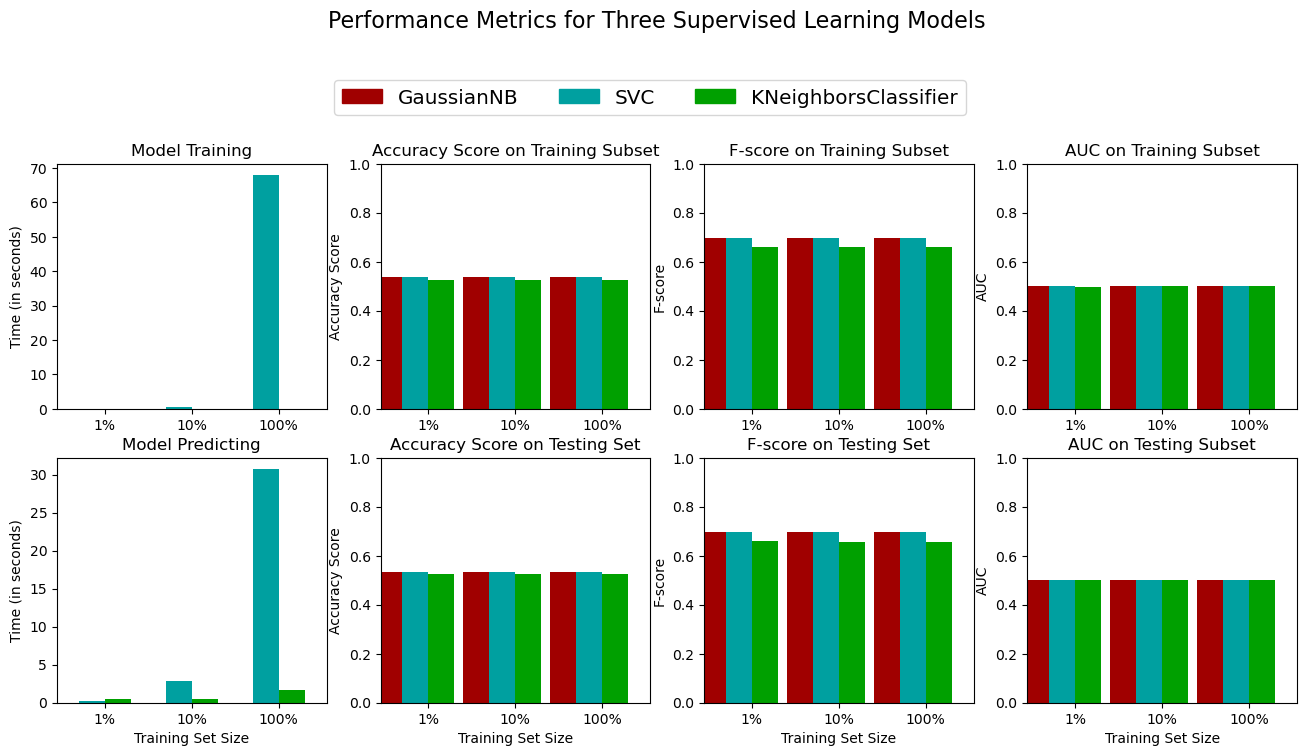
All the negative values were converted into positive values and then the Box-Cox method was applied to handle the outliers.

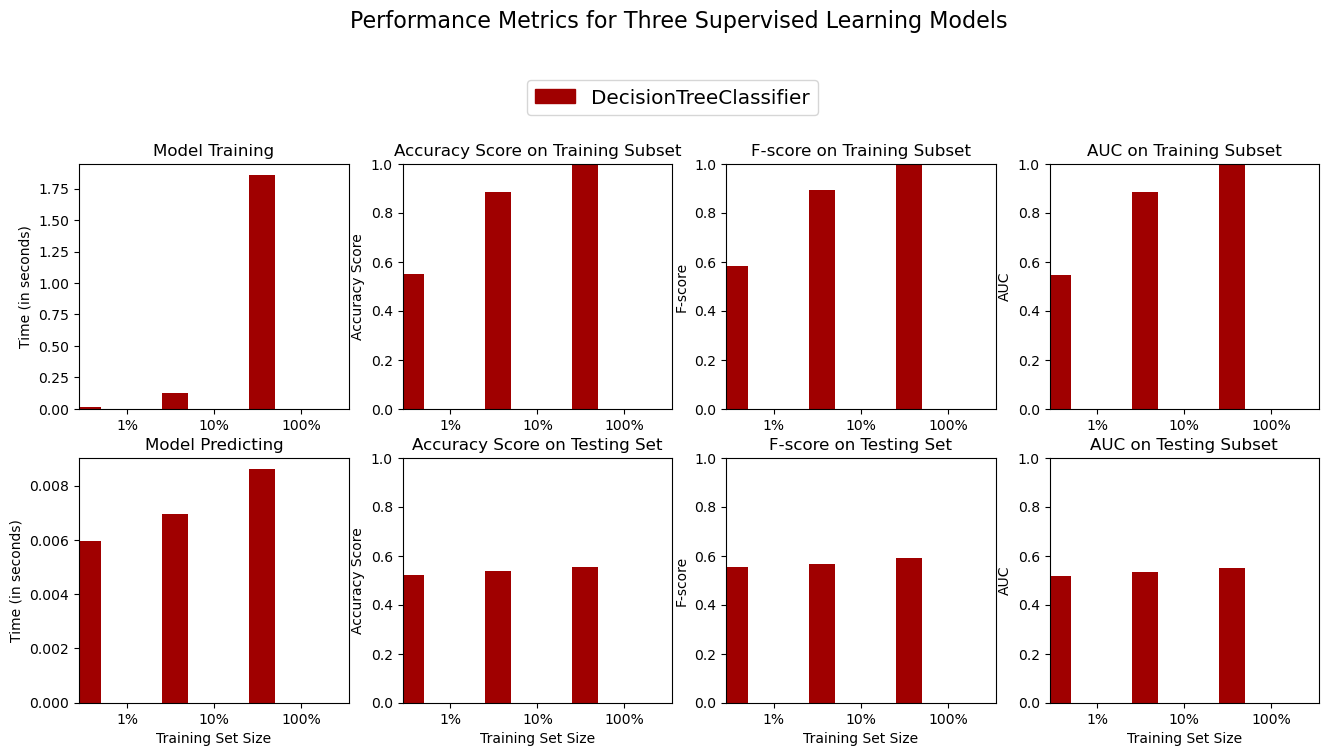
* 1. *Model Training and Evaluation*: Various classification models, including AdaBoost Classifier, Logistic Regression, Random Forest, GaussianNB, Support Vector Classifier, KNeighborsClassifier, and Decision Tree, are trained. Each model is trained on the training set (and tuned be further using appropriate hyperparameters).

The models' performance is evaluated using metrics such as accuracy, F1-score and AUC (precision and recall). The experimental results demonstrate that all the employed classifiers achieved credible enough accuracy in predicting the popularity of online news articles.

AdaBoost Classifier and Random Forest outperformed other classifiers, achieving the highest accuracy rates of 62.2% and 61.2%, respectively. Logistic Regression and Decision Tree also exhibited satisfactory performance with accuracy rates of 53.36% and 55.46%, respectively. GaussianNB, Support Vector Classifier, and KNeighborsClassifier achieved relatively lower accuracy rates ranging from 51% to 54%.





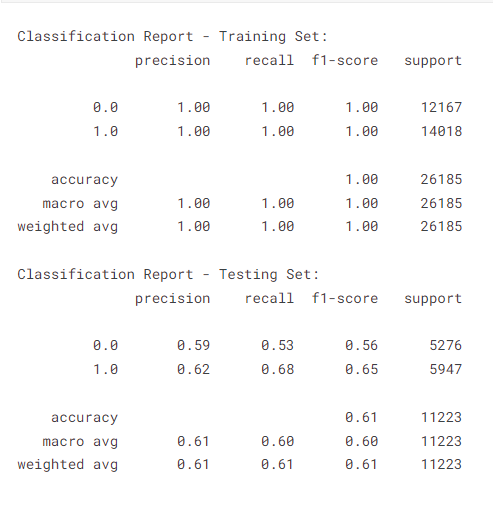


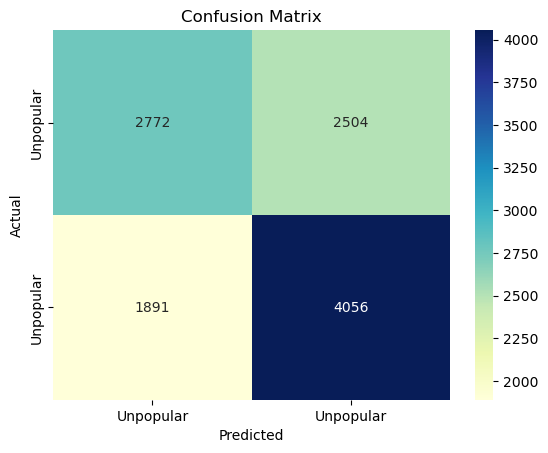
*Feature Extraction*: Furthermore, the feature importance analysis revealed that the number of words, social media shares, article category, and sentiment analysis scores significantly influenced the prediction of article popularity across all classifiers. These findings indicate the importance of content length, social engagement, and topic relevance in determining the popularity of online news articles.

Meaningful features are extracted from the preprocessed data. This includes techniques like RFE, TF-IDF for textual features, statistical features, sentiment analysis scores, and any other relevant information that can contribute to the prediction of article popularity.

1. **Conclusion & Future Work**

In conclusion, the proposed methodology successfully applies various machine learning algorithms, including AdaBoost Classifier, Logistic Regression, Random Forest, GaussianNB, Support Vector Classifier, KNeighborsClassifier, and Decision Tree, to classify the popularity of online news articles using the UCI Online News Popularity dataset. The models exhibit high enough accuracy rates (especially Random Forest after performing feature extraction) and provide valuable insights into the factors influencing article popularity.





For future work and advancements we can focus on ensemble methods, hyperparameter tuning, and exploring additional features to further enhance the accuracy of the classification models.

The findings from this project can assist publishers and content creators in making more data-driven decisions to improve the reach and engagement of their online news articles.

1. **References:**
2. <https://www.researchgate.net/publication/339909346_A_Comprehensive_Review_on_Online_News_Popularity_Prediction_using_Machine_Learning_Approach>
3. <https://www.ijraset.com/research-paper/online-news-articles-popularity-prediction-system>
4. <https://scikit-learn.org/stable/supervised_learning.html>
5. <https://youtu.be/A3gClkblXK8>