# **Classfication Report**

### **Online Articles popularity**

# **1. Preprocessing Techniques:**

In this section, we detail the preprocessing techniques applied to our dataset and their implementation:

* Data Cleaning: We removed any duplicate entries and handled missing values by using mode function
* Feature Scaling: data scaled using min max scaler .
* Feature Encoding: in our data set we have 3 categorical columns (“channel type ”,”weekday”,” isWeekEnd”) and we used LabelEncoder and fit transform to make it numerical data .
* Outlier Detection/Removal: we only droped “url” and “titles” and "time delta"an columns because it’s unique for every row

# **2.Features Used in Classfication Models:**

Features Selection (drop mincorrelation)

top feauter : ([' kw\_avg\_avg', ' LDA\_02', 'isWeekEnd', ' kw\_min\_avg', ' num\_hrefs', ' LDA\_03', ' kw\_max\_avg', ' num\_imgs', 'weekday', ' self\_reference\_avg\_sharess', ' global\_subjectivity', ' LDA\_04', ' self\_reference\_max\_shares', ' self\_reference\_min\_shares', ' num\_keywords', ' LDA\_01', ' n\_non\_stop\_unique\_tokens', ' rate\_negative\_words', ' global\_sentiment\_polarity']

# **3. Classfication Techniques Used:**

* Random Forest Classifier:
* Gradient Boosting
* LogisticRegression :
* Support Vector Machine (SVM):

# **4. Differences Between Models:**

* **Random Forest:** Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees.It randomly selects subsets of features at each split in the decision trees, which helps to reduce overfitting and increase robustness.Random Forests are less prone to overfitting compared to individual decision trees.They can handle both classification and regression tasks.

n\_estimators=50, max\_depth=None

Random Forest Accuracy (tuned): 0.4596474203461103

Random Forest Training Time (tuned): 28.04355478286743

Random Forest Testing Time (tuned): 0.23241138458251953 ----------------------------------------------------- n\_estimators=50, max\_depth=10

Random Forest Accuracy (tuned): 0.4661167717936277

Random Forest Training Time (tuned): 23.682842254638672

Random Forest Testing Time (tuned): 0.14710140228271484 ----------------------------------------------------- n\_estimators=50, max\_depth=20

Random Forest Accuracy (tuned): 0.46207342713892935

Random Forest Training Time (tuned): 31.54683518409729

Random Forest Testing Time (tuned): 0.30992603302001953 ----------------------------------------------------- n\_estimators=100, max\_depth=None

Random Forest Accuracy (tuned): 0.47016011644832606

Random Forest Training Time (tuned): 71.473788022995

Random Forest Testing Time (tuned): 1.6842584609985352 ----------------------------------------------------- n\_estimators=100, max\_depth=10

Random Forest Accuracy (tuned): 0.4725861232411451

Random Forest Training Time (tuned): 38.27916693687439

Random Forest Testing Time (tuned): 0.22829961776733398 -----------------------------------------------------

Random Forest Accuracy (tuned): 0.47711466925440726 Random Forest Training Time (tuned): 88.08925294876099 Random Forest Testing Time (tuned): 0.6250133514404297

* **Gradient Boosting:** Gradient Boosting is also an ensemble learning method, but instead of training multiple models independently like in Random Forests, it builds trees sequentially.Each tree corrects the errors made by the previous one, focusing on the mistakes of the preceding model.Gradient Boosting typically uses shallow trees as weak learners, which are combined to create a strong learner.It can often yield higher predictive accuracy than Random Forests but might be more prone to overfitting if not properly tuned.

Gradient Boosting Accuracy: 0.4735565259582727

Gradient Boosting Training Time: 420.0677237510681

Gradient Boosting Testing Time: 0.5921001434326172

* **LogisticRegression** :Logistic Regression is a linear model used for binary classification problems.It models the probability that a given input belongs to a particular class using a logistic Despite its name, logistic regression is a classification algorithm, not a regression algorithm.It's relatively simple, interpretable, and efficient for small to medium-sized datasets

Logistic Regression Accuracy: 0.45511887433284814

Logistic Regression Training Time: 2.918429136276245

Logistic Regression Testing Time: 0.1850569248199463

* **Support Vector Machine (SVM):**SVM is a powerful supervised learning algorithm capable of performing classification, regression, and outlier detection tasks.It finds the optimal hyperplane that best separates classes in high-dimensional space.SVM is effective in high-dimensional spaces and when the number of features exceeds the number of samples.It can handle both linear and non-linear classification tasks by using different kernel functions

SVM Accuracy: 0.46644023936600354

SVM Training Time: 399.1364076137543

SVM Testing Time: 122.15789699554443

# **5.Plots:**

# 

# 

# 