ONLINE NEWS POPULARITY PREDICTION

IE7300: FALL 2023 FINAL PROJECT

Presented by :Aditi Chadha

ABSTRACT

In the hyper-competitive social media landscape, content creators grapple with the daunting task of not only reaching targeted audiences but also maximizing audience engagement. This project harnesses machine learning to predict the popularity of online content, framing the task as a binary classification—distinguishing content as **"Popular"** or **"Unpopular"** based on share counts relative to the median. Through exhaustive analysis of pivotal content features such as title, category, sentiment, and keywords, etc., our approach achieved an accuracy of 66%. This predictive model equips creators with invaluable insights, enabling informed decisions regarding resource allocation, content creation, and audience targeting. Such informed strategies empower creators to thrive and excel in the dynamic digital age, where the ability to decode audience preferences is paramount for content success.

INTRODUCTION/OBJECTIVE

The modern landscape of online news consumption is characterized by an abundance of information and a constant struggle to capture audience attention. In this environment, predicting the popularity of online news articles has become crucial for content creators, news organizations, and marketers alike. Understanding which articles are likely to resonate with audiences and generate high engagement allows for informed decision-making, ultimately leading to increased reach, brand recognition, and success.

This project leverages the power of machine learning to predict the popularity of online news articles based on the UCI Online News Popularity Dataset. This dataset provides a collection of features for each article, including various statistics for the new content such as words count, number of links, number of keywords, etc. By analyzing these features with machine learning algorithms, we aim to develop a robust model capable of accurately classifying articles as "Popular" or "Unpopular" based on their potential to generate high shares and engagement.

At its essence, this project is geared toward illuminating the potential popularity of online news articles through a classification framework. Such predictive insights serve as navigational beacons, guiding content creators, news organizations, and marketers toward content strategies that align closely with audience preferences.

DATA OVERVIEW

Source: [UCI Machine Learning Repository - Online News Popularity (Mashable)]

Size: 39,644 records, 61 features (60 + target)

Note: "url" & "timedelta" features dropped (non-predictive)

Column Separation:

- Organized features into categories based on their nature: Words, Links, Digital Media, Keywords, Publication Day, NLP (LDA, Subjectivity & Polarity, Positive & Negative), Channels, and Target.
- This structure helped in analysis and understanding of feature relationships.

Data Preparation:

Target variable ("shares") converted into a binary classification problem:

- **Popular**: Shares > 1,400
- **Unpopular**: Shares <= 1,400

No missing values in the data.

Data types: Numerical

Category	S. No.	Feature Name	Feature Description
1. WORDS	1	n_tokens_title	Number of words in the title
	2	n_tokens_content	Number of words in the content
	3	n_unique_tokens	Rate of unique words in the content
	4	n_non_stop_words	Rate of non-stop words in the content
	5	n_non_stop_unique_tokens	Rate of unique non-stop words in the content
	6	average_token_length	Average length of the words in the content
2. LINKS	1	num_hrefs	Number of links
	2	num_self_hrefs	Number of links to other articles published by Mashable
	3	self_reference_min_shares	Min. shares of referenced articles in Mashable
	4	self_reference_max_shares	Max. shares of referenced articles in Mashable
	5	self_reference_avg_sharess	Avg. shares of referenced articles in Mashable
3. DIGITAL MEDIA	1	num_imgs	Number of images
	2	num_videos	Number of videos
4. KEYWORDS	1	num_keywords	Number of keywords in the metadata
	2	kw_min_min	Worst keyword (min. shares)
	3	kw_max_min	Worst keyword (max. shares)
	4	kw_avg_min	Worst keyword (avg. shares)
	5	kw_min_max	Best keyword (min. shares)
	6	kw_max_max	Best keyword (max. shares)
	7	kw_avg_max	Best keyword (avg. shares)
	8	kw_min_avg	Avg. keyword (min. shares)
	9	kw_max_avg	Avg. keyword (max. shares)
	10	kw_avg_avg	Avg. keyword (avg. shares)
5. CHANNELS	1	data_channel_is_lifestyle	Is data channel 'Lifestyle'?
	2	data_channel_is_entertainment	Is data channel 'Entertainment'?
	3	data_channel_is_bus	Is data channel 'Business'?
	4	data_channel_is_socmed	Is data channel 'Social Media'?
	5	data_channel_is_tech	Is data channel 'Tech'?
	6	data_channel_is_world	Is data channel 'World'?

Category	S. No.	Feature Name	Feature Description
6. PUBLICATION DAY	1	weekday_is_monday	Was the article published on a Monday?
	2	weekday_is_tuesday	Was the article published on a Tuesday?
	3	weekday_is_wednesday	Was the article published on a Wednesday?
	4	weekday_is_thursday	Was the article published on a Thursday?
	5	weekday_is_friday	Was the article published on a Friday?
	6	weekday_is_saturday	Was the article published on a Saturday?
	7	weekday_is_sunday	Was the article published on a Sunday?
	8	is_weekend	Was the article published on the weekend?
7.1 NLP (LDA)	1	LDA_00	Closeness to LDA topic 0
	2	LDA_01	Closeness to LDA topic 1
	3	LDA_02	Closeness to LDA topic 2
	4	LDA_03	Closeness to LDA topic 3
	5	LDA_04	Closeness to LDA topic 4
7.2 NLP (Subjectivity & Polarity)	1	global_subjectivity	Text subjectivity
	2	global_sentiment_polarity	Text sentiment polarity
	3	title_subjectivity	Title subjectivity
	4	title_sentiment_polarity	Title polarity
	5	abs_title_subjectivity	Absolute subjectivity level
	6	abs_title_sentiment_polarity	Absolute polarity level
7.3 NLP (Positive & Negative)	1	global_rate_positive_words	Rate of positive words in the content
	2	global_rate_negative_words	Rate of negative words in the content
	3	rate_positive_words	Rate of positive words among non-neutral tokens
	4	rate_negative_words	Rate of negative words among non-neutral tokens
	5	avg_positive_polarity	Avg. polarity of positive words
	6	min_positive_polarity	Min. polarity of positive words
	7	max_positive_polarity	Max. polarity of positive words
	8	avg_negative_polarity	Avg. polarity of negative words
	9	min_negative_polarity	Min. polarity of negative words
	10	max_negative_polarity	Max. polarity of negative words
8. TARGET	1	shares	Number of shares (target)

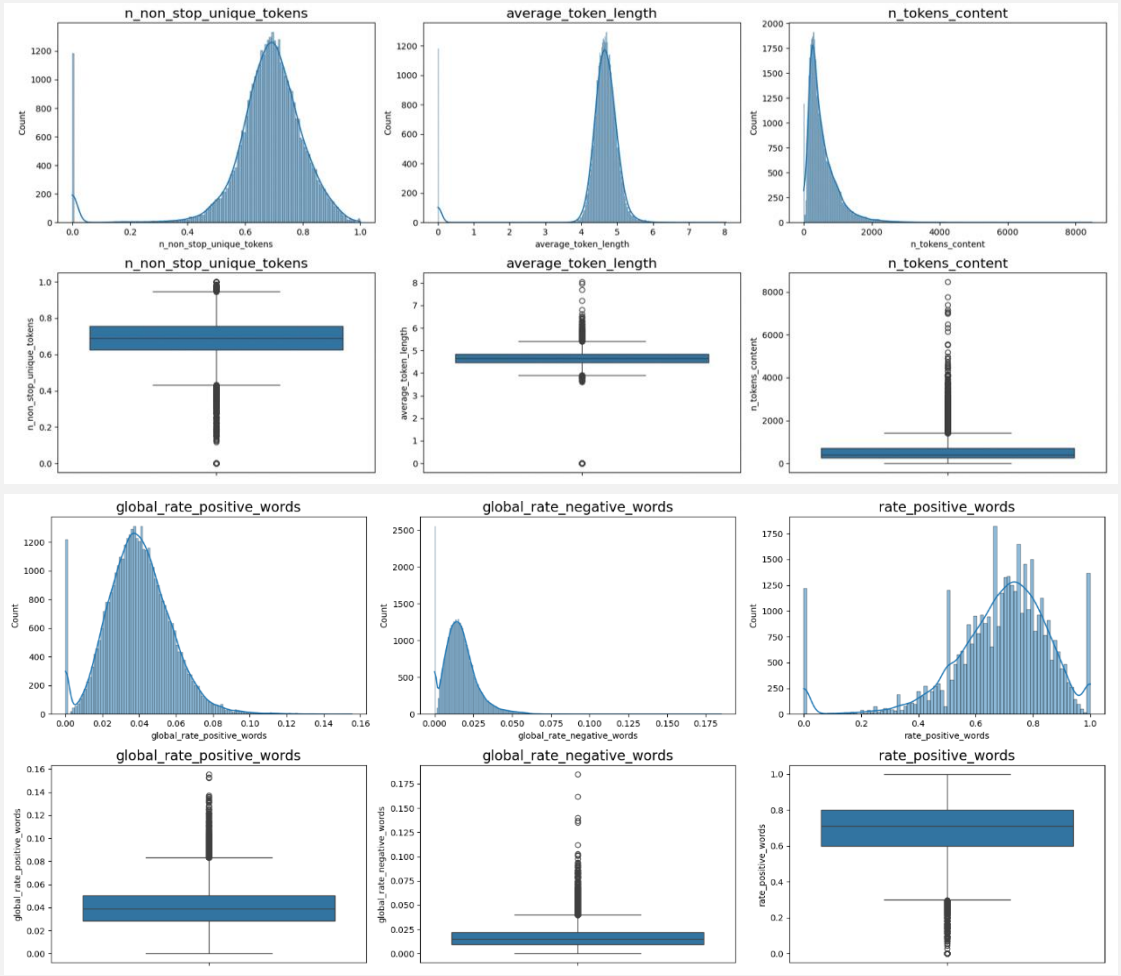
DATA EXPLORATION : STATISTICAL ANALYSIS

1. **Exploring Continuous Variables**
    - **Correlation Analysis:** Utilized Pearson's Correlation matrix to understand relationships between continuous variables.
    - **Visualization:** Employed box plots and histograms to assess normal distribution within the data.
2. **Chi-square Test for Categorical Variables**
    - **Purpose:** Assess association between categorical features and target variable ("shares_c").
    - **Hypothesis:** Null hypothesis: no association; Alternate hypothesis: significant association.
    - **Interpretation:**
        - **P-value < 0.05:** Reject null hypothesis, indicating significant association.
        - **P-value > 0.05:** Fail to reject null hypothesis, suggesting no association.
3. **Outlier Treatment: KNN Imputer**
    - **Approach:** Utilized KNN Imputer to handle outliers within continuous features.
    - **Method:** KNNImputer employs k-nearest neighbors to estimate and replace outlier values based on neighboring data points.
    - **Advantage:** Preserves data structure by leveraging similarities among instances, leading to more accurate imputation and representative results.
4. **Basic Statistics**
    - Checked basic statistics (mean, std, min, max, median, 10th and 90th percentiles) for features in each category to understand feature characteristics

	count	mean	std	min	10%	25%	50%	75%	90%	max
n_tokens_title	39644.0	10.398749	2.114037	2.0	8.000000	9.000000	10.000000	12.000000	13.000000	23.000000
n_tokens_content	39644.0	546.514731	471.107508	0.0	152.000000	246.000000	409.000000	716.000000	1090.000000	8474.000000
n_unique_tokens	39644.0	0.548216	3.520708	0.0	0.406378	0.470870	0.539226	0.608696	0.676714	701.000000
n_non_stop_words	39644.0	0.996469	5.231231	0.0	1.000000	1.000000	1.000000	1.000000	1.000000	1042.000000
n_non_stop_unique_tokens	39644.0	0.689175	3.264816	0.0	0.553379	0.625739	0.690476	0.754630	0.818841	650.000000
average_token_length	39644.0	4.548239	0.844406	0.0	4.302621	4.478404	4.664082	4.854839	5.036971	8.041534

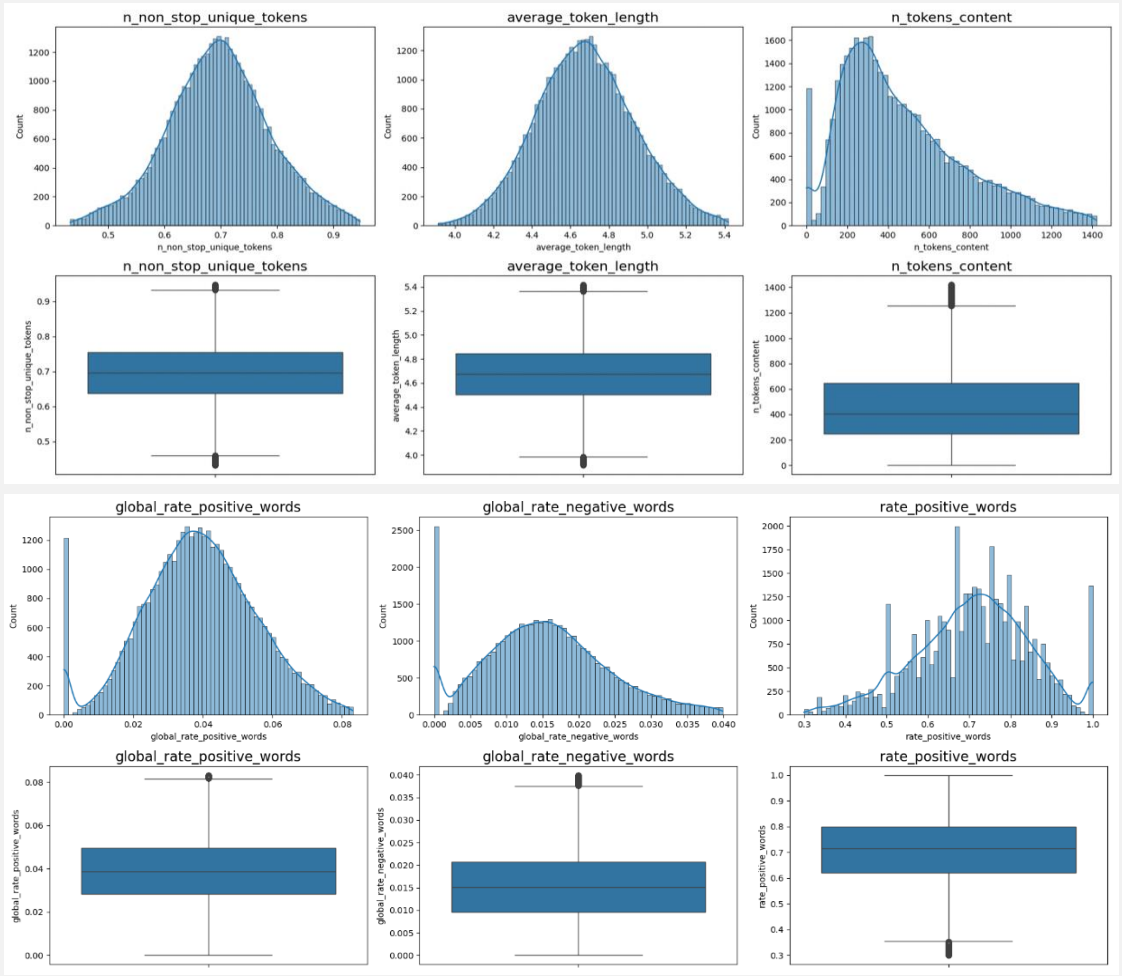
DATA EXPLORATION : OUTLIER TREATMENT

Before



- Histograms and box plots visually demonstrate the skewed distribution of features and the presence of outliers.
- Outliers can significantly impact the performance of machine learning models.
- Almost all continuous features in the dataset contain outliers.

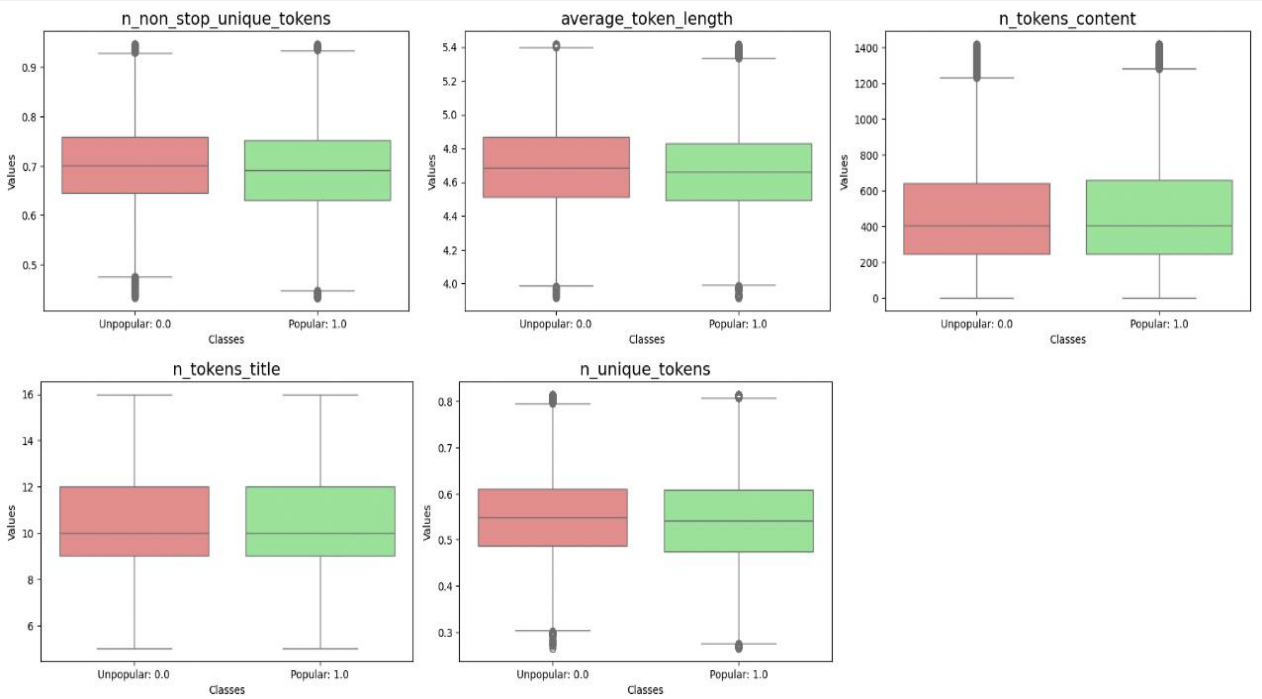
After



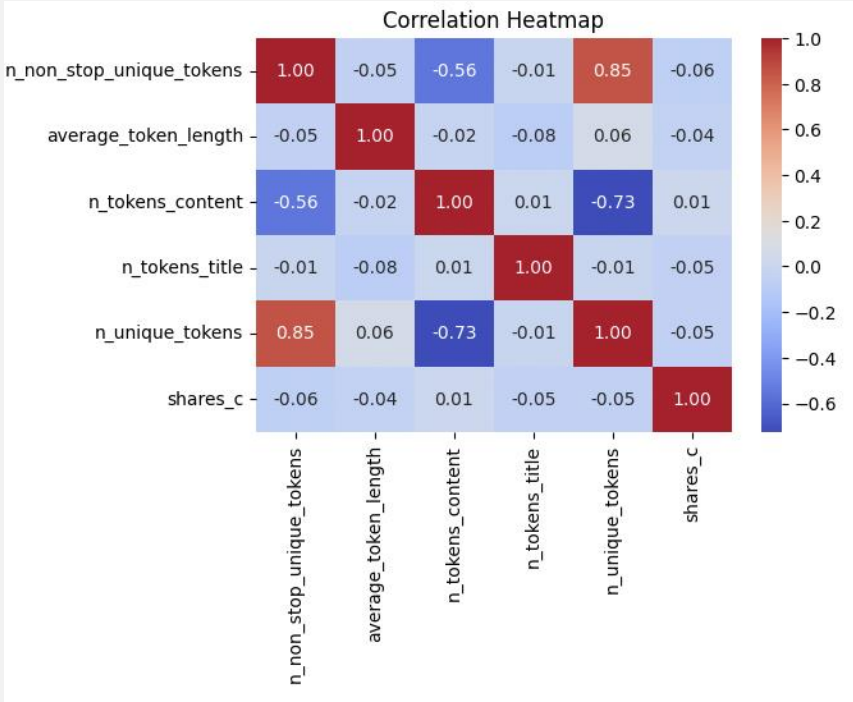
- Outlier treatment using KNN Imputation, are employed to address outliers and normalize the data distribution.
- This ensures that models are not biased by extreme values and can learn more accurate patterns from the data.

DATA EXPLORATION : CONTINUOUS VARIABLES

Boxplot

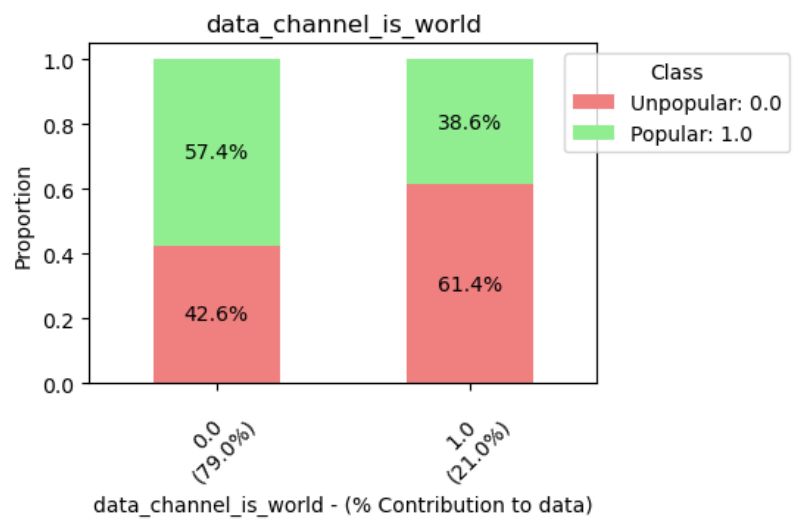


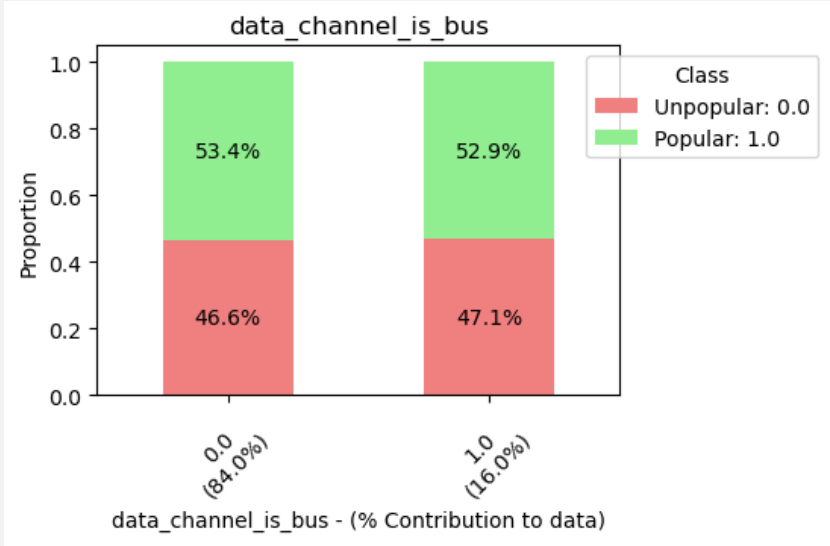
Correlation plot



- The methodology for continuous feature analysis involves visualizing their relationship with the target class to gauge predictive influence. Box plots showcase feature distributions across target classes, while heatmaps reveal feature interrelationships.
- Box plots aid in understanding feature distributions concerning the target class. They provide insights into potential discriminatory power, indicating whether features exhibit distinct behavior for different target categories.
- Correlation heatmaps highlight relationships among continuous features. They assist in identifying multicollinearity, indicating pairs of features highly correlated with each other, aiding in feature selection and model performance.
- The displayed visuals (Boxplot and correlation heatmap) represent a subset, demonstrating the absence of strong feature-target relationships when we see the boxplot and the identification of multicollinearity when we see heatmap, guiding subsequent feature selection and modeling strategies.
- Features with multicollinearity were dropped from the dataset, in this case 'n_unique_tokens' was dropped

DATA EXPLORATION : CATEGORICAL VARIABLES





HYPOTHESES for **data_channel_is_world** and **shares_c**:

- *Null Hypothesis (H0):* There is no association between data_channel_is_world and shares_c.
- *Alternative Hypothesis (H1):* There is an association between data_channel_is_world and shares_c.

STATISTICS:

- **Chi-squared for data_channel_is_world:**941.6751392485123
- **P-value for data_channel_is_world:** 8.55797814281514e-207
- **Degrees of Freedom for data_channel_is_world:** 1

Hypothesis Test Report for data_channel_is_world and shares_c:

- *Result:* **Reject** the null hypothesis
- *Conclusion:* **Significant association** between data_channel_is_world and shares_c.

HYPOTHESES for **data_channel_is_bus** and **shares_c**:

- *Null Hypothesis (H0):* There is no association between data_channel_is_bus and shares_c.
- *Alternative Hypothesis (H1):* There is an association between data_channel_is_bus and shares_c.

STATISTICS:

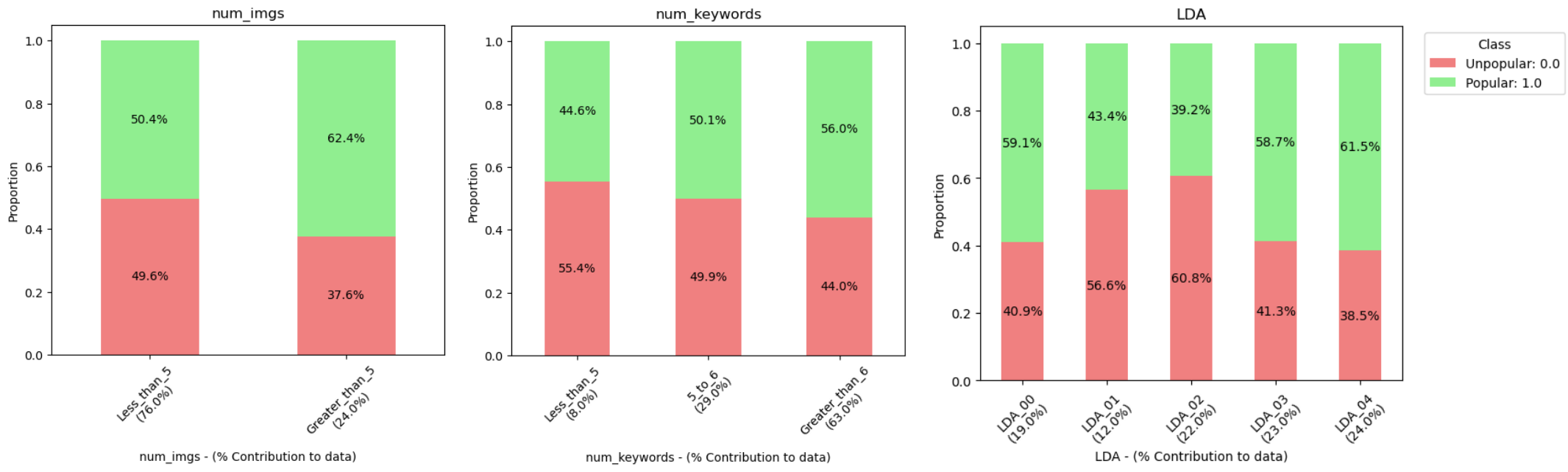
- **Chi-squared for data_channel_is_bus:**0.5430902335655362
- **P-value for data_channel_is_bus:** 0.46115484764846093
- **Degrees of Freedom for data_channel_is_bus:** 1

Hypothesis Test Report for data_channel_is_bus and shares_c:

- *Result:* **Fail** to reject the null hypothesis
- *Conclusion:* **No significant** association between data_channel_is_bus and shares_c.

- Discrete feature analysis involves assessing their distribution across target classes to determine discriminatory power. The chi-squared test helps confirm feature relevance for predictive modeling.
- The feature 'data_channel_is_world' exhibit noticeable distribution variations among classes, indicating potential relevance in prediction.
- Conversely, 'data_channel_is_bus' displays nearly identical distributions across classes, suggesting limited discriminatory power. To assess its relevance, a chi-squared test was conducted.
- Results from the chi-squared test indicate no significant associations for 'data_channel_is_bus' with the target class, failing to reject the Null Hypothesis. As a result, 'data_channel_is_bus' is dropped due to its lack of significant association with the target variable.
- Similarly other discrete features were also analyzed and dropped based on the statistics

FEATURE ENGINEERING: TRANSFORMING INSIGHTS



Feature engineering involves altering or creating new features to extract valuable insights or improve model performance. It plays a pivotal role in extracting more informative patterns from raw data. The three depicted feature engineered features shows the transformation of original discrete or continuous features into more informative or simplified representations, emphasizing their impact on data understanding and model interpretability

1.'num_imgs' Binning for Discriminatory Feature:

1. Custom binning based on 'num_imgs' demonstrates its potential as a discriminative discrete feature.
2. Binning into 'Less_than_5' and 'Greater_than_5' categories highlights significant distribution disparities, simplifying the feature and emphasizing divergence within data.

2.'num_keywords' Discretization:

1. Discretization based on custom bin edges creates concise categories to capture nonlinear relationships with the target variable.

3.Transformation of LDA Features:

1. Addressed skewness and outliers in LDA features by transforming them into a discrete format.
2. Mapping records to LDA topics with maximum probability enhances discriminatory power, as depicted in the bar chart.
3. Bar chart showcases the impact of feature engineering on LDA components, highlighting a clear distinction in the distribution of 'Popular' articles across different LDA components.

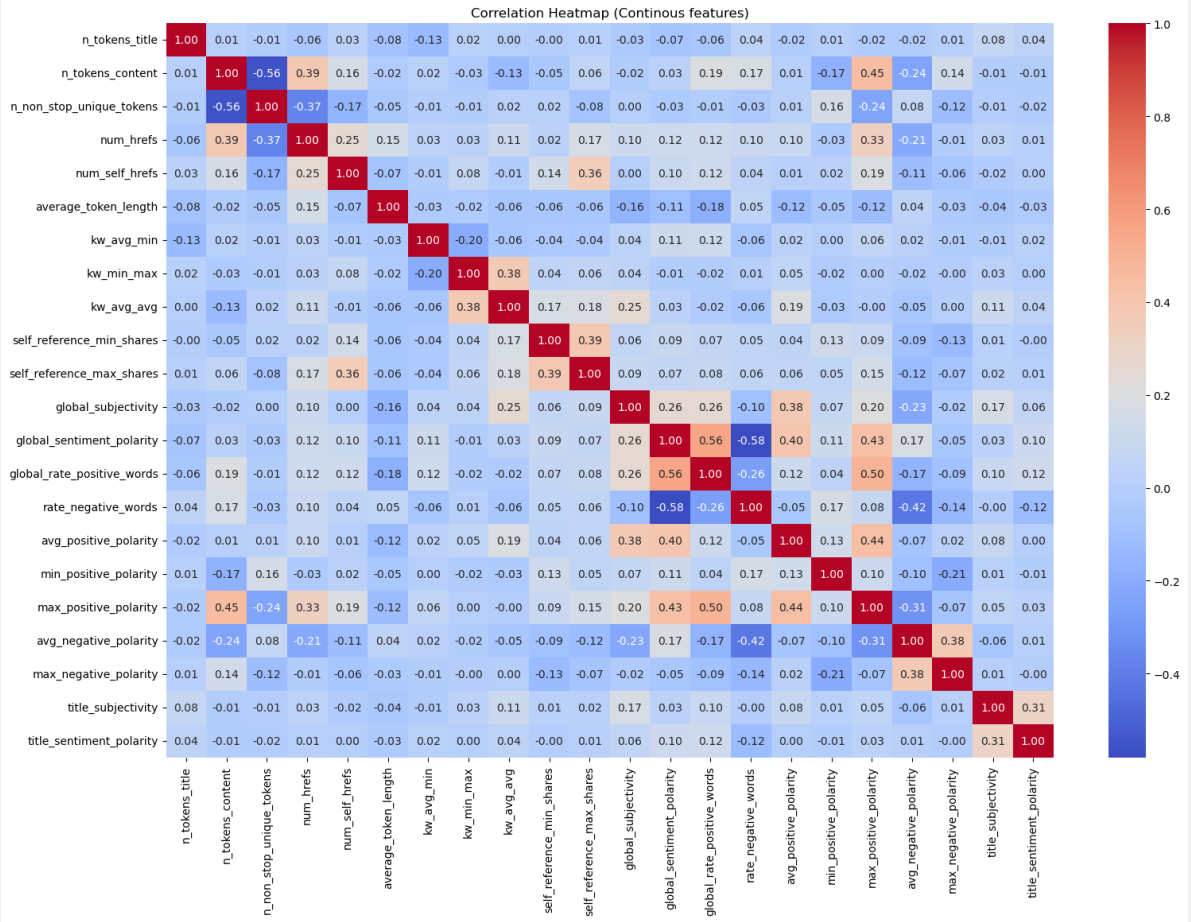
DATA AFTER EDA

Based on the above analysis, we have **dropped 22 features** which are as follows:

n_unique_tokens	self_reference_avg_shares	LDA_00
kw_min_min	kw_max_max	LDA_02
kw_avg_max	kw_max_avg	LDA_04
kw_min_avg	kw_max_min	LDA_01
data_channel_is_bus	weekday_is_saturday	LDA_03
weekday_is_sunday	weekday_is_friday	abs_title_subjectivity
abs_title_sentiment_polarity	global_rate_negative_words	
rate_positive_words	min_negative_polarity	

- **Newly engineered features:**
  - num_imgs
  - num_videos
  - num_keywords
  - LDA

We are now **left with 38 features**, which will be used for prediction.



Upon reviewing the correlation heatmap for the finalized set of continuous features, a noteworthy observation is the absence of any noticeable multicollinearity among these selected features. This outcome is a direct reflection of the comprehensive nature of our Exploratory Data Analysis (EDA) phase. During the EDA, meticulous scrutiny was applied to ensure the inclusion of only the most relevant features, thereby eliminating redundant or highly correlated ones. The absence of multicollinearity in this correlation heatmap strongly indicates the efficacy of our feature selection process, affirming its robustness and the precision of our decision-making during the exploratory phase.

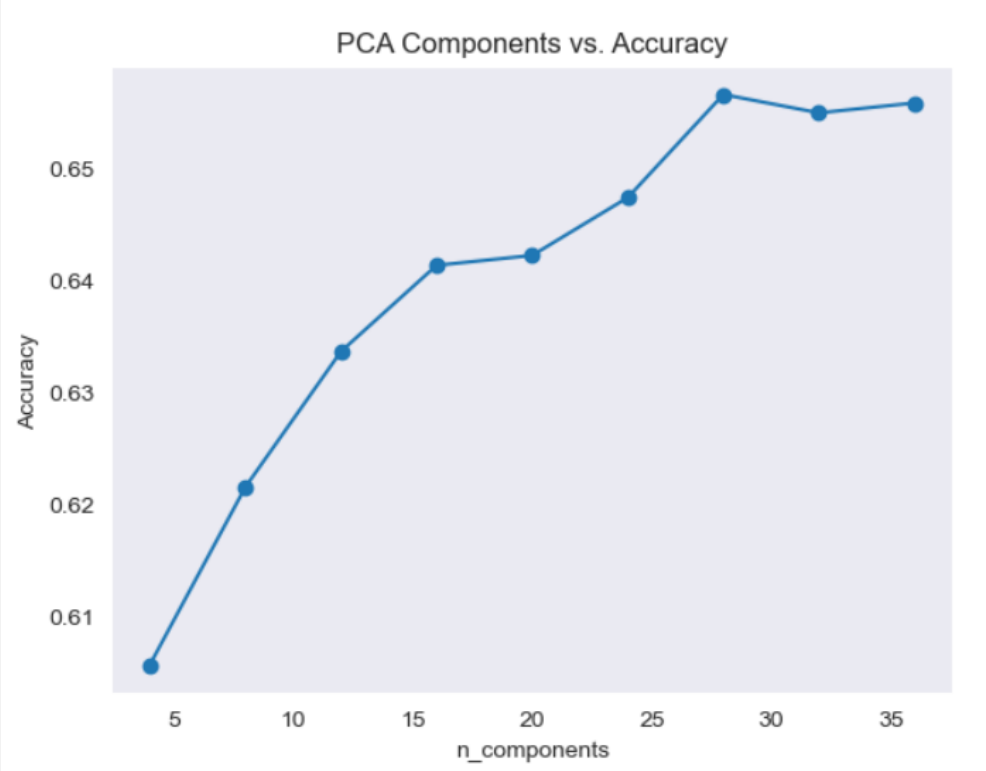
MODEL EVALUATION: LOGISTIC REGRESSION

1. **Approach:**
   - Logistic Regression is employed for binary classification problems, transforming real values into probabilities using the sigmoid function.
   - Key assumptions involve linearity between log-odds and continuous variables and the absence of multicollinearity.
2. **Variable Selection and Model Assessment:**
   - Identified and dropped variables violating linearity assumptions through GLM and verified multicollinearity absence via correlation heatmap.
   - Gradient descent experiments led to a fixed learning rate of 0.000001; no substantial change observed in model performance.
   - PCA experiments indicated optimal accuracy with 28 components but no significant improvement over baseline.
3. **Regularization Analysis (L2 - Ridge):**
   - Implementation of L2 regularization to control overfitting and improve model generalizability.
   - Minimal impact observed on model performance; similar bias-variance tradeoff and generalization to unseen data.
4. **PCA and Model Performance:**
   - PCA reduced dimensionality without drastically impacting accuracy; observed saturation point around 28 components.
   - Plateau in accuracy post 28 components suggests diminishing returns in dimensionality reduction.
5. **Model Observations:**
   - Table showcasing performance metrics for different logistic regression models.
   - Models LR1 and LR2 exhibit similar accuracy and performance; LR3 shows a slight tradeoff between precision and recall post-PCA.
6. **Summary of Findings:**
   - Overall, PCA implementation slightly influences precision and recall while maintaining accuracy and bias-variance tradeoff similar to non-PCA models.
   - Applying PCA alongside regularization doesn't significantly alter model performance.

Model	Regularization (L2)	PCA(28)	Accuracy	Precision	Recall	F1-Score	Train F1 Score	Time(sec)
LR0	Baseline		65.38%	0.667	0.699	0.683	0.683	7.855
LR1			65.67%	0.667	0.710	0.688	0.684	4.0472
LR2	✓		65.67%	0.667	0.710	0.688	0.684	4.0375
LR3		✓	65.65%	0.690	0.646	0.667	0.661	2.4923
LR4	✓	✓	65.65%	0.690	0.646	0.667	0.642	1.9860

PRINCIPAL COMPONENT ANALYSIS AND MODEL ACCURACY

1. **Introduction to Principal Component Analysis (PCA)**
   - PCA serves as a dimensionality-reduction technique, condensing variables while retaining essential information.
   - Balances reduction with data patterns, impacting accuracy with a trade-off.
2. **Relationship between Components and Model Accuracy:**
   - Initial reduction in components leads to a gradual decline in accuracy, indicating information loss in feature space reduction.
   - Notable observation between 28 and 40 components: Accuracy stabilizes, forming a plateau.
   - Beyond 28 components, minimal improvement in accuracy, suggesting diminishing returns in enhancing model performance.
3. **Threshold of Principal Components and Accuracy:**
   - Dropping below 28 components leads to a significant accuracy drop, emphasizing critical information encapsulated within these components.
   - Plateau around 28 to 40 components implies a saturation point, where additional components fail to substantially enhance accuracy.
4. **Implications and Model Optimization:**
   - Balance between dimensionality reduction and retained information is crucial for accurate predictions.
   - Finding the optimal balance to maximize predictive power while minimizing computational complexity is vital for model optimization.
5. **Visual Representation:**
   - n_components vs Accuracy plot illustrating the relationship between the number of principal components and model accuracy.



MODEL EVALUATION: NAÏVE BAYES

1. **Approach:**
   - Based on Bayes' theorem, Naïve Bayes assumes conditional independence among features given the class label.
   - Efficient predictions in high-dimensional spaces due to the "naive" assumption.
2. **Model Application:**
   - Chosen due to the absence of observed collinearity during Exploratory Data Analysis (EDA).
   - Utilized Bernoulli Naïve Bayes for discrete features and Gaussian Naïve Bayes for continuous features, aligning with their respective distributions.
3. **Model Performance Analysis:**
   - Training set performance: Achieved approximately 63-65% accuracy for class 0 and 61-63% for class 1 instances.
   - Consistent performance observed on the test set, demonstrating stable accuracy rates comparable to the training data.
4. **Bias and Variance Evaluation:**
   - Comparable performance between training and test sets indicates a balanced model, avoiding extremes of being overly simplistic or excessively complex.
   - Reflects a reasonable bias-variance tradeoff, capturing data patterns without significant overfitting or underfitting.
5. **Considerations for Improvement:**
   - Although the model showcases a balanced bias-variance tradeoff, there's potential for improving precision and recall for both classes, which could enhance overall model performance.

```
Execution time: 365.3136 seconds


Accuracy for train set: 0.634546257173488
              precision    recall  f1-score   support

         0.0       0.60      0.65      0.62     14732
         1.0       0.67      0.63      0.65     16982

    accuracy                           0.63     31714
   macro avg       0.63      0.64      0.63     31714
weighted avg       0.64      0.63      0.64     31714

Accuracy for test set: 0.6328666918905285
              precision    recall  f1-score   support

         0.0       0.60      0.65      0.63      3758
         1.0       0.66      0.61      0.64      4171

    accuracy                           0.63      7929
   macro avg       0.63      0.63      0.63      7929
weighted avg       0.64      0.63      0.63      7929
```

MODEL EVALUATION: NEURAL NETWORK

1. **Layers and Activation Functions:**
   - Configured with 2 dense layers (32 nodes each), ReLU activation, and a single-node output layer with sigmoid activation for binary outcome prediction.
   - Dropout regularization implemented to enhance model generalizability by preventing overfitting.
2. **Model Complexity vs. Test Metrics:**
   - Increased complexity led to significant impact on test metrics, signifying overfitting.
   - Test accuracy and F1 score declined with rising epochs and batch sizes, while train metrics reached over 95%.
3. **Overfitting Issue:**
   - High train accuracy contrasting with decreasing test metrics indicates overfitting.
   - Model memorizes training data patterns instead of generalizing, resulting in poor performance on new data.
4. **Balancing Bias and Variance:**
   - Models exhibit a balance between bias and variance, as seen in consistent train accuracy across different configurations.
   - Overfitting evident with high train accuracy but diminishing test metrics, suggesting a high-variance scenario.

Model No.	Epochs	Batch Size	Dense Layers	Regularization	Accuracy	Precision	Recall	F1 Score	Time(sec)
NN1	50	32	2		61.48%	0.635	0.654	0.644	336.41
NN2	100	32	2	✓	64.90%	0.648	0.746	0.694	322.98
NN3	150	64	3	✓	60.31%	0.648	0.561	0.601	324.03
NN4	150	128	3	✓	61.69%	0.625	0.703	0.662	206.76
NN5	150	256	3	✓	60.72%	0.631	0.635	0.633	101.06
NN6	150	512	2	✓	61.12%	0.639	0.63	0.633	83.67

MODEL PERFORMANCE COMPARISON

Model No.	Model	Accuracy	Precision	Recall	F1 Score	Time(sec)
LR1	Logistic Regression	65.67%	0.667	0.710	0.688	4.0472
NB1	Naïve Bayes	63.29%	0.660	0.610	0.640	365.313
NN2	Neural Network	64.90%	0.648	0.746	0.694	322.980

1. **Accuracy:**
   - Logistic Regression achieved the highest accuracy of 65.67%, depicting its ability to make correct predictions.
2. **Precision:**
   - Naïve Bayes and Logistic Regression showcased a precision of 0.66, indicating fewer false positives in popular news prediction.
3. **Recall:**
   - Neural Network model (NN2) exhibited the highest recall (sensitivity) of 0.746, ensuring fewer missed popular articles.
4. **F1 Score:**
   - Neural Network model (NN2) achieved the best balance between precision and recall, yielding an F1 score of 0.694.
5. **Time:**
   - Logistic Regression stood out in terms of efficiency, taking significantly less time (4.05 secs) for model fitting and predictions compared to other models.

Conclusion:

- **Selecting the Right Metric:** While Logistic Regression performed best in accuracy, Neural Network (NN2) excelled in recall and F1 Score, offering a more balanced precision-recall tradeoff.
- **Time Efficiency:** Logistic Regression notably outperformed other models in terms of computational time, making it a preferable choice for faster predictions

BIAS-VARIANCE TRADE-OFF AND MODEL GENERALIZABILITY

Variance	Bias	Interpretation	Models
High	Low	Overfitting	NN3, NN4, NN5, NN6
High	High	Underfitting	-
Low	Low	Good Generalizability	LR1, LR2, LR3, LR4,
Low	High	Underfitting	-

Understanding the Trade-off:

- **Bias:** Errors from oversimplified assumptions, leading to underfitting and poor performance on training and new data.
- **Variance:** Sensitivity to fluctuations in training data, causing overfitting with high training accuracy but poor performance on new data.

Factors Impacting Model Generalizability:

1. **Diverse Data:** Our models (LR1, LR2, LR3, LR4) exhibit low variance and low bias, showing good generalizability despite the complexity of the news dataset..
2. **Feature Engineering:** Feature engineering techniques helped retain critical features, allowing logistic regression models to generalize well to new data.
3. **Regularization:** Integration of L2 regularization in logistic regression models ensured a balanced trade-off between simplicity and complexity, contributing to better generalization.

Conclusion:

- **Model Insights:** The Logistic Regression models (LR1, LR2, LR3, LR4) strike a balanced bias-variance trade-off, showcasing a robust performance on unseen data.
- **Key Consideration:** While Neural Network models (NN3, NN4, NN5, NN6) excel in learning complex patterns, they suffer from overfitting, impacting their ability to generalize to new instances.

FINAL OUTCOME AND INSIGHTS

1. **Feature Selection and Engineering:**
   - Comprehensive feature selection identified valuable features and eliminated irrelevant ones, leading to improved model performance.
   - Engineered features, such as "articles published on weekends" and "binned number of images," captured key insights and enhanced prediction accuracy.
   - These findings emphasize the importance of temporal factors and domain knowledge for content engagement prediction.
2. **Model Selection:**
   - Extensive model evaluation identified Logistic Regression as the most suitable choice for predicting article shareability based on accuracy.
   - Logistic Regression demonstrated robust and fast performance across various metrics, making it a reliable and efficient solution.
3. **Model Development Considerations:**
   - Effective handling of outliers to data ensures model integrity and prevents biases.
   - Capturing intricate feature interactions allows the model to learn complex relationships within the data.
   - Leveraging domain research guided feature engineering and improved model effectiveness by reducing the number of unwanted features.
4. **Regularization:**
   - Implementing L2 or dropout regularization techniques helps mitigate overfitting and promotes generalizability, however in our case it didn't have much impact as we effectively perform EDA.
   - Striking the right balance between model complexity and generalization is crucial for real-world application.
5. **Overall Outcome:**
   - This comprehensive approach has resulted in an accurate and effective model for predicting online news popularity.
   - Valuable insights gained from the analysis provide a deeper understanding of content popularity factors.

Final note:

This project provided a valuable opportunity to apply the concepts learned in IE7300 to a real-world problem. By incorporating techniques like outlier treatment, feature engineering, regularization, and model selection, we were able to develop a robust and accurate model for predicting article shareability. While the achieved accuracy of 66.7% may seem relatively low, it is consistent with the findings of other researchers working with this challenging dataset. The inherent difficulty of this dataset, characterized by outliers and complex relationships, highlights the importance of the theoretical foundations and practical skills gained through this course.

THANK YOU

My heartfelt gratitude to Professor and teaching assistants for helping me throughout the semester and guiding me whenever I got stuck and needed help.