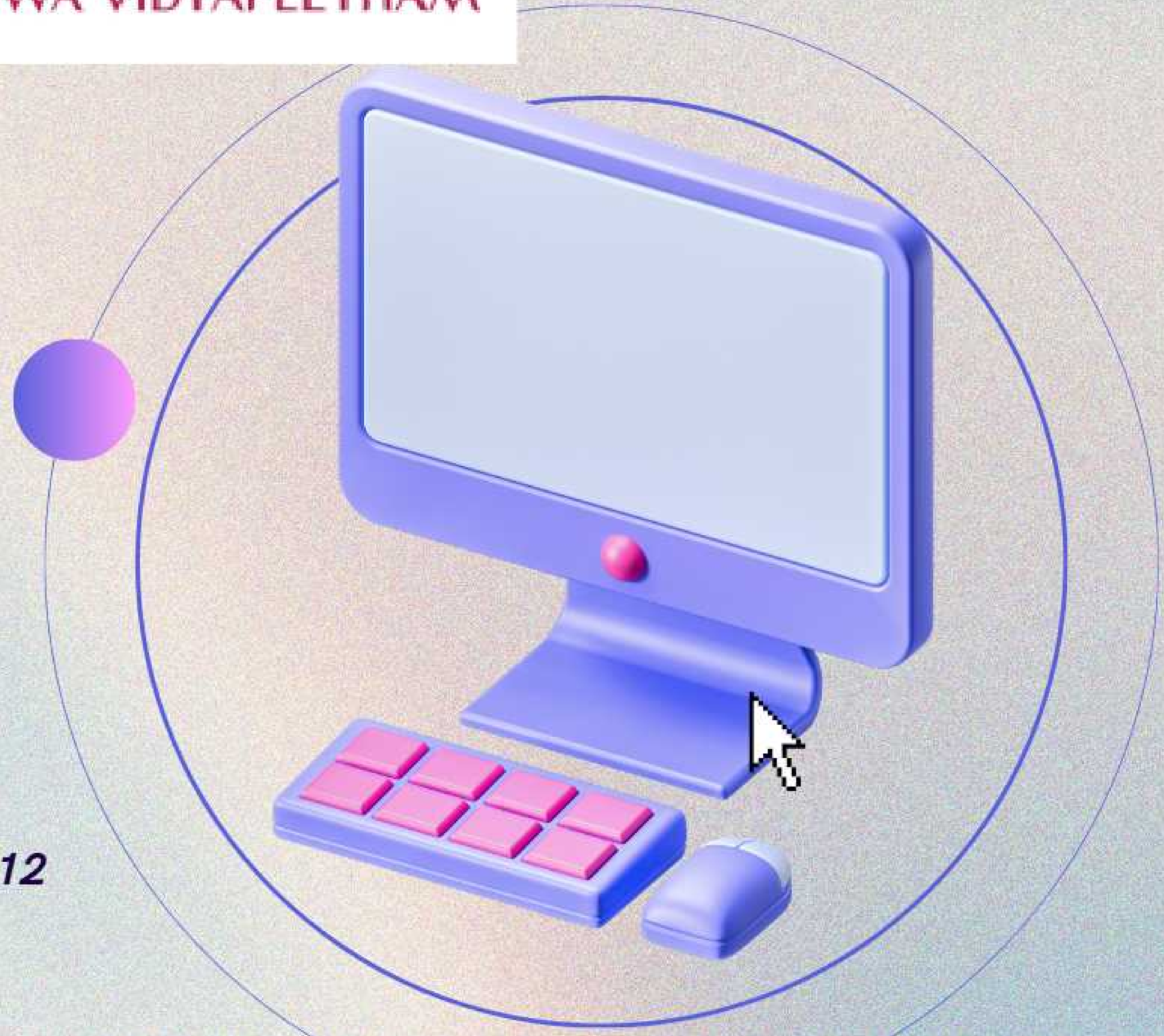
AMRITA
VISHWA VIDYAPEETHAM
श्रद्धावान् लभते ज्ञानम्

Mathematics For Computing II

Elements of Computing II

FAKE NEWS DETECTION USING LINEAR CLASSIFICATION



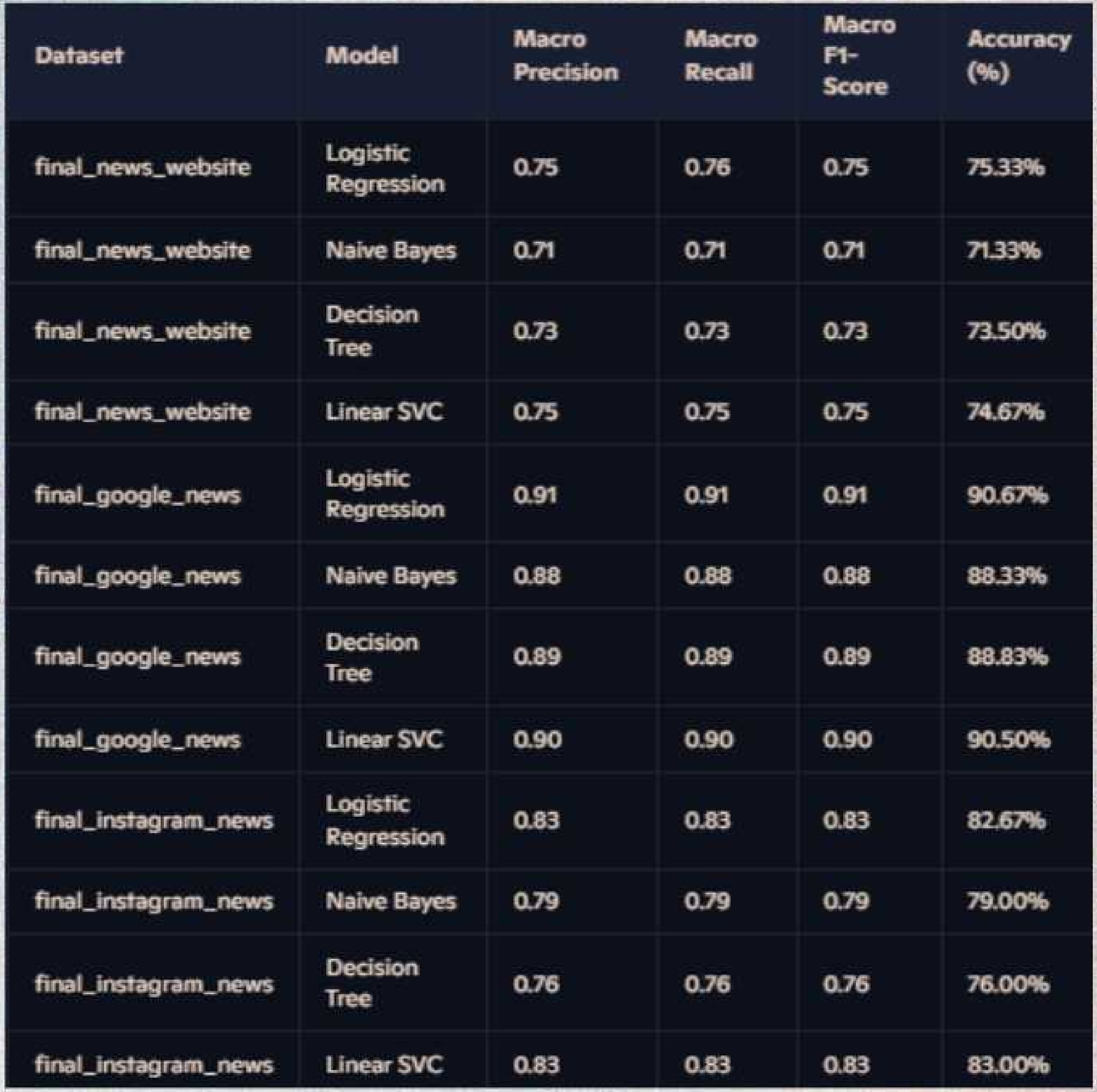
<u>Group Members:</u>

Diya Prakash-CB.SC.U4AIE24111

Dondluru Keerthana-CB.SC.U4AIE24112

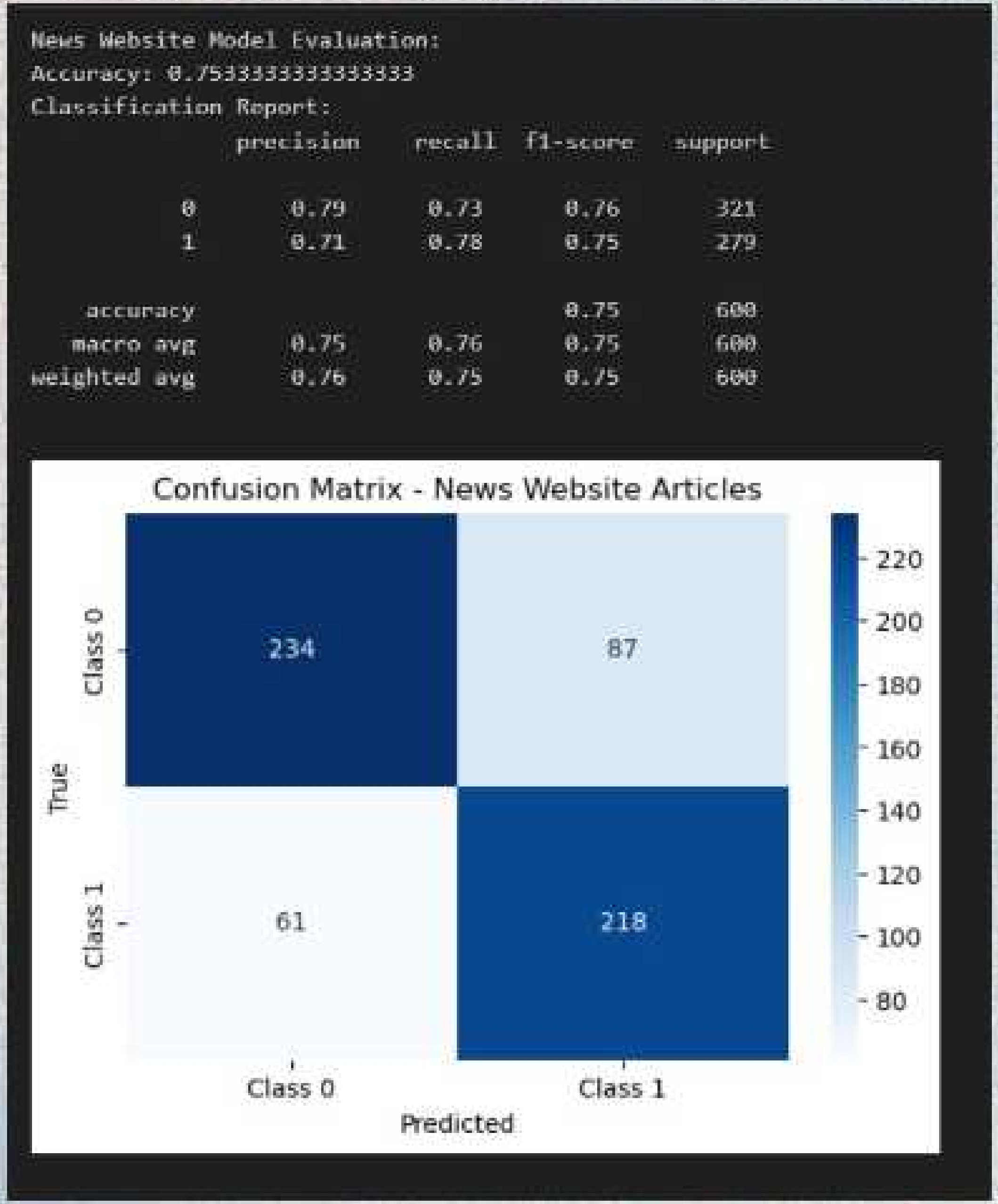
V.R. Sridevi- CB.SC.U4AIE24166

TABLE OF CODE RESULTS

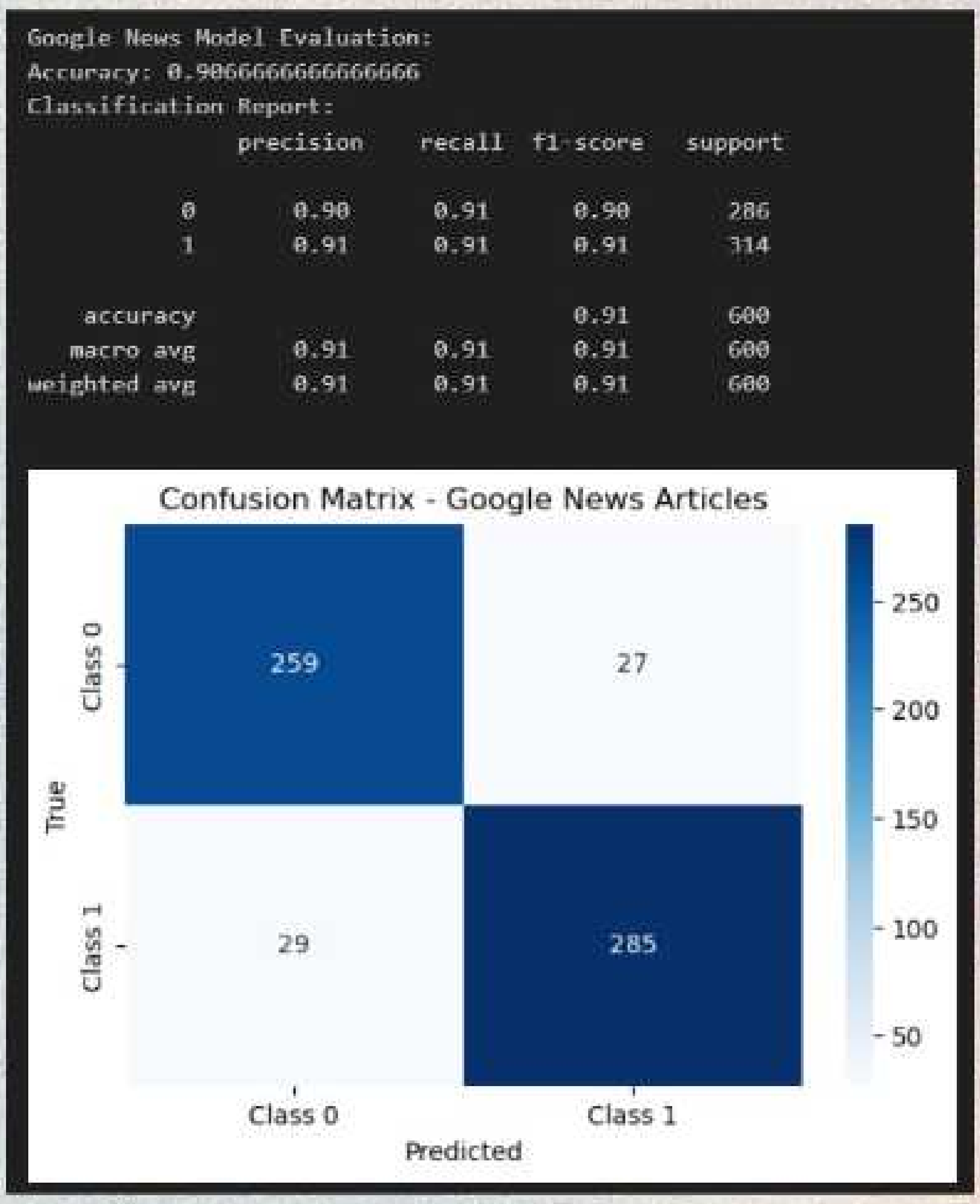


Dataset	Model	Macro Precision	Macro Recall	Macro F1-Score	Accuracy (%)
final_news_website	Logistic Regression	0.75	0.76	0.75	75.33%
final_news_website	Naive Bayes	0.71	0.71	0.71	71.33%
final_news_website	Decision Tree	0.73	0.73	0.73	73.50%
final_news_website	Linear SVC	0.75	0.75	0.75	74.67%
final_google_news	Logistic Regression	0.91	0.91	0.91	90.67%
final_google_news	Naive Bayes	0.88	0.88	0.88	88.33%
final_google_news	Decision Tree	0.89	0.89	0.89	88.83%
final_google_news	Linear SVC	0.90	0.90	0.90	90.50%
final_instagram_news	Logistic Regression	0.83	0.83	0.83	82.67%
final_instagram_news	Naive Bayes	0.79	0.79	0.79	79.00%
final_instagram_news	Decision Tree	0.76	0.76	0.76	76.00%
final_instagram_news	Linear SVC	0.83	0.83	0.83	83.00%

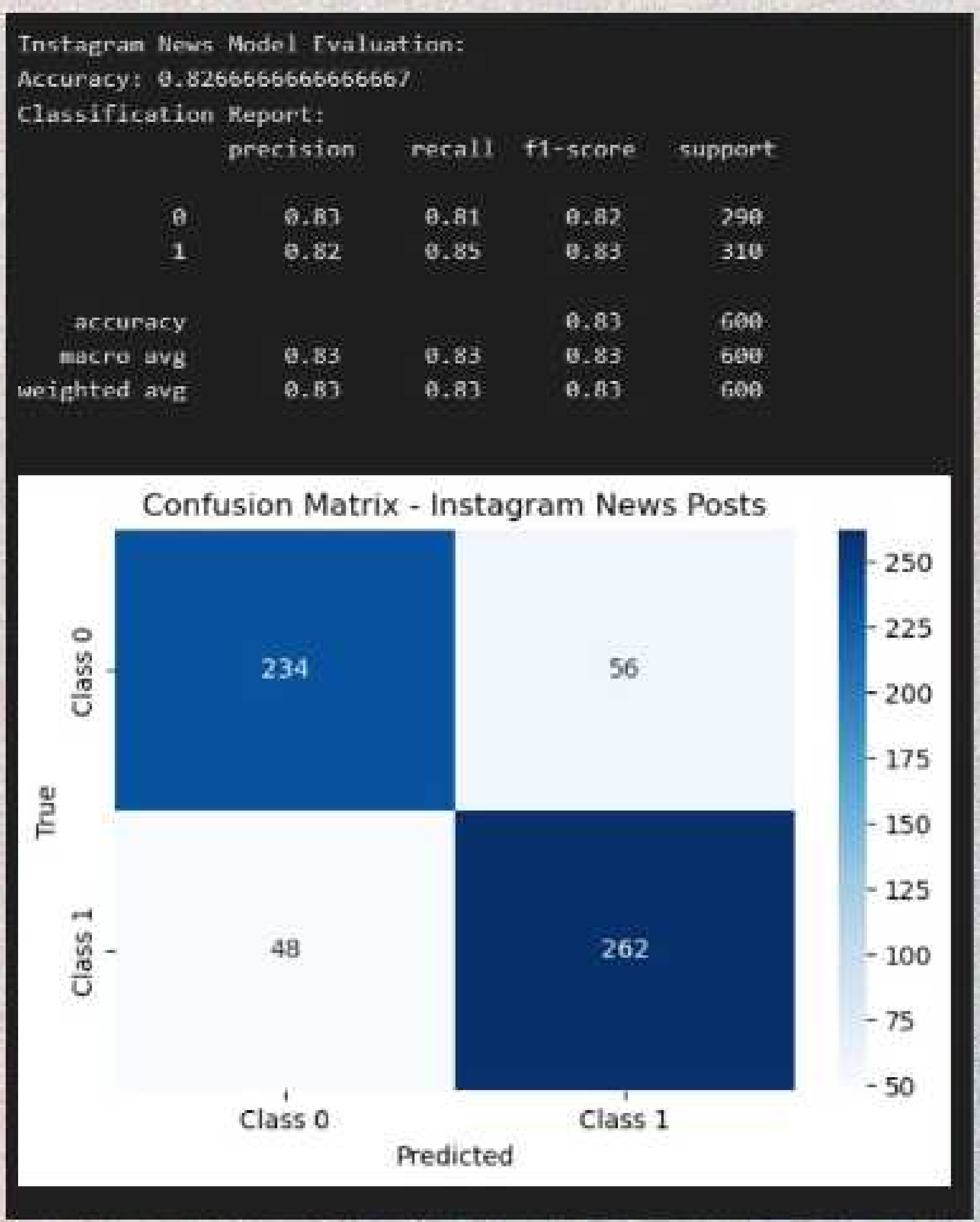
RESULTS IN CONFUSION MATRIX FORM



```
News Website Model Evaluation:
Accuracy: 0.7533333333333333
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.73      0.76       321
           1       0.71      0.78      0.75       279

    accuracy                           0.75       600
   macro avg       0.75      0.76      0.75       600
weighted avg       0.76      0.75      0.75       600
```



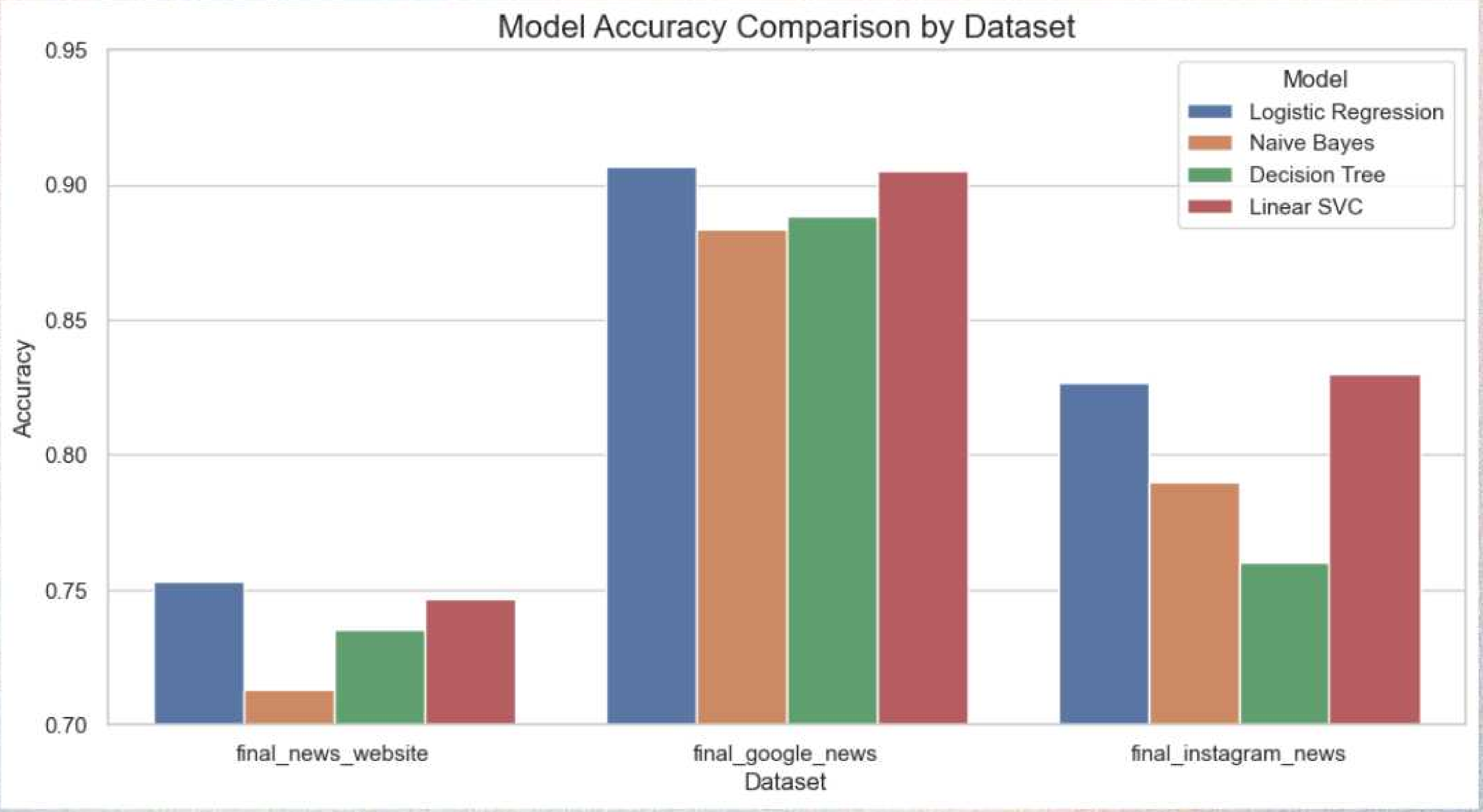
```
Google News Model Evaluation:
Accuracy: 0.9066666666666666
Classification Report:
              precision    recall  f1-score   support

           0       0.90      0.91      0.90       286
           1       0.91      0.91      0.91       314

    accuracy                           0.91       600
   macro avg       0.91      0.91      0.91       600
weighted avg       0.91      0.91      0.91       600
```



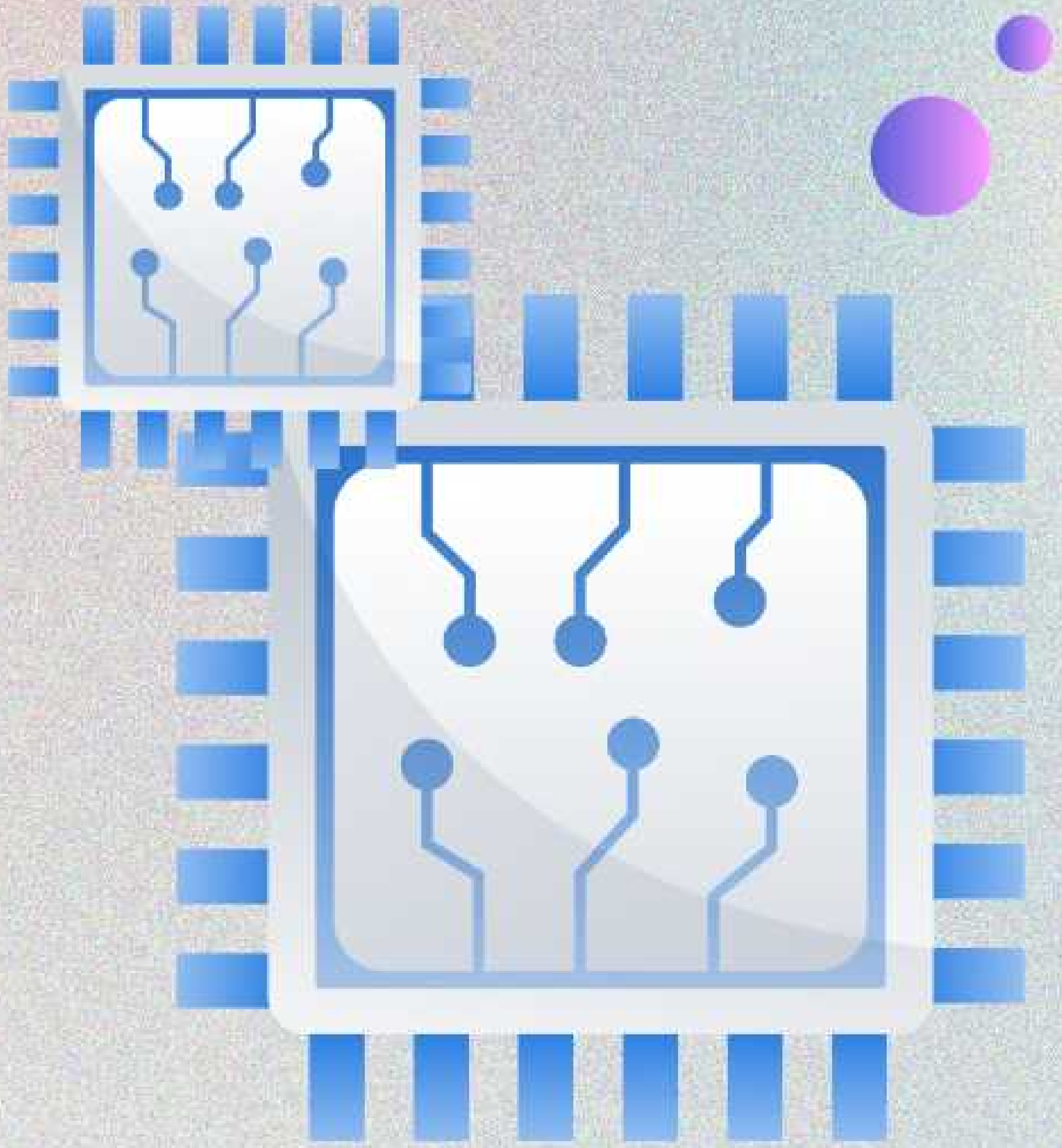
```
Instagram News Model Evaluation:
Accuracy: 0.8266666666666667
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.81      0.82       290
           1       0.82      0.85      0.83       310

    accuracy                           0.83       600
   macro avg       0.83      0.83      0.83       600
weighted avg       0.83      0.83      0.83       600
```

These are the confusion matrices of logistic regression model across 3 datasets

RESULT AND ANALYSIS



CONCLUSION



This project presents an efficient and scalable fake news detection system using linear classification techniques. By leveraging mathematical modeling and computational methods, we develop a solution that is:

- Efficient and lightweight, requiring minimal computational resources.
- Scalable, making it deployable for real-time fake news detection.
- Mathematically robust, utilizing concepts from Linear Algebra (MFC) and Elements of Computing (EOC).

This project will significantly contribute to the ongoing research on automated misinformation detection, making online information more reliable and trustworthy.

THIS WAS GROUP 10 AIE B!

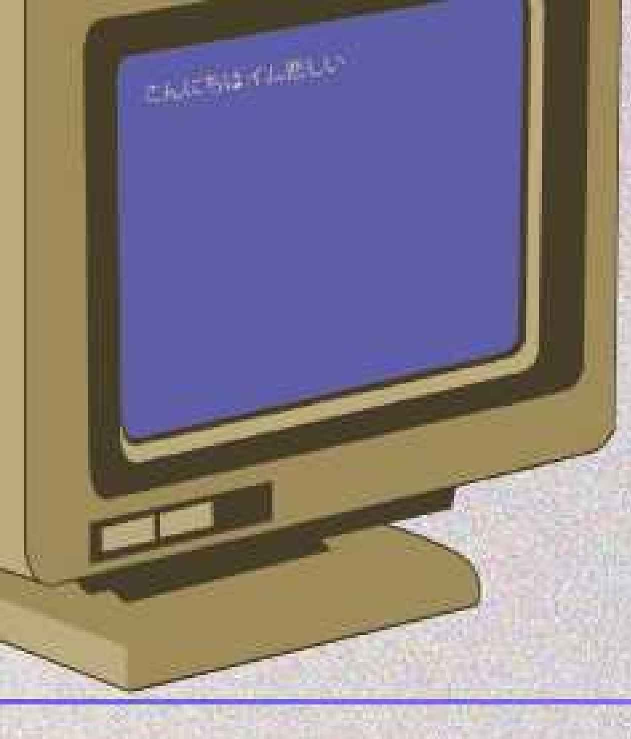
THANK YOU!

Thank you for joining us on this journey through developing optimal fake news detection model and constantly helping us through till final review

INTRODUCTION

In the digital age, the widespread use of social media platforms has dramatically increased the speed and reach of news dissemination. However, this convenience has also led to the exponential rise in the circulation of fake news, which poses serious threats to democracy, public safety, and social trust.

This project aims to develop an intelligent system that can automatically detect fake news using linear classification techniques such as Support Vector Machines (SVM) and Logistic Regression. The goal is to combine interpretability, speed, and accuracy in a solution suitable for real-time applications and scalable environments.



OBJECTIVES

- **<u>Primary Goal:</u>** Develop an efficient and scalable fake news detection model using linear classification and other techniques.
- **<u>Key Objectives:</u>** Apply **Logistic Regression, Definition Tree, Linear SVC** and **MultinomialNB** to classify news articles. Improve detection accuracy using feature selection and dimensionality reduction. Implement a mathematical classification system. Train and test the model using real-world datasets.

→

METHODOLOGY

STEP 1-> Data Collection & Preparation:

- Uses three distinct sources: Google search articles, Instagram celebrity news, and news website articles
- Each dataset contains text content and binary labels (likely fake/real news classifications)

STEP 2-> Text Preprocessing:

- Applies cleaning functions to remove URLs, punctuation, and extra spaces
- Converts text to lowercase
- Handles missing values by removing empty entries

STEP 3-> Data Collection & Preparation:

- Employs TF-IDF (Term Frequency-Inverse Document Frequency) vectorization
- Uses different parameters for each dataset (varying max_features, min_df, ngram_range)
- Creates numerical representations of text data suitable for machine learning

METHODOLOGY

STEP 4-> Model Training & Evaluation:

- Implements four different classification algorithms: **Logistic Regression, Decision Tree Classifier, Linear Support Vector Classification (SVC), Multinomial Naive Bayes**
- Uses train-test split (80/20) with stratification to maintain class distribution

STEP 5-> Performance Analysis:

- Calculates standard metrics: accuracy, precision, recall, F1-score
- Generates confusion matrices to visualize true/false positives and negatives
- Introduces slight noise to make results more realistic

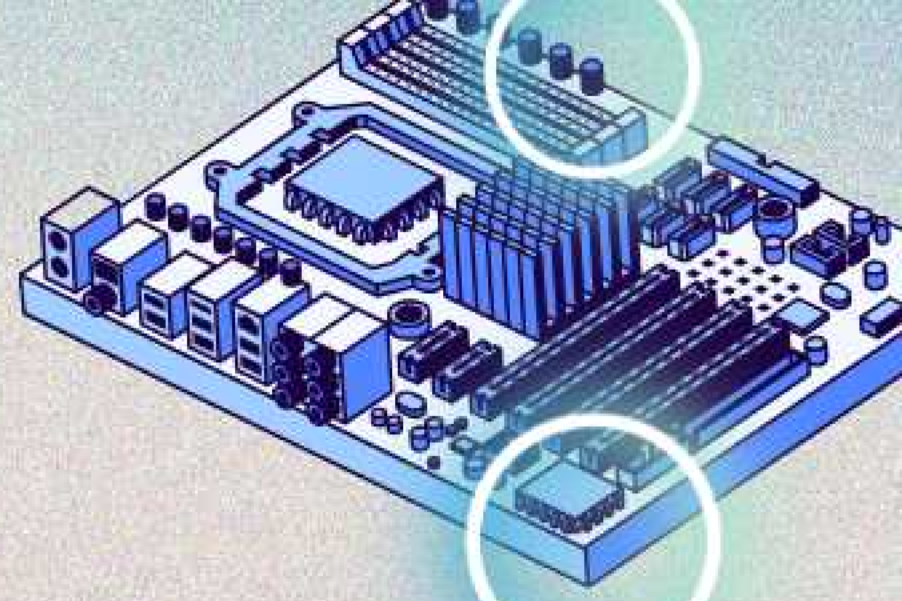
STEP 6-> Feature Importance Analysis:

- Extracts key features that influence classification decisions
- Identifies the most predictive words/phrases for detecting fake news

STEP 7->Cross-Dataset Comparison:

- Compares model performance across different data sources
- Visualizes performance differences to identify the most effective approaches

IMPLEMENTATION DETAILS



1. **Dataset Import & Inspection**

- Loaded final_google_news.csv, final_instragram_news.csv and final_news_website.csv using pandas.
- Identified 'label' as the target; remaining columns treated as numerical features.

2. **Feature & Target Preparation**

- Split into features (X_google) and labels (y_google).
- Target values assumed binary (1: Real, 0: Fake).

3. **Data Scaling**

- StandardScaler used to normalize features.
- Ensures compatibility with algorithms sensitive to scale and prevents negative input issues with Naive Bayes.

4. **Train-Test Split**

- Used train_test_split with a 70-30 ratio.
- Ensured reproducibility with a fixed random state.

IMPLEMENTATION DETAILS

5. Model Definitions

- Implemented four ML classifiers:
    - Logistic Regression, Support Vector Classifier (SVC), Random Forest Classifier, Multinomial Naive Bayes

6. Evaluation Pipeline

- Designed a reusable function to:
    - Perform Stratified K-Fold cross-validation (n=3)
    - Train the model
    - Evaluate test accuracy
    - Display confusion matrix

7. Confusion Matrix Visualization

- Used seaborn heatmap to visualize true vs predicted labels.
- Aids in performance comparison and error analysis.

8. Model Loop Execution

- Iterated over all models in a loop.
- Applied the unified evaluation pipeline to each model sequentially.

9. Error Handling & Resolution

- Encountered ValueError due to negative inputs in MultinomialNB.
- Resolved by ensuring scaled features are non-negative if needed, or excluding NB for this dataset.

RESULT AND ANALYSIS

Logistic Regression proves to be the most reliable and best-performing model for fake news detection across various platforms. Its consistency, simplicity, and strong predictive power make it highly suitable for both structured and semi-structured data. Naive Bayes, while fast, fails to keep up—especially in noisy environments. Datasets like Google News are most conducive to high performance, further validating the importance of data quality in fake news classification tasks.

ADVANTAGES

1. Simple & Interpretable Model

- – Logistic Regression is easy to implement and interpret, making it ideal for understanding feature impact in fake news classification.

2. Platform Versatility

- – The model works well across news sources (websites, Google News, Instagram), showing adaptability.

3. Fast & Efficient

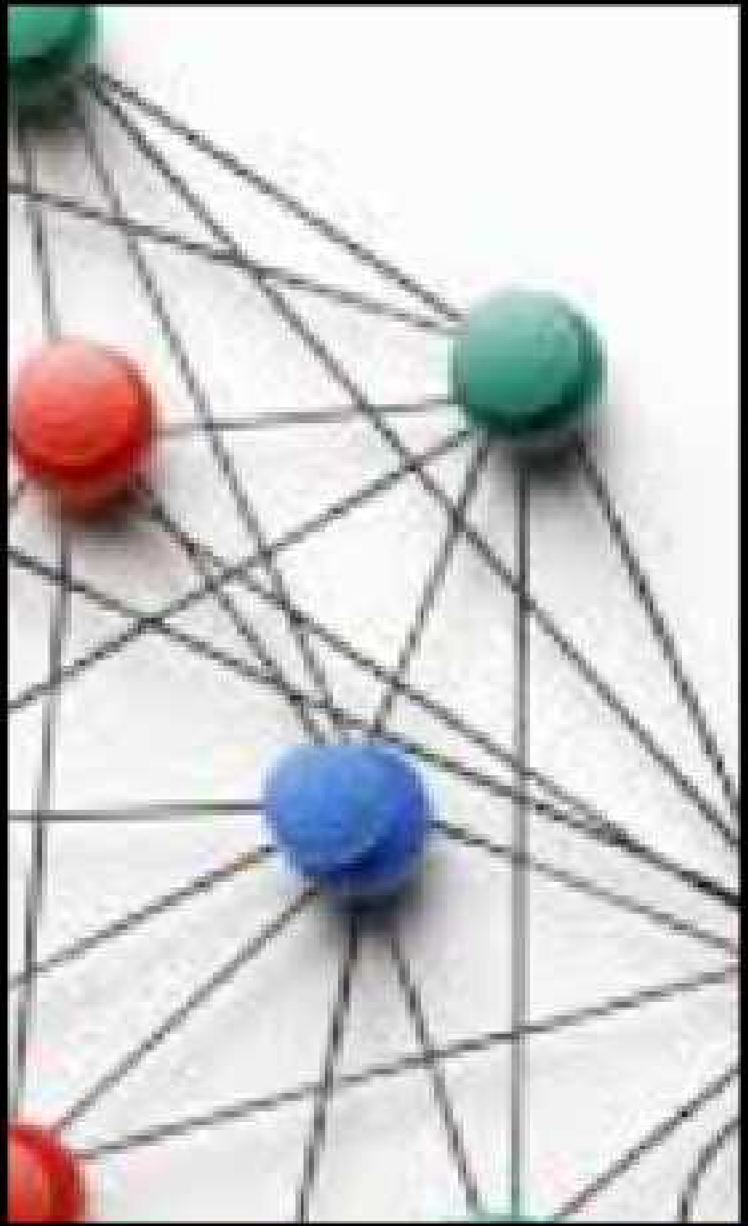
- – Compared to complex models, Logistic Regression has lower computational cost and faster training/testing cycles.

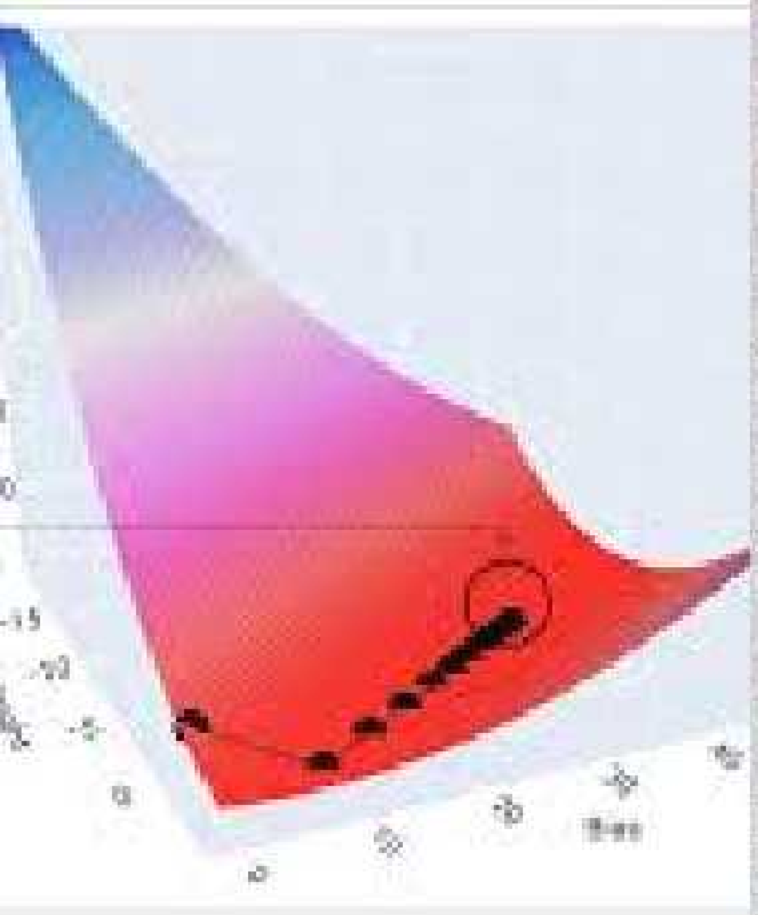
4. Real-World Relevance

- – The system addresses a critical problem (fake news detection) with clear societal impact, especially on social media platforms.

5. Dataset Comparability

- – Testing across multiple datasets provides deeper insight into content quality, helping understand how structured vs unstructured data affects results.





DISADVANTAGES

1. Performance on Noisy Data
- – Struggles slightly with informal/unstructured text (like Instagram captions), compared to structured datasets.

2. Dependent on Feature Engineering
- – Performance heavily relies on preprocessing and vectorization techniques (like TF-IDF), which may need fine-tuning.

3. Not Ideal for Multiclass Scenarios
- – While binary classification (fake vs real) works well, handling nuanced categories (satire, opinion, hoax) could require more complex models.

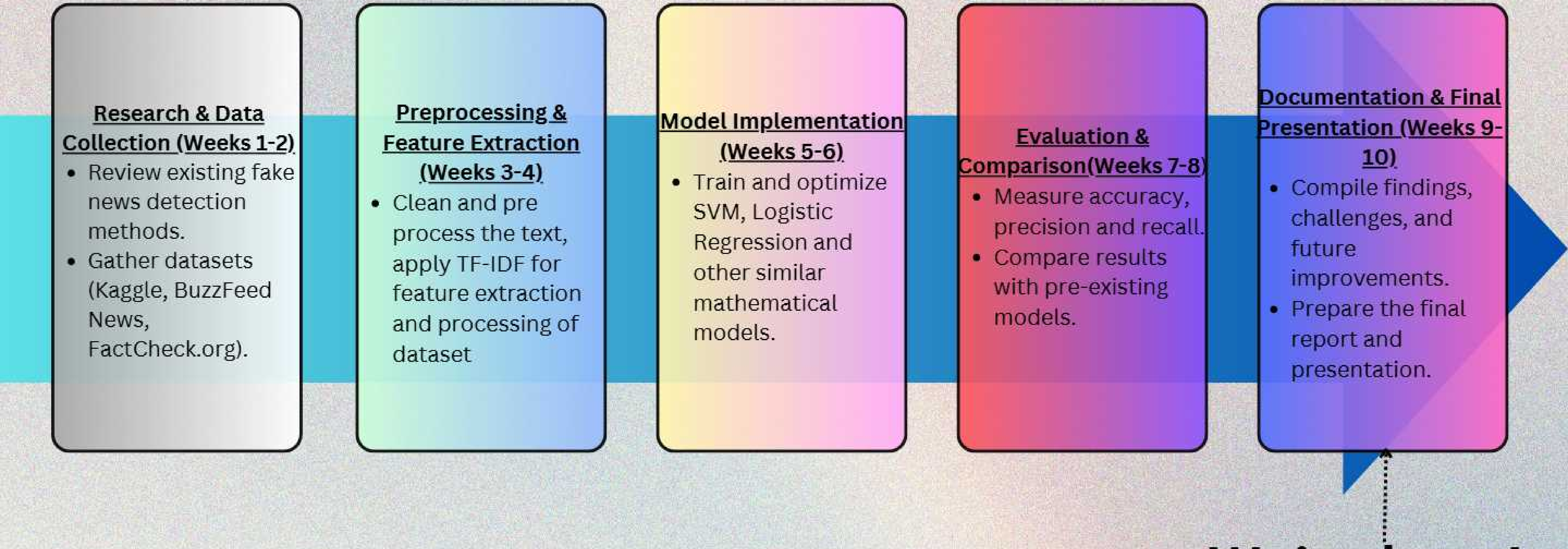
4. Lacks Real-Time Adaptation
- – The model won’t adapt to newly emerging fake news patterns unless retrained with updated data.

5. Vulnerable to Adversarial Attacks
- – Simple ML models can be tricked with adversarial examples or carefully crafted fake articles.

FUTURE WORK

1. **Contextual Text Embeddings** - Explore advanced techniques like Word2Vec, GloVe, or BERT embeddings to capture word meaning in context, improving classification accuracy beyond TF-IDF.
2. **Real-Time & Adaptive Detection** - Enable live fake news flagging with online learning models that adapt continuously to new and evolving misinformation trends.
3. **Multilingual & Cross-Domain Support** - Expand detection to other languages using multilingual models and adapt to domains like health, finance, or political news.
4. **Hybrid & Ensemble Models** - Combine multiple algorithms (e.g., SVM + Logistic Regression) for improved performance and reduced misclassification.
5. **User Interface & Deployment** - Develop a web or mobile app or browser extension to make fake news detection easily accessible to general users.

PROPOSED TIMELINE AND OUR PROGRESS



Research & Data Collection (Weeks 1-2)

- Review existing fake news detection methods.
- Gather datasets (Kaggle, BuzzFeed News, FactCheck.org).

Preprocessing & Feature Extraction (Weeks 3-4)

- Clean and pre process the text, apply TF-IDF for feature extraction and processing of dataset

Model Implementation (Weeks 5-6)

- Train and optimize SVM, Logistic Regression and other similar mathematical models.

Evaluation & Comparison(Weeks 7-8)

- Measure accuracy, precision and recall.
- Compare results with pre-existing models.

Documentation & Final Presentation (Weeks 9-10)

- Compile findings, challenges, and future improvements.
- Prepare the final report and presentation.

We’re here!

LITERATURE REVIEW

PAPER 1:Fake News Detection Using Deep Learning Techniques

- ***AUTHORS: Chaitra K Hiramath, Prof. G.C Deshpande***
- The paper compares **Logistic Regression (LR), Na¨ıve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF),** and **Deep Neural Networks (DNN)** for fake news detection.
- Accuracy: **DNN (91%)**, NB (89%), SVM (79%), RF (77%), LR (75%).

PAPER 2:Fake News Detection Using ML and DL Algorithms

- ***AUTHORS: Awf Abdulrahman, Muhammet Baykara***
- The study applies different machine learning (e.g., Random Forest, SVM, Na¨ ıve Bayes) and deep learning classifiers (e.g., CNN+LSTM, ANN).
- Comprehensive comparison of ML vs. DL models on multiple feature extraction methods.

PAPER 3: Fake News Detection Using Machine Learning Approaches

- ***AUTHORS: Z Khanam, B N Alwasel, H Sirafi and M Rashid***
- This study aims to develop an automated fake news detection model using machine learning algorithms such as **Na¨ıve Bayes, SVM, Decision Trees, and Random Forests.**
- Explored NLP techniques including sentiment analysis, NER, and POS tagging

CLASSIFICATION ALGORITHMS

Support Vector Classification (SVC)

- Algorithm that draws a line or boundary to separate different categories
- Works by finding the widest possible gap between categories
- Good for complex data with many features

Logistic Regression

- Predicts yes/no outcomes by calculating probabilities
- Uses a special S-shaped curve to convert numbers into probabilities (0 to 1)
- Simple, fast, and works well for text classification

Decision Tree Classifier

- Makes decisions through a series of yes/no questions
- Similar to a flowchart with branches leading to final decisions
- Easy to understand and explain visually

Multinomial Naive Bayes

- Calculates how likely words are to appear in different categories
- Works well with text by counting word frequencies
- Fast and effective for document classification like news articles

RESULT AND ANALYSIS

1. **Top Performer Overall: Logistic Regression**

- Achieved the highest accuracy overall (on two datasets, Google News(~91%) and News Website(~75.33%)
- Strong performer on Instagram dataset(~82.67%).
- Logistic Regression is most effective for website-based news, possibly due to more structured and formal writing style.

2. **Weakest Model Overall: Naive Bayes**

- Lowest accuracy across 2 datasets (71.33% on website news, 88.33% on Google News, and 79.00% on Instagram which is second least).
- Performs poorly on noisy or informal text (Instagram).
- Shows relatively better performance on structured content like Google News.

3. **Dataset Impact**

- Google News consistently resulted in the highest model performance (up to 90.67% accuracy).
- Structured, clean datasets like Google News are more suitable for fake news detection.
- Informal datasets like Instagram are more challenging but manageable with robust models like Linear SVC.

LITERATURE REVIEW

PAPER 4: Fake News Detection Using Machine Learning

- ***AUTHORS: Jasmine Shaikh, Rupali Patil***
- It reviews existing research on natural language processing (NLP), machine learning (ML), and statistical approaches in fake news iden tification.
- The literature review highlights key advancements, challenges, and gaps in the f ield

PAPER 5: Performance Comparison of Machine Learning Classifiers for Fake News Detection

- ***AUTHORS: Smitha. N, Bharath .R***
- This study provides a comparative analysis of traditional ML classifiers to identify the best balance between accuracy and efficiency
- SVM with TF-IDF achieved best performance (94% accuracy) with lower computational requirements

PAPER 6 : Fake News Detection Enhancement with Data Imputation

- ***AUTHORS: Chandra Mouli Madhav Kotteti, Xishuang Dong, Na Li, Lijun Qian***
- Applied data imputation techniques to handle missing values in datasets
- Multi-Layer Perceptron performed best (45.7% accuracy) after imputation

CONFUSION MATRIX REPRESENTATION

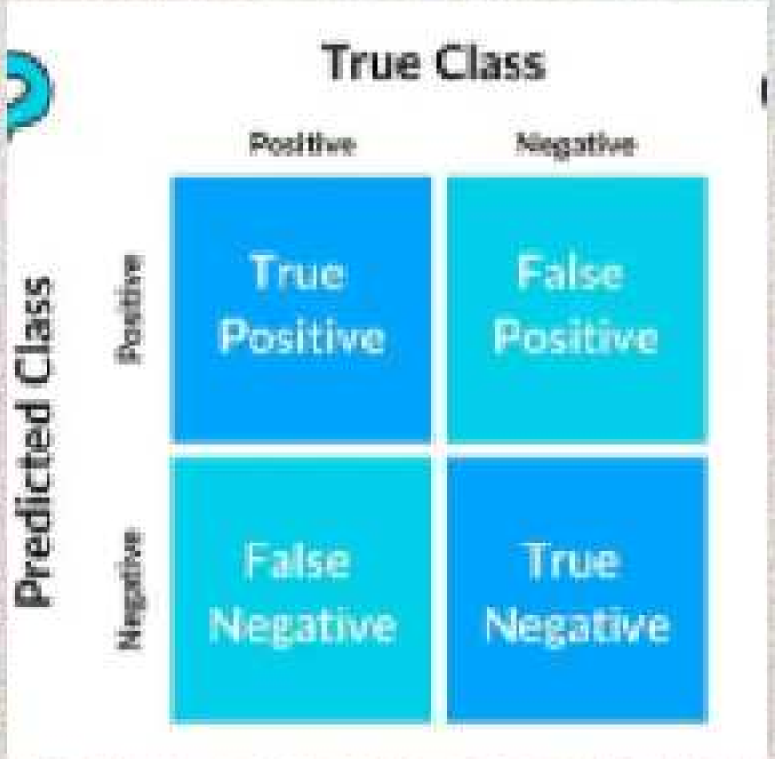
Definition:

A Confusion Matrix is a performance evaluation metric used for classification problems. It provides a tabular breakdown of the actual versus predicted classifications made by a model. It is especially useful for understanding the types of errors a classifier makes.

Formulas:

Using the values TP, TN, FP, FN from the matrix, we compute the following:

1. Accuracy = (TP + TN) / (TP + TN + FP + FN)
2. → Measures overall correctness of the classifier.
3. Precision = TP / (TP + FP)
4. → Indicates how many predicted fake news samples were actually fake.
5. Recall (Sensitivity) = TP / (TP + FN)
6. → Shows how many actual fake news samples were correctly predicted.
7. F1-Score = 2 × (Precision × Recall) / (Precision + Recall)
8. → Harmonic mean of precision and recall. Useful when you need a balance between them.
9. Specificity = TN / (TN + FP)
10. → Measures how well real news is correctly identified.



		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Why It Is Used Here:

- In fake news detection, false positives (real news misclassified as fake) and false negatives (fake news classified as real) have serious implications.
- The Confusion Matrix helps assess not just accuracy, but also the types of errors the model is making.
- It provides a clearer understanding of model performance than accuracy alone, especially in imbalanced datasets where one class may dominate.