Machine Learning Projects
FAKE NEWS
DETECTION

INTRODUCTION

- With the rise of digital media, fake news has become a significant issue, leading to misinformation and manipulation. Identifying and addressing fake news is critical for maintaining trust in information sources. This project focuses on predicting whether a news article is fake or genuine, based on features such as text content, headlines, and other metadata.
- By leveraging machine learning and historical data, the model classifies news articles as "Fake" or "Real." This initiative aims to help readers, platforms, and fact-checkers navigate the media landscape more effectively, promoting informed decision-making and reducing misinformation.

Dataset Overview

The dataset contains 30,938 rows and 4 columns with the following attributes:

- **title**: The headline of the article.
- **author**: The author of the article.
- **text**: The full content of the news article.
- **label**: The target variable (1 = fake news, 0 = real news).

Data Cleaning

1. Missing values were filled:
   title: Replaced with "No Title Available."
   author: Replaced with "Unknown Author."
   text: Replaced with "No Text Available."
2. Duplicate rows were identified (110) and removed, resulting in a final dataset of 30,828 rows

Preprocessing Steps

1. Text cleaning: Removed punctuation and special characters.
2. Tokenization: Split text into individual tokens.
3. Stopword removal: Excluded non-informative words.
4. Lemmatization: Converted words to their base forms.
5. Combined **title** and **text** into a new feature, **News**, for model training.
6. Transformed text using CountVectorizer, resulting in a feature matrix of shape (30,828, 192,409).

Addressing Class Imbalance

- The dataset's class distribution was relatively balanced:
- **Real News (0)**: 16,084 articles (52%)
- **Fake News (1)**: 14,744 articles (48%)

Model Selection and Performance

- For this project, four classification algorithms were evaluated for fake news detection: Multinomial Naive Bayes, Support Vector Classifier (SVC), RandomForestClassifier, and Gradient Boosting. Among these, **RandomForestClassifier** achieved the highest accuracy at **82%**. After applying 5-fold cross-validation, the accuracy of the RandomForestClassifier improved to **83%**, demonstrating its stability and robustness across different data splits. This makes RandomForestClassifier the most effective model for fake news detection, balancing accuracy, precision, and recall while minimizing overfitting.

Conclusion

The Fake News Detection system helps users identify false information online and make more informed decisions. Using machine learning, it provides a reliable tool to spot misinformation, protecting users from misleading content. This system shows how data science can solve real-world problems, making it an important tool in today’s digital world. It also promotes media literacy by encouraging users to critically evaluate the information they come across.