Presented by EG3

News Scope: Unveiling Insights Across Domains - An End-to-End Analysis of Diverse News Articles

Meet the team

Mieke Spaans

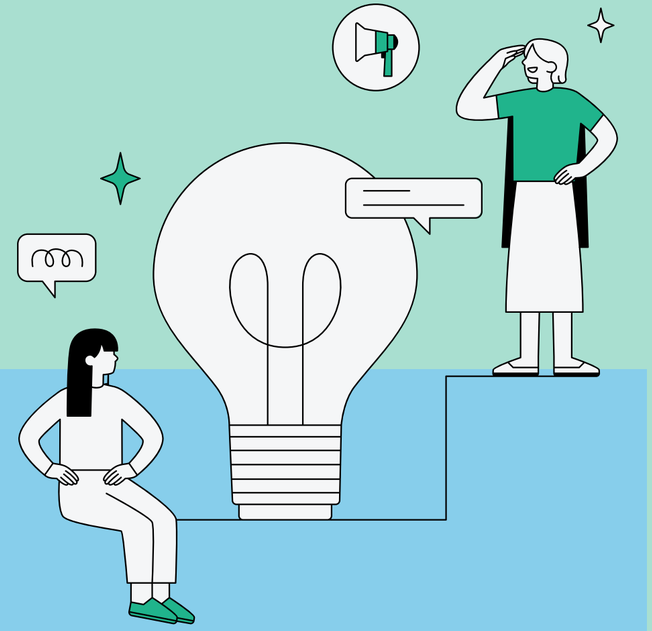
Pfarelo Ramunasi

Zakhele Mabuza

Coceka Keto

Sinawo Londa

Simphiwe Khoza

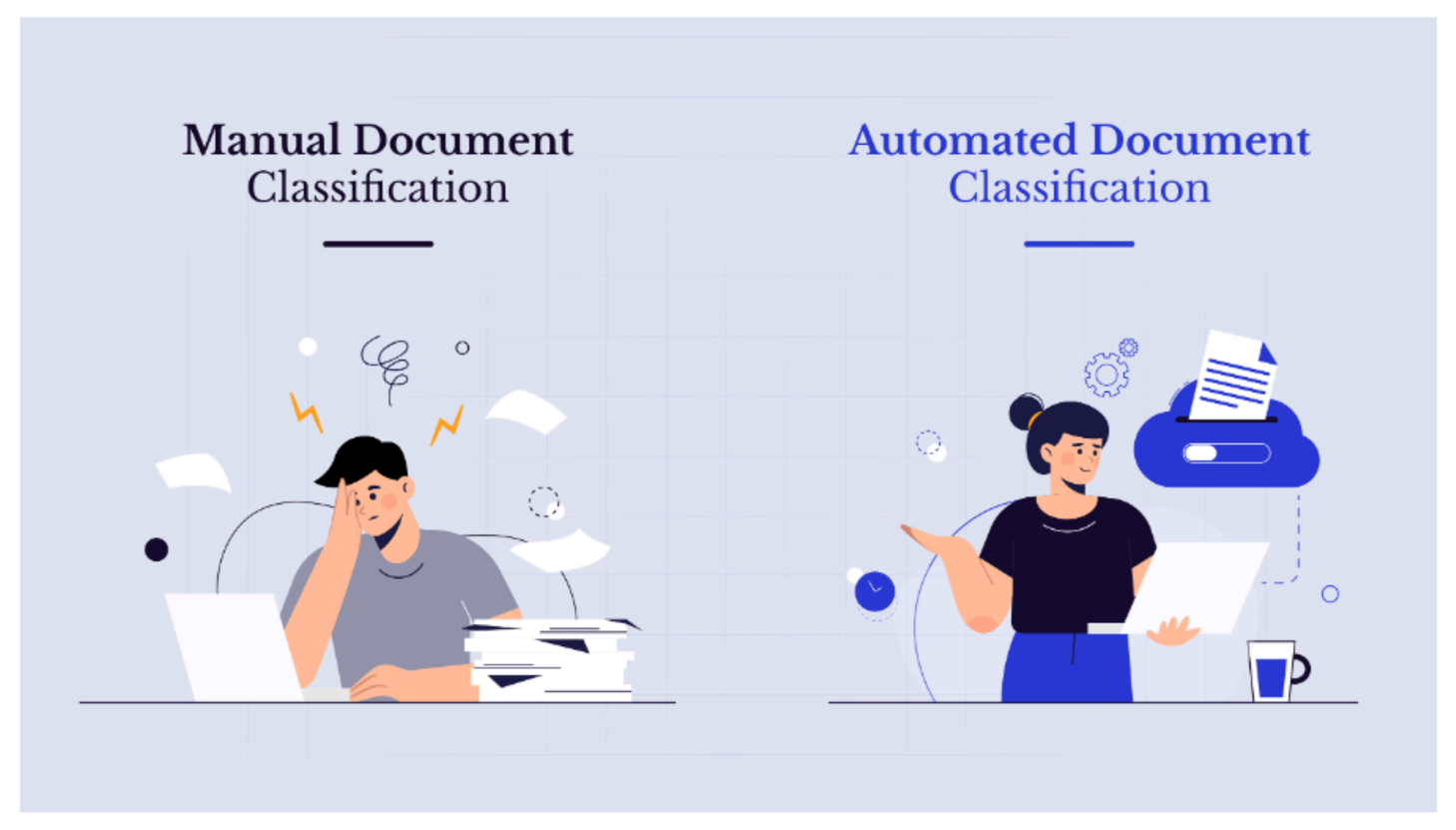


Introduction

- Today's fast-paced digital landscape needs effectively categorizing/delivering
- news content is crucial for enhancing user experience
- Must develop robust classification models using advanced machine learning techniques.
- demonstrate the application of natural language processing (NLP) methodologies through an end-to-end workflow, using Python and Streamlit

Problem statement

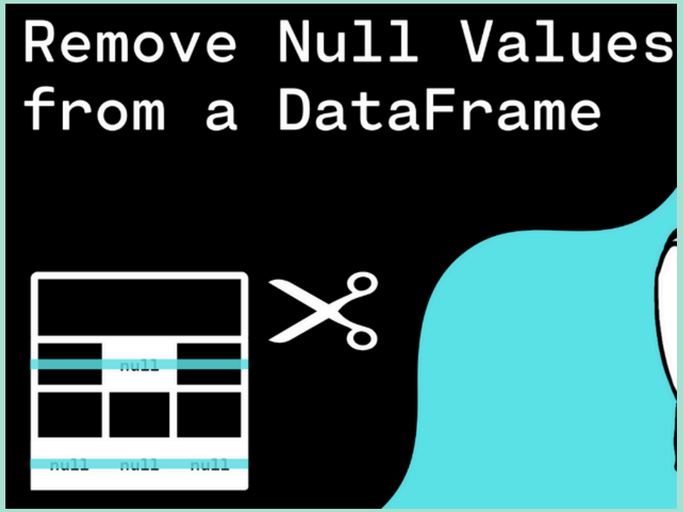
- manually categorizing articles is resource-intensive
- limits the scalability of content management systems.
- need for automated classification mechanisms becomes increasingly urgent.

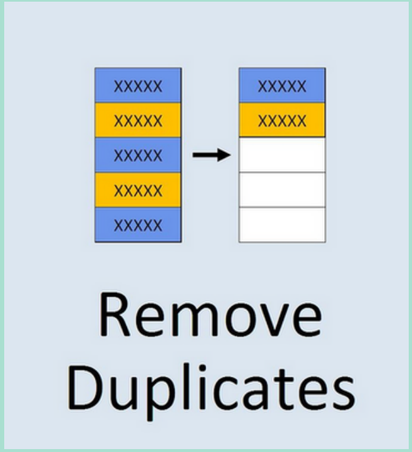


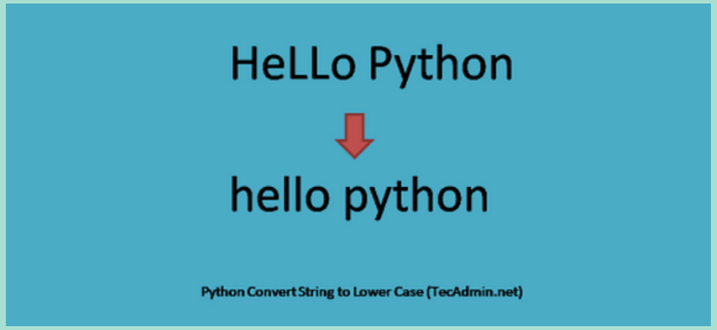
Project walkthrough

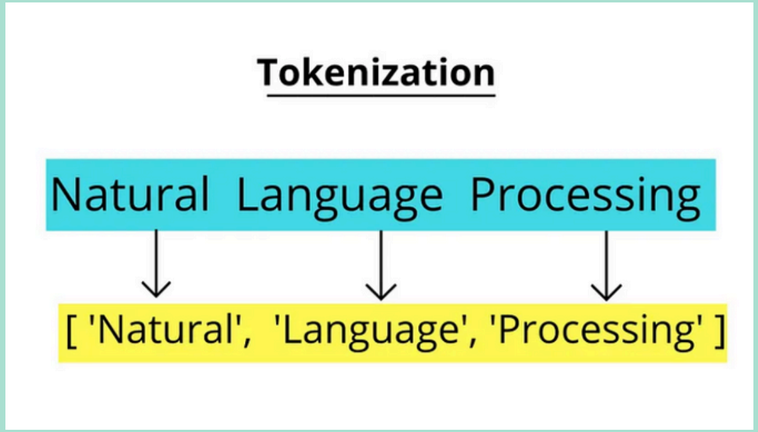
- Data cleaning and pre processing
- Exploratory Data Analysis
- Preparing data for model training
- Model training and evaluation
- Streamlit app demo
- Conclusions and Insights
- Recommendations

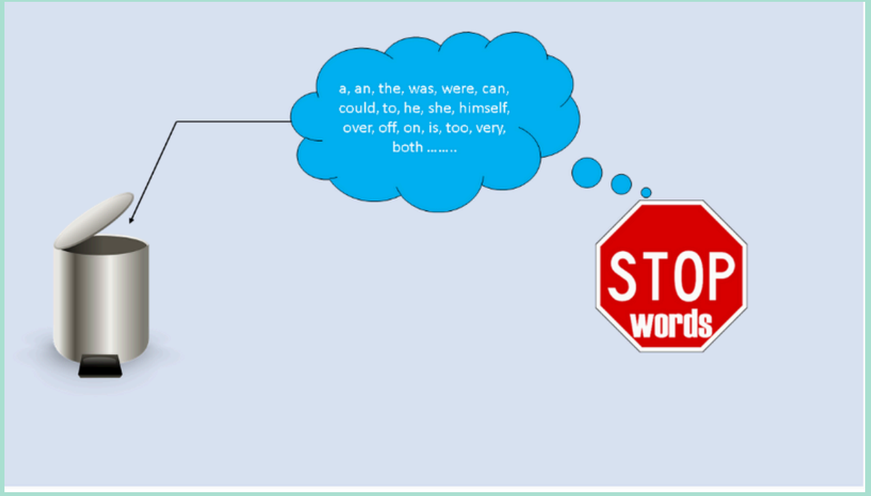
Data Cleaning and pre processing

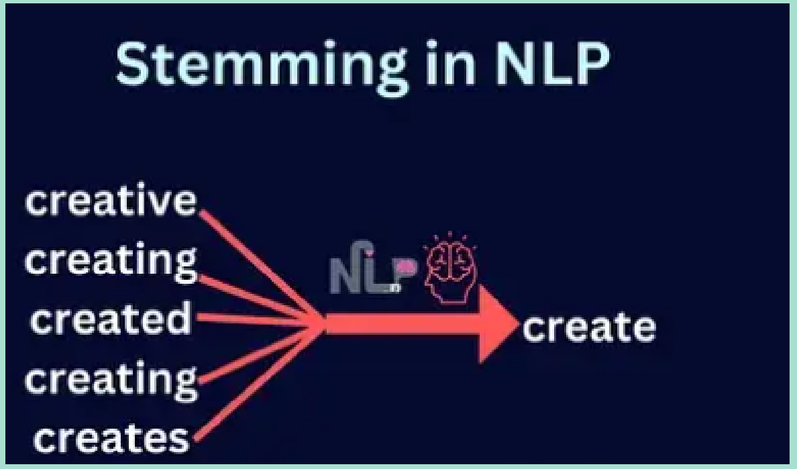




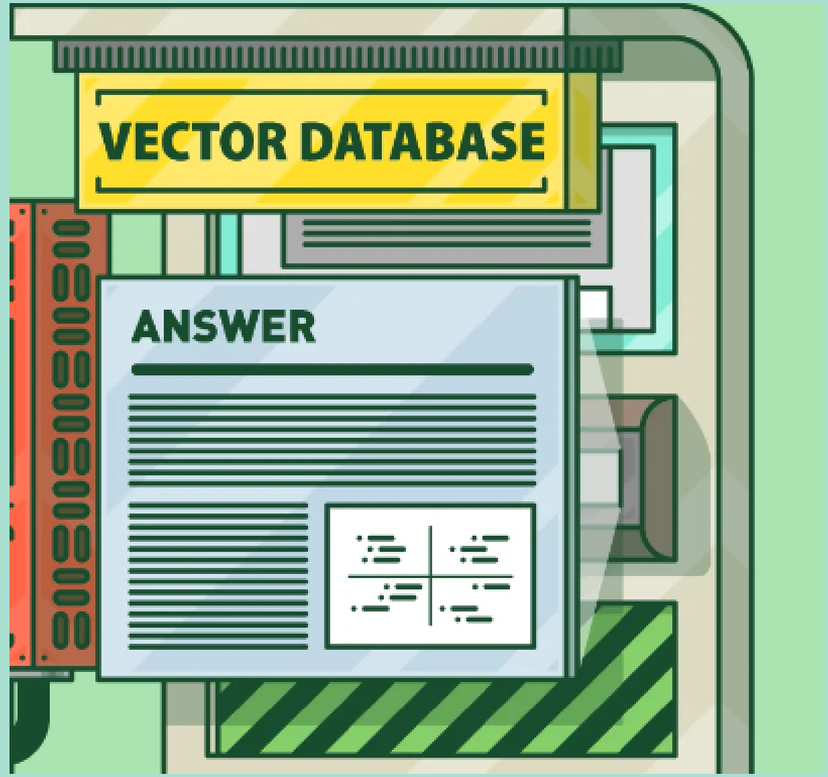




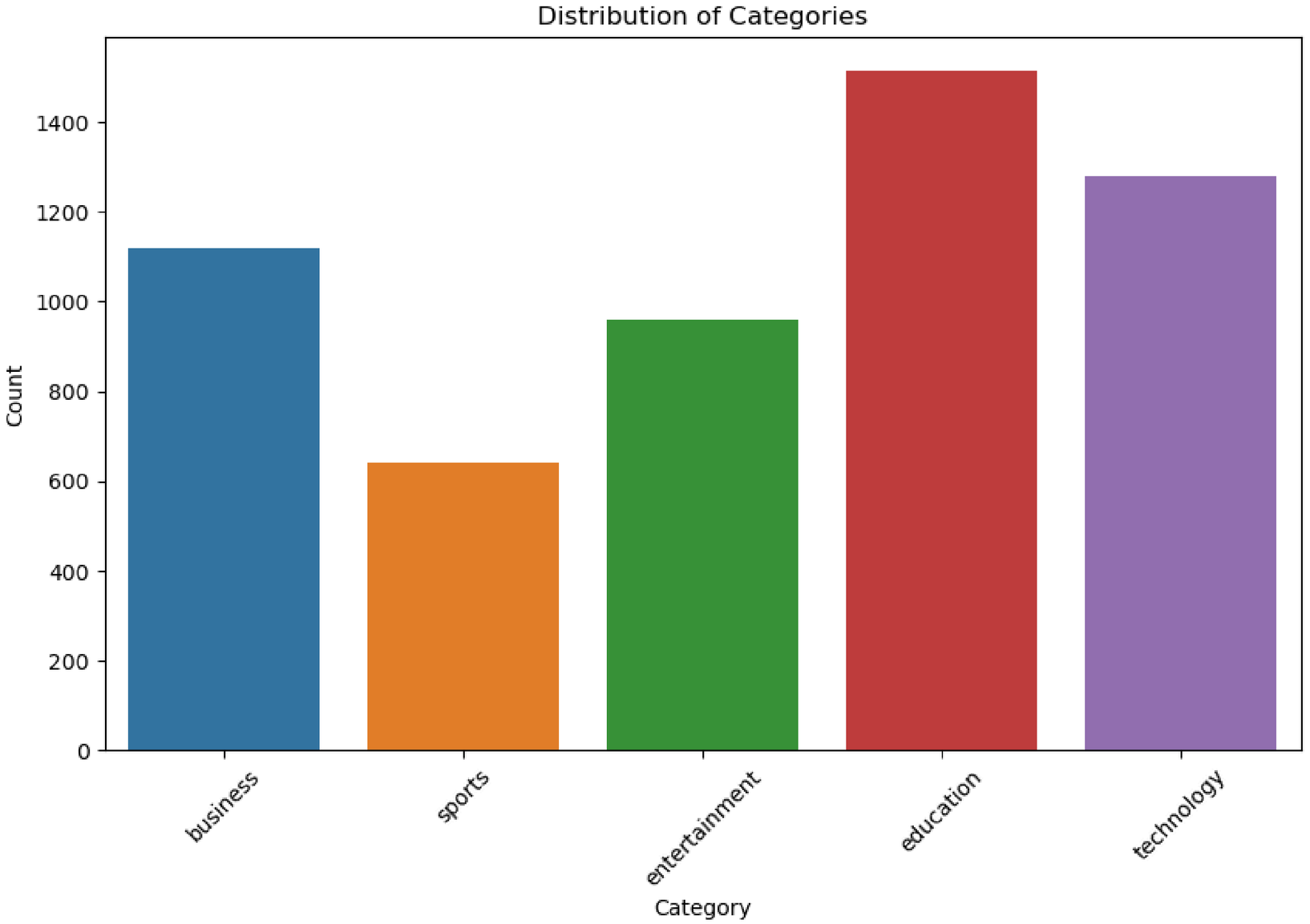




Applying vectorizer and balancing data set

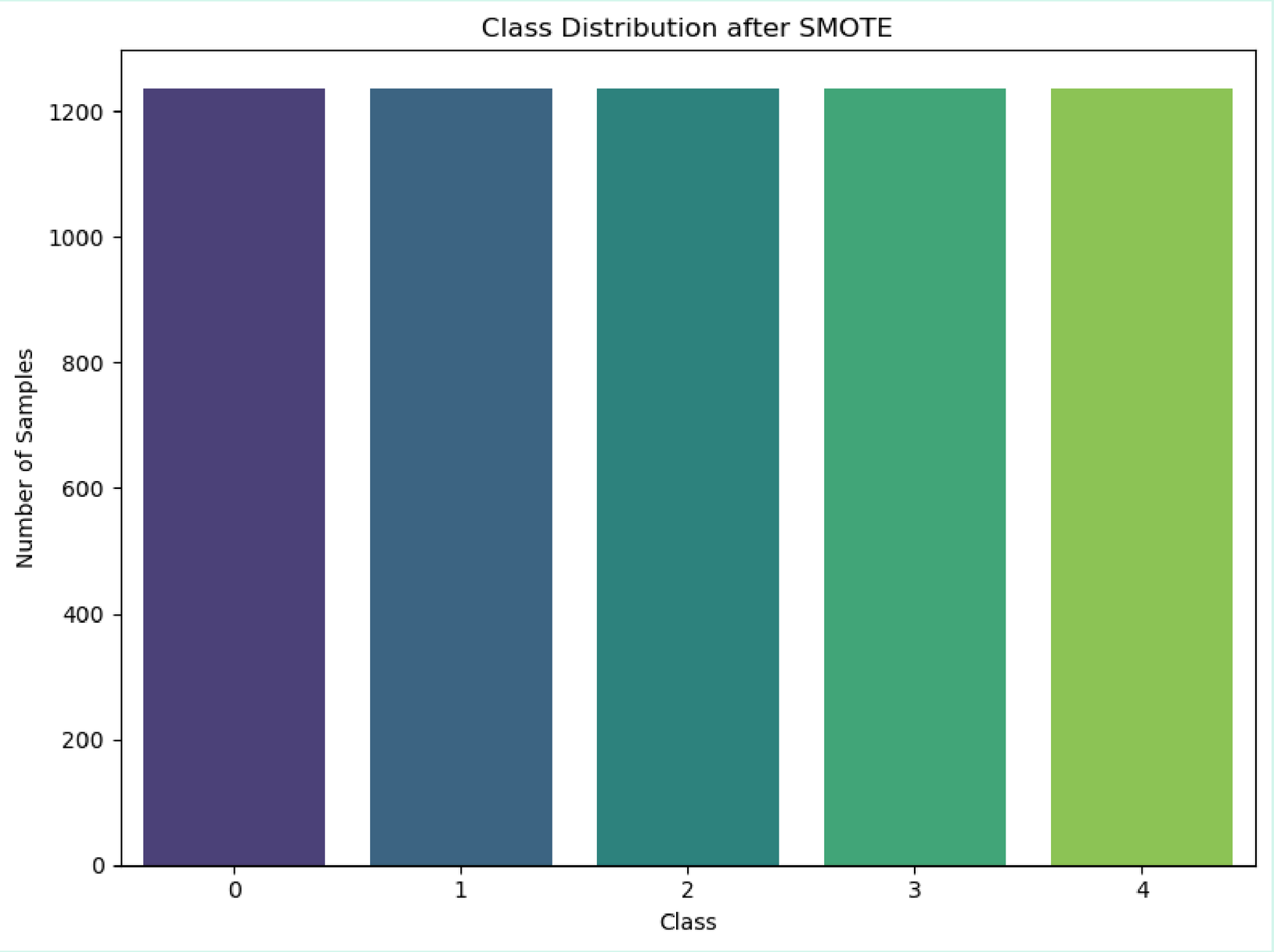


Class distribution



- There is a noticeable class imbalance in the dataset.
- Education category has significantly more instances than the sports category.
- Imbalance can bias them towards the more frequent classes.

Balanced class distribution



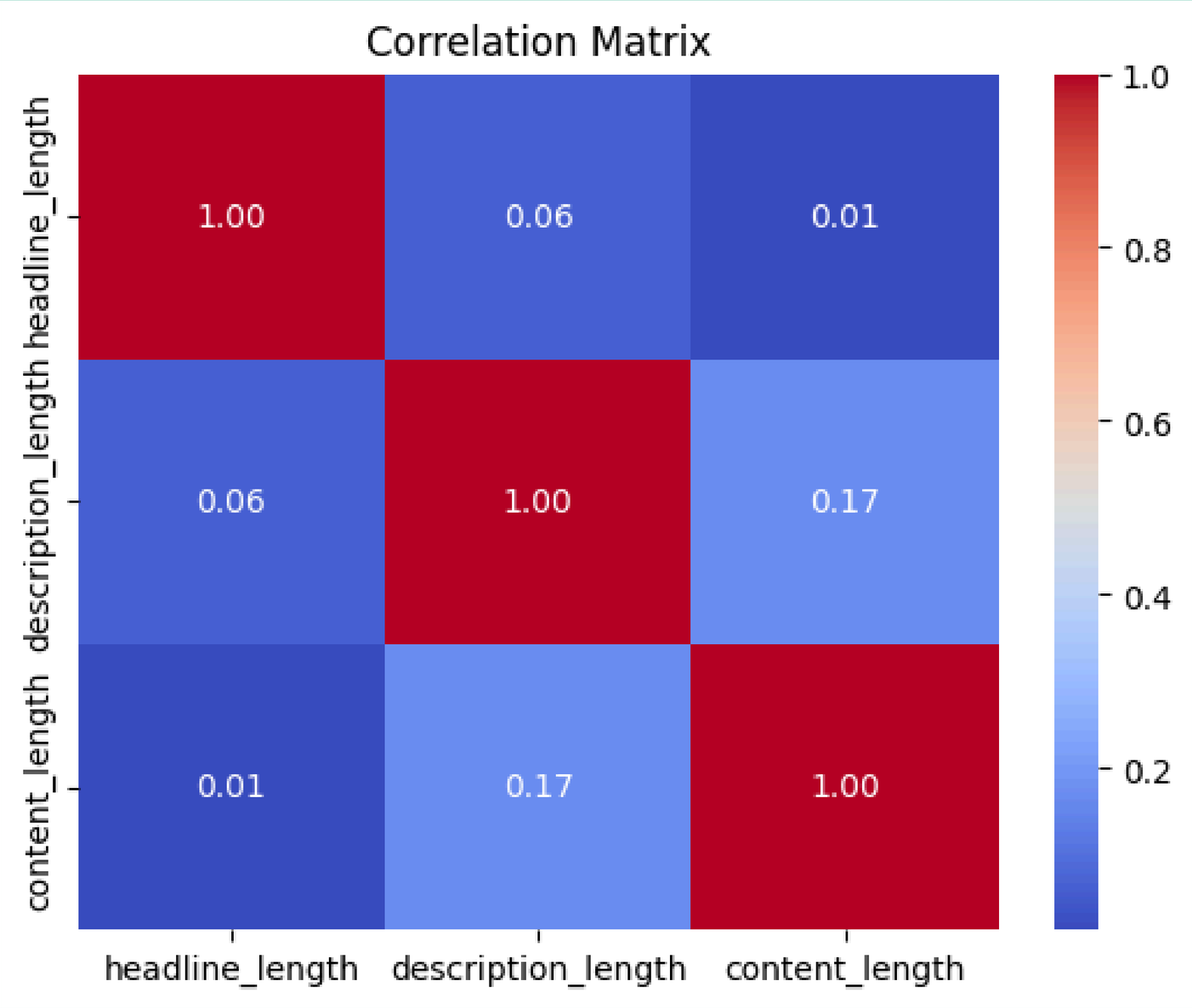
Class distribution is now balanced

Exploratory data analysis

Word Cloud

- mention of tech giant show focus on tech advancements,
- words like said imply quotes from athletes
- Prominence of student suggests focus on educational content
- register may relate to admissions,
- story/advertisement related to narratives in movies, TV shows,

Correlation heatmap



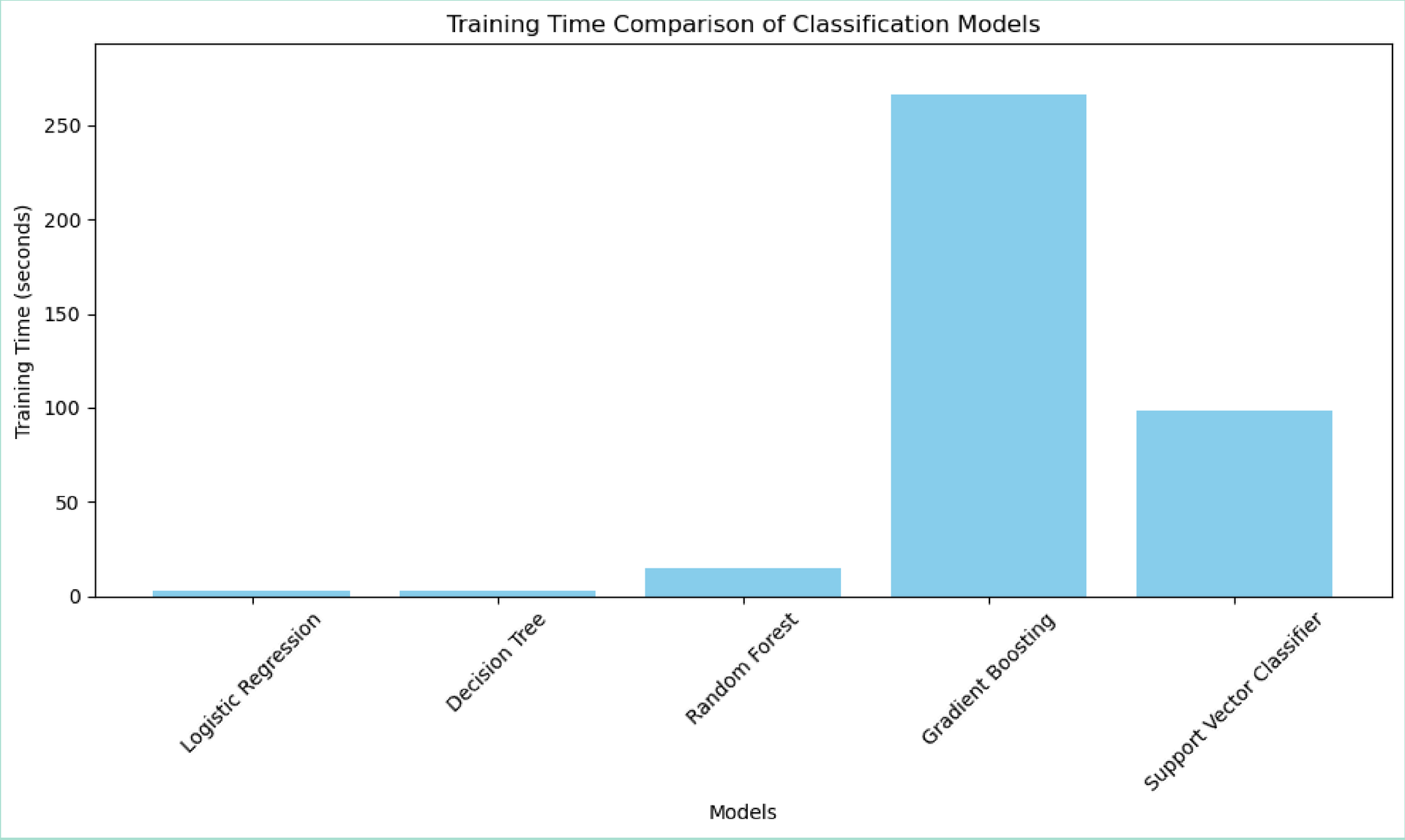
- Headlines follow standard grammatical structures with use of punctuation for clarity
- Focus on words like "India" and "2023" suggests the headlines are relevant to current events,
- Range of headline lengths shows a mix of concise and detailed headlines,

Model training

We trained 5 classifier models:

- Logistic regression
- Decision tree
- Random forest
- Gradient boosting classifier
- Support vector classifier

Models Perfomance



- **Gradient boosting has the highest testing time**
- **Decision, logistic and random forest tree regression have the shortest time**
- **Thus most suitable to use**

Model evaluation

We decided to use the ROC-AUC score to evaluate each of the five model's performance.

Results:

- **Logistic Regression: 0.9985** 
- **Decision Tree Classifier: 0.9180**
- **Random Forest Classifier: 0.9976** 
- **Gradient Boosting Classifier: 0.9972**
- **Support Vector Classifier: 0.9987** 

Top three models

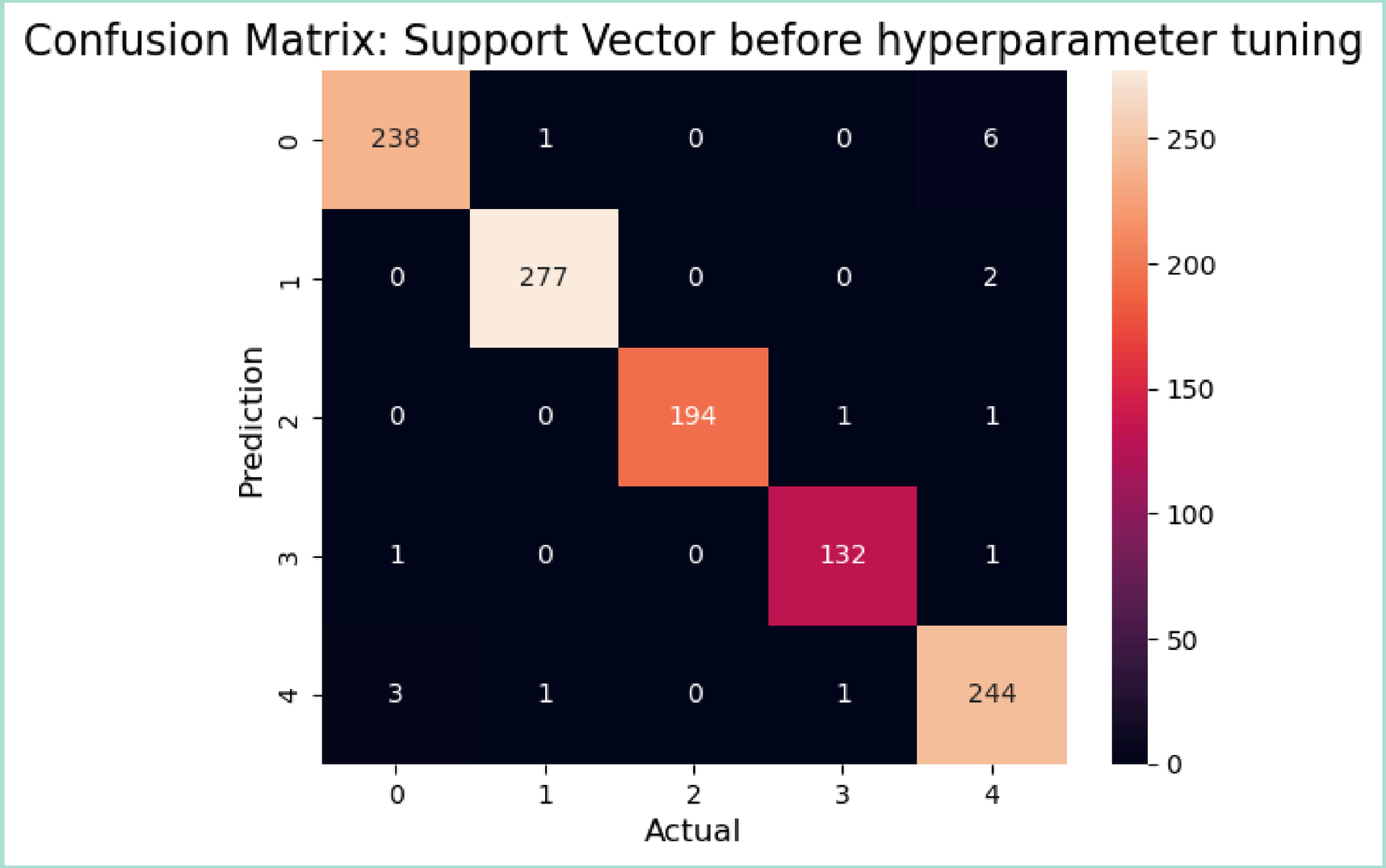
Model hyperparameter tuning

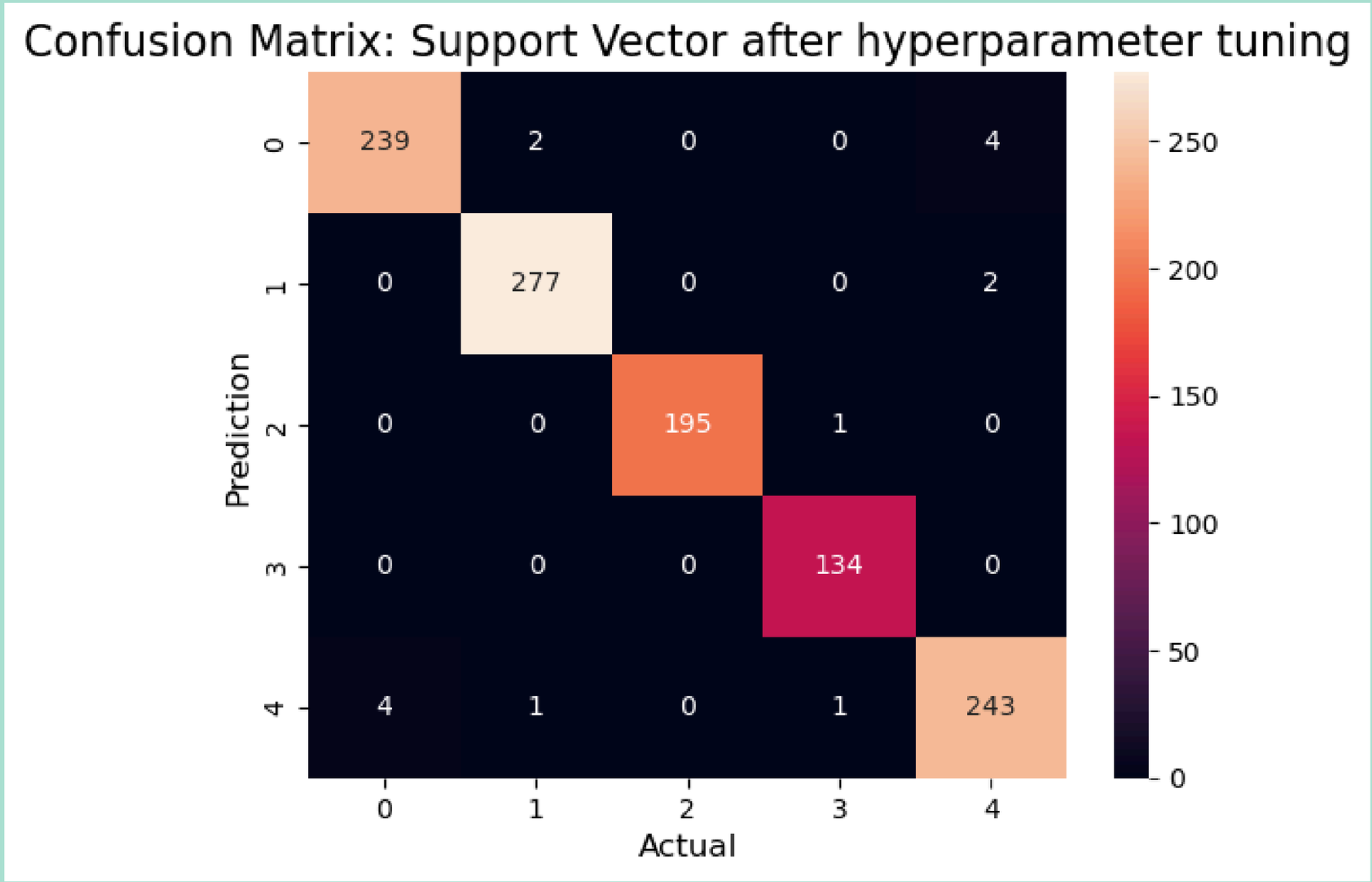
By tuning these three models:

- **Logistic regression**
- **Random forest**
- **Support vector classifier**

Results:

- **SVC displayed the best performance**
- **Increase the roc_auc score to 0.9992**
- **Increase true positives from 1085 to 1088**


Confusion Matrix: Support Vector before hyperparameter tuning
Prediction
Actual
0
1
2
3
4
238
1
0
0
6
0
277
0
0
2
0
0
194
1
1
1
0
0
132
1
3
1
0
1
244
250
200
150
100
50
0


Confusion Matrix: Support Vector after hyperparameter tuning
Prediction
Actual
0
1
2
3
4
239
2
0
0
4
0
277
0
0
2
0
0
195
1
0
0
0
0
134
0
4
1
0
1
243
250
200
150
100
50
0

Streamlit App demo

Conclusions and Insights

Reccomendations

Thank you for your time. Any questions?

Any questions?