ML Projects Milestone 2

Online Articles Popularity Prediction

SC_7

فرح صفوت عز الرجال علي 2021170392

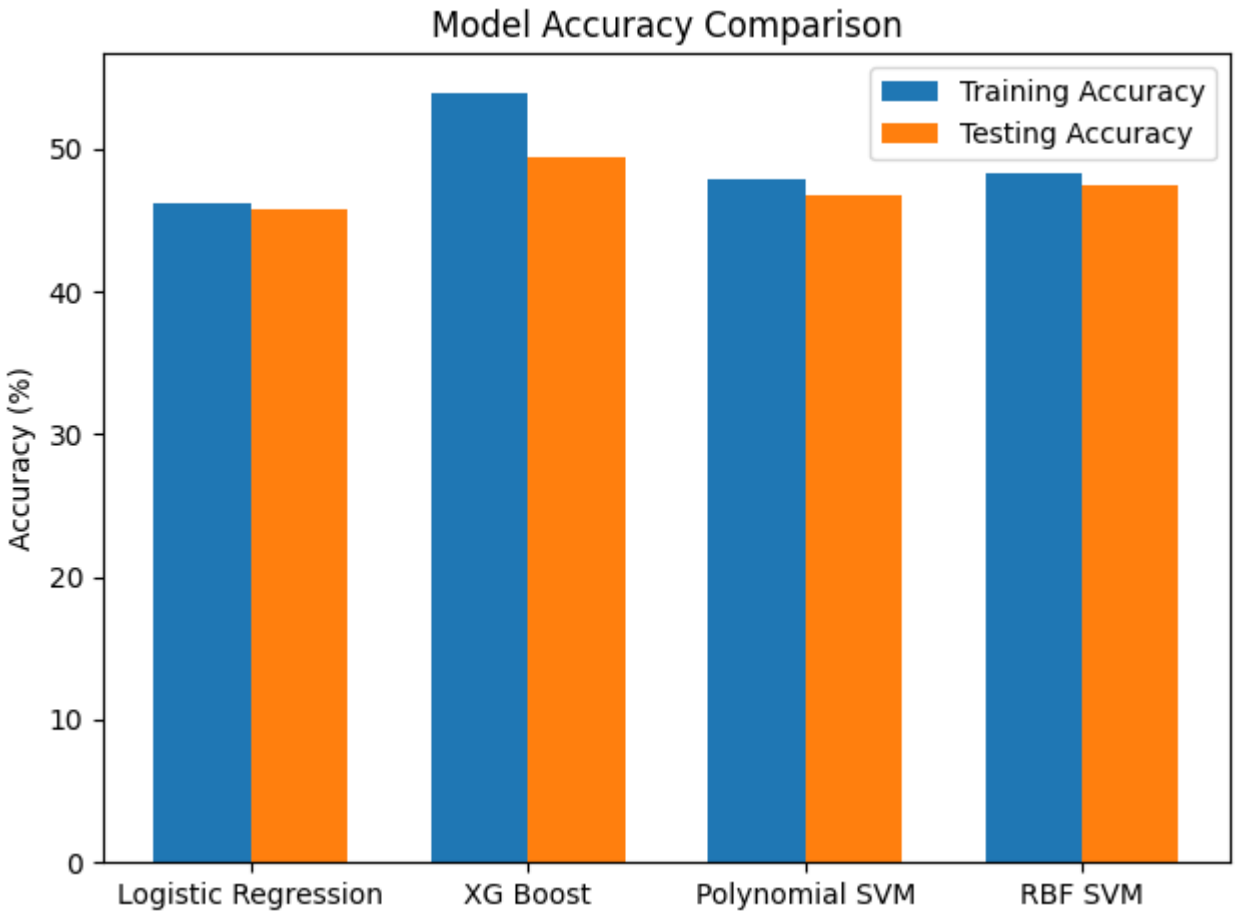
سهير خالد محمد السيد 2021170245

محمد خالد توفيق محمد سعد 2021170463

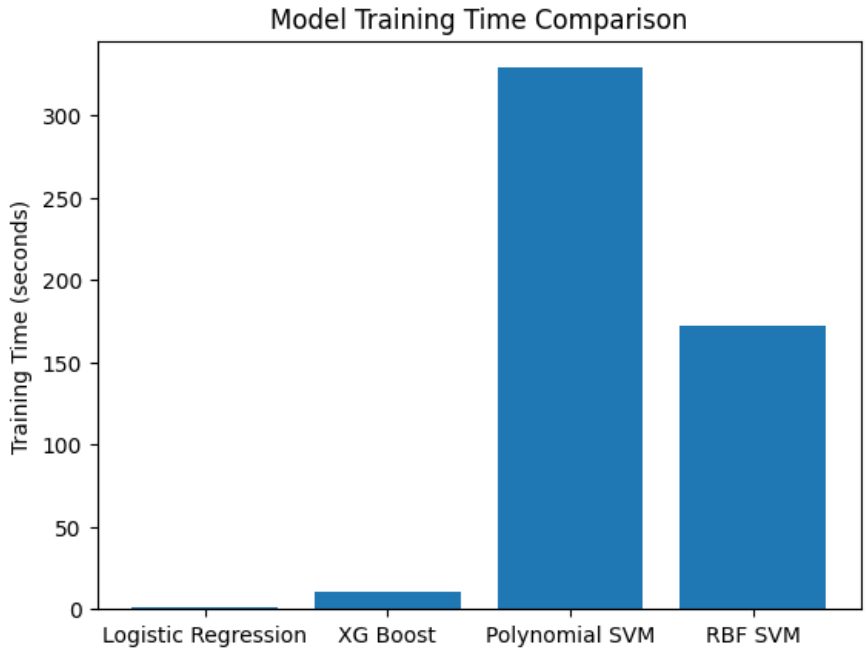
نور سامي محمود ياغي 2021170592

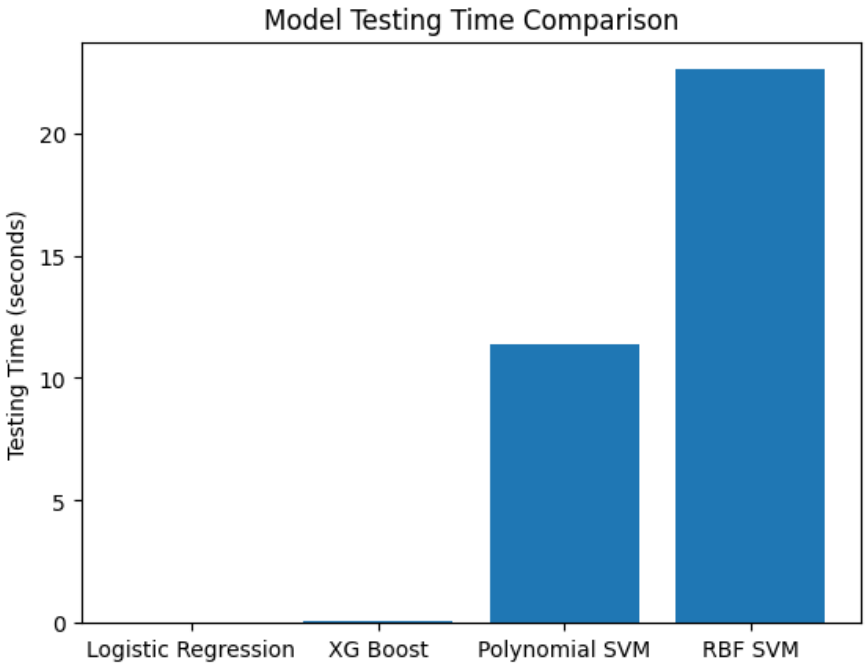
مارك ماجد موريس عزيز 2021170425

مازن محمد عبد الحميد ابراهيم 2021170661



This graph shows the accuracies of 4 different models: XG Boost (which had the highest training and testing accuracies), Logistic Regression, RBF SVM and Polynomial SVM.





These graphs represent the total time needed to train and test each classification model and as we can see the Polynomial SVM has the highest training time yet the RBF SVM has the highest testing time.

• Feature Selection Process

In this phase of the project, the feature selection process involved:

❖ Calculation of ANOVA F-scores and p-values for each feature to select the top 40 features based on their F-scores.

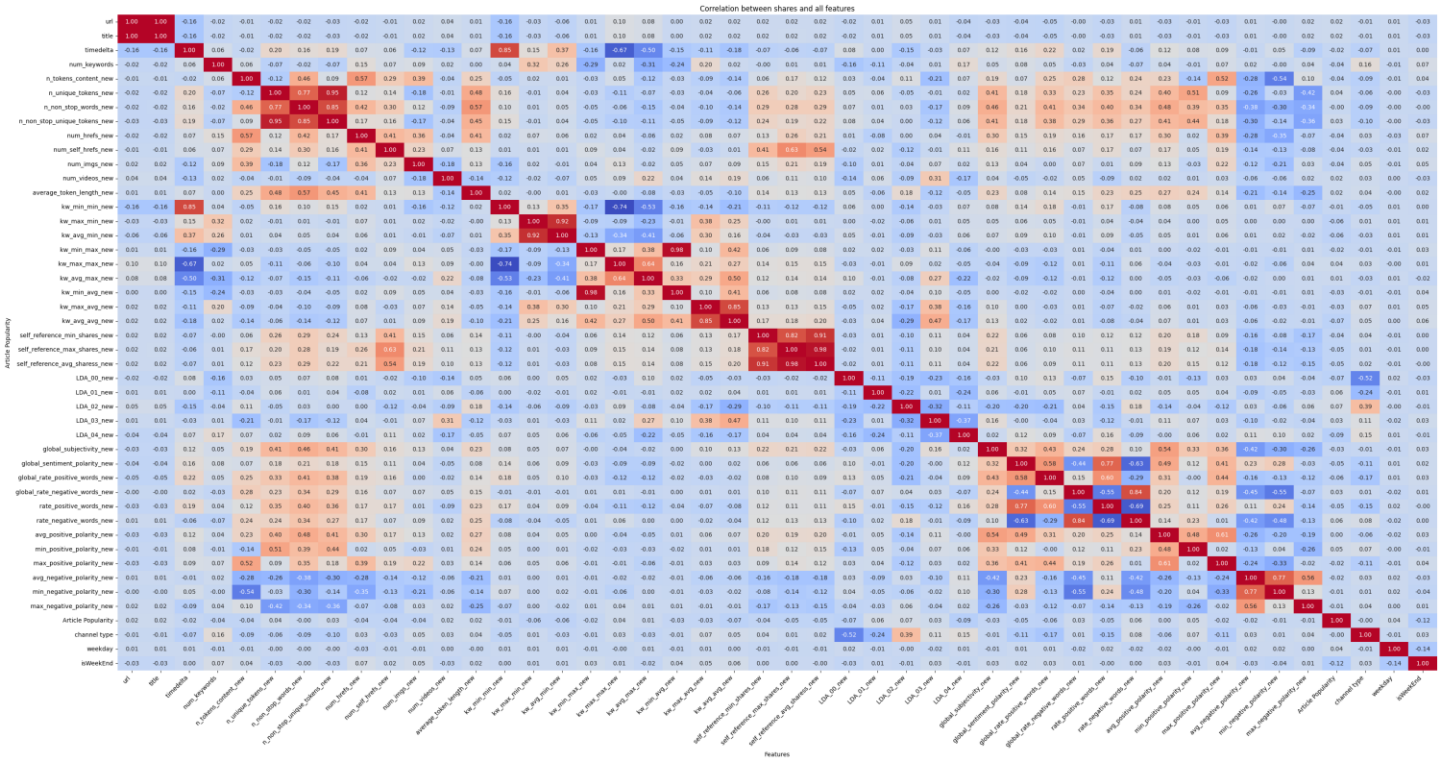
```
f_scores, p_values = f_classif(X, y)

anova_results = pd.DataFrame({'Feature': X.columns, 'F-Score': f_scores, 'p-value': p_values})

anova_results_sorted = anova_results.sort_values(by='F-Score', ascending=False)

top_40_features = anova_results_sorted.head(40)['Feature'].tolist()
print(top_40_features)
```

• Correlation Heat Map:



• Hyperparameter Tuning

The hyperparameter tuning was performed differently for each model:

SVM			
kernel = Poly	C=0.1	C=20	C=100
Training	43.5%	51.4%	57.8%
Testing	41.5%	44.8%	44.1%
Kernel = RPF	C=0.001	C=20	C=1000
Training	43.5%	50.9%	77%
Testing	41.5%	45.3%	41%

Logistic Regression	C=0.001	C=0.1	C=100
Training	43.5%	46.5%	46.6%
Testing	41.4%	44.7%	44.6%

• Conclusion

In this milestone, we changed the target to be categorical and applied different classification models and methods to classify our data according to Article Popularity column, applied hyperparameter tuning on the models and ensembling and in the end we found that XG Boost model has the best accuracy out of all the models used.