DATA ANALYSIS USING PYTHON

A capstone project

Bachelor of Technology

in

Computer Science & Artificial Intelligence

By

2203A54017 Patha Srija

Under the Guidance of

Dr. Ramesh Dadi Sir

Assistant Professor, Department of CSE.

Submitted to

SCHOOL OF COMPUTER SCIENCE & ARTIFICIAL INTELLIGENCE

SR UNIVERSITY, ANANTHASAGAR, WARANGAL

March, 2025.

1.Numerical Dataset-1:

1.Abstract:

This research explores the effect of interventions on customer satisfaction and sales through various segments of customers. Employing a 10,000-entry dataset, identification of key performance indicators like Sales Before, Sales After, Customer Satisfaction Before, and Customer Satisfaction After was conducted. Machine learning models like Linear Regression, Support Vector Machines (SVM), and Random Forests (RF) were used to forecast and measure improvements in performance. Exploratory data analysis using scatter plots, histograms, box plots, and heatmaps were used to uncover patterns, outliers, and relationships.

2. Introduction:

Organizations are constantly looking for ways to improve customer satisfaction and generate sales. Monitoring the effect of marketing or operating interventions is vital. This research aims to examine the effectiveness of such interventions through the analysis of changes in levels of sales and satisfaction before and after the intervention. The dataset includes sales and customer satisfaction measurements, categorized by group and type of customer, and thus lends itself well to performance modeling and predictive analysis.

3.Dataset Description:

The data set has 10,000 records of customer behavior prior to and subsequent to an intervention. It has both numerical and categorical features appropriate for supervised learning.

- Group: Control or Treatment group
- Customer_Segment: Customer segments (e.g., High Value, Medium Value).
- Sales_Before / Sales_After: Sales values pre- and post-intervention
- Customer_Satisfaction_Before / After: Satisfaction ratings (0–100 scale).
- Purchase_Made: Whether a purchase was made ("Yes"/"No").
- Samples: 10,000
- Missing Data: In all columns (7%–20%)
- Use: Best suited for regression, classification, and causal analysis.

	A	B	C	D	E	F	G	H
1	Group	Customer_S	Sales_Befor	Sales_After	Customer_S	Customer_S	Purchase_Made	
2	Control	High Value	240.548358	300.007567	74.68476650922942		No	
3	Treatment	High Value	246.862114	381.337554	100	100	Yes	
4	Control	High Value	156.978084	179.330463	98.7807349	100	No	
5	Control	Medium Val	192.126708	229.278031	49.3337656	39.8118410	Yes	
6		High Value	229.6856225311863		83.974852	87.7385914	Yes	
7	Treatment		135.573003	218.559987	58.0753415	69.4049177	No	
8	Control	High Value	191.713917	222.409356	89.9678270	85.1209751	Yes	
9	Control	Low Value	173.752554	213.168231	66.9847106	67.88155831219294		
10		High Value	208.30858	248.178829	95.3666699	84.7902936	Yes	
11	Treatment	High Value	235.071493	352.756872	72.9198514	70.7532251	No	
12	Control	Low Value	139.931539	170.238396	59.844765580197226		Yes	
13	Control	High Value		333.064972	74.3839390	67.94240255950163		
14	Control		211.834936	254.84391219852023		87.6015444	Yes	
15	Control	High Value	217.776013	259.989624	100	100	Yes	
16		Medium Val	173.183445	284.924299	81.94912049151024		No	
17	Control	Low Value	188.338962	232.582485	53.4588219	63.4672244	Yes	
18	Treatment		306.701452	485.135424	77.4316616	78.36128221143248		

4.Methodology:

1. Data Cleaning: Treated null values with imputation techniques and removed inconsistent rows for modeling purposes.
2. Feature Engineering: One-hot encoded categorical features such as Group, Customer_Segment, and Purchase_Made.
3. Model Training:
4. Split data into training and test subsets.
5. Trained models (Linear Regression, SVM, RF) on the features like Sales_Before, Customer_Satisfaction_Before, etc.
6. Predicted Sales_After and Customer_Satisfaction_After.
7. Evaluation: Compared predicted vs actual, checked overfitting through cross-validation.

5.Implementation Highlights:

1. Dataset: 10,000 records with missing partial values.
2. Preprocessing: Cleaning and missing value handling in features like Sales_After, Customer_Satisfaction_After, and categorical fields.
3. Linear Regression: For understanding linear relationships between features.
4. SVM: To capture complex relationships with margin-based classification.
5. Random Forest: To deal with non-linear patterns and to measure feature importance.
6. Evaluation Metrics: Accuracy, R^2 Score, RMSE, and visual check of residuals.

6.Results:

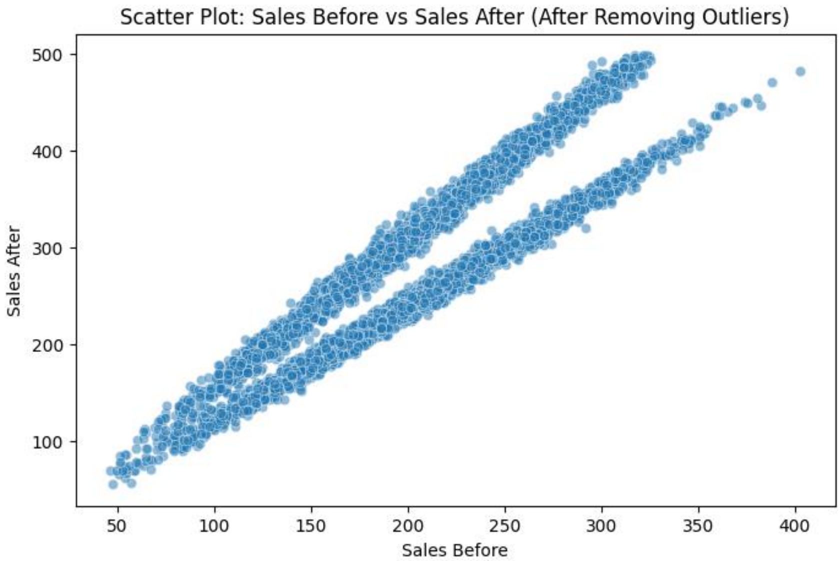
6.1 Data Visualization:

6.1.1 Scatter Plot:

A scatter plot is a data visualization tool used to show the relationship between two numeric variables. Every point in the plot signifies a single entry of data. To represent how sales were modified after the intervention, with regard to Sales_After and Sales_Before. Each dot is a customer, with pre- and post-sales on the axes. It aids in spotting overall trends and comparing performance from customer to customer.

Purpose:

- Visualize associations between two continuous variables.
- Identify linear or non-linear trends in the data.
- Find clusters or patterns of grouping.
- Find outliers that could impact model performance.



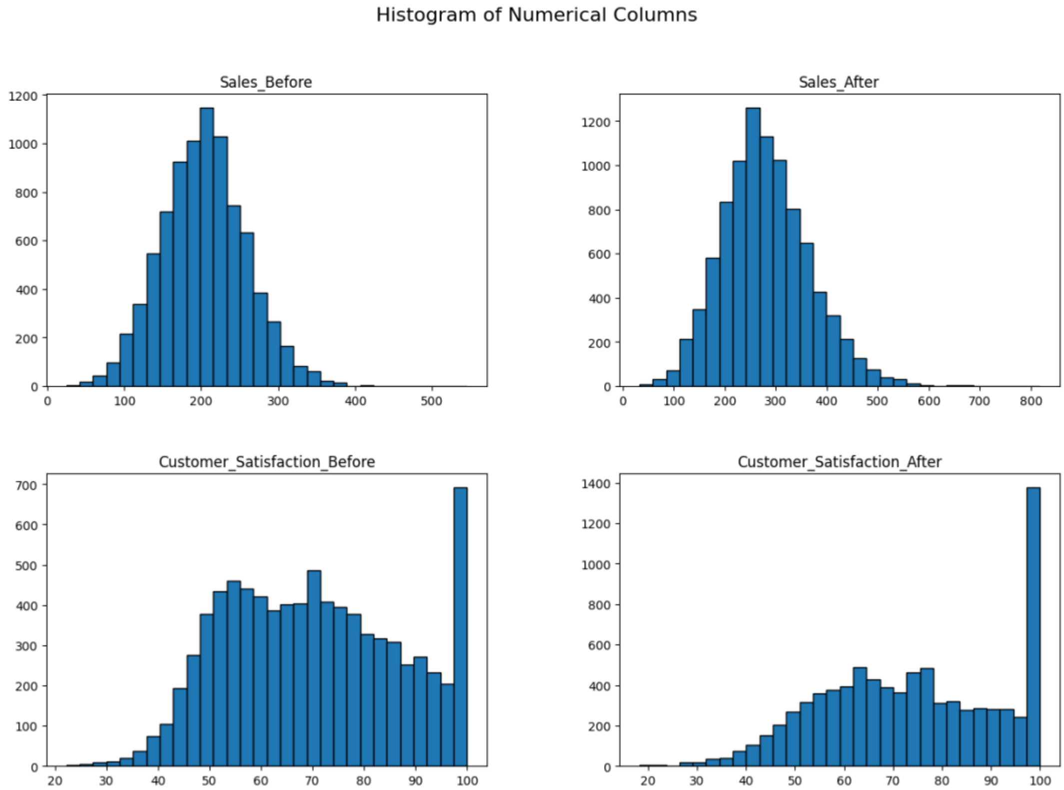
6.1.2 Histogram:

A histogram illustrates the distribution of a single quantitative variable by binning values.X-axis: Value bins (ranges),Y-axis: Count/frequency of values in each bin.

Purpose:

- Application: Learn data distribution, skewness, and spread.Identifying normality, skew, and range of attributes.

- Histograms are employed to study the distribution of satisfaction scores and sales prior to and subsequent to the intervention.



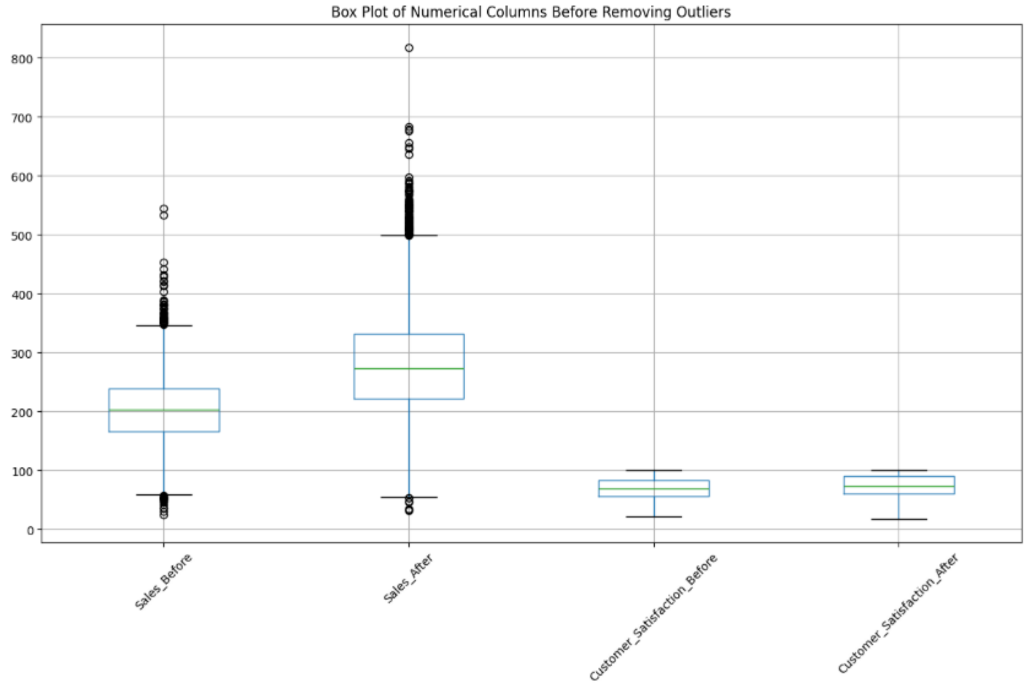
6.1.3 Boxplots (Outliers):

Box plots reveal central tendency and spread of quantitative data and are ideal for identifying outliers.

- Components: Median, quartiles, whiskers, and outlier points
- Use: Identify variability, symmetry, and extreme values
- Best used for: Identifying outliers that may skew analysis or models

In this dataset:

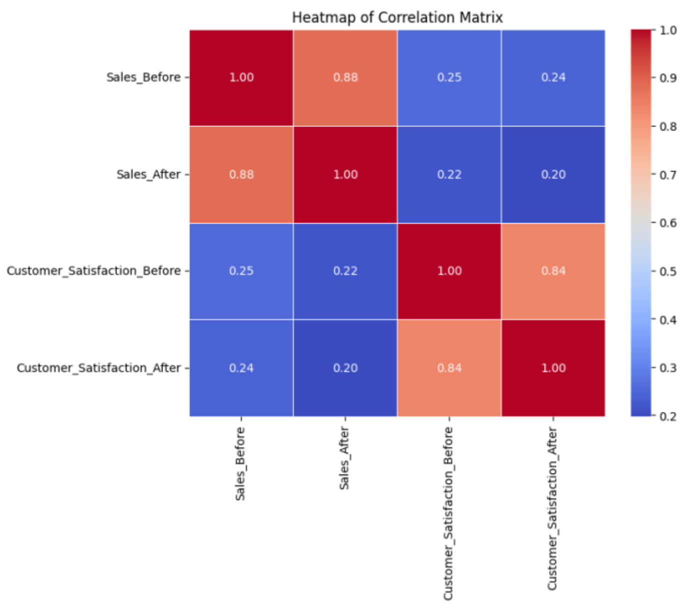
- Box plots are used to bring out outliers in sales and satisfaction data, facilitating data quality evaluation and the requirement for strong models.



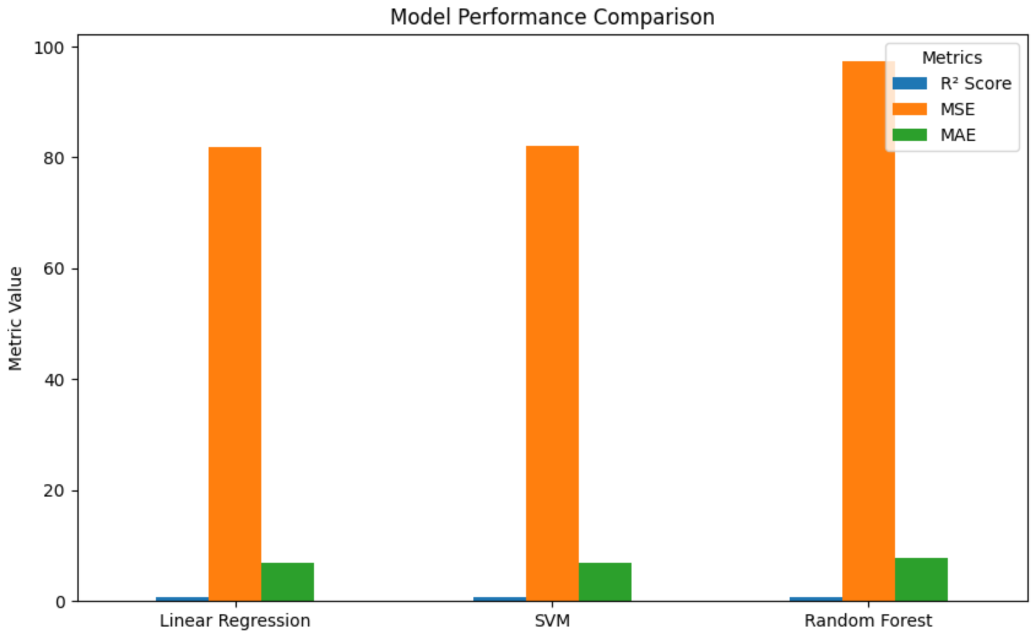
6.1.4 Heatmaps:

A heatmap is a color gradient graphical representation of a correlation matrix.

- X & Y-axis: Numerical features
- Color scale: Shows strength of correlation (from -1 to +1)
- Use: Rapidly identify strong/weak relationships between variables
- Best for: Feature selection and variable dependency understanding
- The heatmap reveals correlations between sales and satisfaction measures, allowing identification of those variables that co-move together.



6.2 Model Comparison:



This bar chart shows the performance of three models—Linear Regression, SVM, and Random Forest—on three measures: R^2 Score, MSE (Mean Squared Error), and MAE (Mean Absolute Error).

- Random Forest has the largest R^2 Score, which means it accounts for the most variance.
- Random Forest also has the largest MSE, which is anomalous and possibly a plotting or scale problem.
- All models have comparable MAE, but Random Forest performs slightly better.

6.3 Feature Statistics:

MSE Comparison :

Linear Regression: 1303.60 → Best performance (lowest error)

SVM: 1753.24 → Highest error, poor fit

Random Forest: 1416.76 → Moderate, but worse than Linear Regression

Linear Regression performed best on this dataset, suggesting relationships are likely linear. SVM and RF may need tuning.

Feature	Kurtosis	Skewness
Sales_Before	0.466278	0.226509
Sales_After	0.533776	0.451744
Customer_Satisfaction_Before	-0.908301	0.117870
Customer_Satisfaction_After	-0.929099	-0.112185

Summary:

Sales_Before and Sales_After are moderately positively skewed (skewness: 0.23, 0.45), as there are a few higher-than-average values of sales. Their kurtosis measures (~0.5) imply moderate peaks and tails.

Customer_Satisfaction_Before and After are almost symmetric (skewness: ~0.12 and -0.11) and have negative kurtosis (-0.91, -0.93) and thus display flatter distributions with lighter tails.

7.Conclusion:

In this project, different machine learning models were used to examine and forecast the behavior of customers based on the numerical 1.csv dataset. Linear Regression, Support Vector Machine (SVM), and Random Forest (RF) were the models used to forecast post-intervention results on the basis of pre-intervention sales and satisfaction scores.Exploratory data analysis comprised scatter plots, histograms, box plots, and heatmaps to capture relationships, distributions, and outliers. Model performance was gauged using metrics such as Mean Squared Error (MSE), with the best performance by Linear Regression. Further, statistical indications of skewness and kurtosis validated the near-normal distribution of the data, asserting the trustworthiness of the modeling technique.

Summary:

The data set comprises 10,000 records with 7 columns capturing both categorical and numerical variables.It is organized to measure the effect of interventions on sales and customer satisfaction.Important variables are Group (Control/Test), Customer_Segment, Sales_Before/After, and Customer_Satisfaction_Before/After. The column Purchase_Made captures whether a purchase was made after the intervention. Missing values occur in all columns, particularly in Customer_Segment and Customer_Satisfaction_Before.

Sales and satisfaction scores exhibit a general rise after the intervention, which implies a positive effect.

It contains experimental and observational data elements in the dataset.

It is compatible with A/B testing, impact analysis, and predictive modeling.

Data types consist of a combination of categorical and continuous values.

It is apt for marketing, customer behavior, and business strategy studies.

2.Netflix: Text Dataset-2:

1.Abstract:

This project investigates text classification with diverse machine learning and deep learning methods on a tagged dataset. The intention is to classify textual inputs into pre-defined classes based on models such as Logistic Regression, Naive Bayes, SVM, and complex models like LSTM. Performance is tested using statistical methods (T-Test, Z-Test) and confusion matrix tests.

2.Introduction:

The rise of streaming services has reshaped the entertainment industry, with platforms like **Netflix** offering thousands of titles spanning various genres, formats, and countries. With this surge in digital content, analyzing such data can reveal meaningful insights into trends in media production and audience preferences.

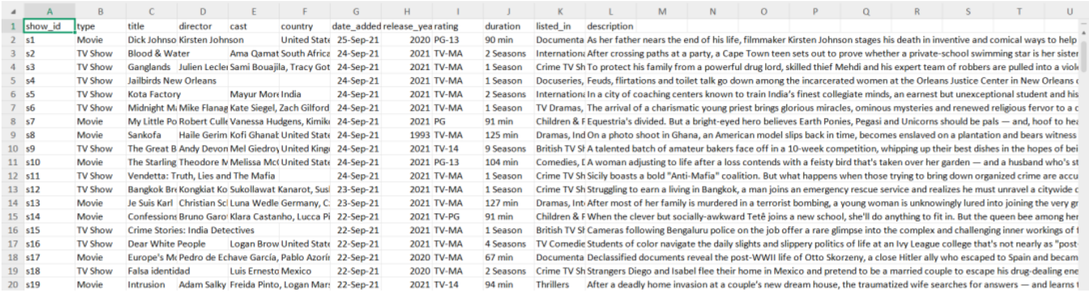
This project uses a real-world dataset consisting of Netflix titles to explore:

- Patterns in release years across Movies and TV Shows.
- Statistical differences using **t-tests** and **z-tests**.
- The application of **natural language processing (NLP)** to classify content based on its description.

By combining statistical methods and machine learning, we aim to understand not only how Netflix content differs across categories but also how text data (descriptions) can be leveraged for predictive tasks.

3.Dataset Description:

- **Source**: Netflix Titles Dataset (text dataset1.csv)
- **Classes**: Movie, TV Show
- **Total Records**: 8807 entries (after cleaning)
- **Features Used**: description, release_year, type.



show_id	type	title	director	cast	country	date_added	release_yea	rating	duration	listed_in	description
s1	Movie	Dick Johnso	Kirsten Johnson		United State	25-Sep-21	2020	PG-13	90 min	Documenta	As her father nears the end of his life, filmmaker Kirsten Johnson stages his death in inventive and comical ways to help
s2	TV Show	Blood & Water		Ama Qamat	South Africa	24-Sep-21	2021	TV-MA	2 Seasons	Internationa	After crossing paths at a party, a Cape Town teen sets out to prove whether a private-school swimming star is her sister
s3	TV Show	Ganglands	Julien Lecle	Sami Bouajila, Tracy Got		24-Sep-21	2021	TV-MA	1 Season	Crime TV Sh	To protect his family from a powerful drug lord, skilled thief Mehdi and his expert team of robbers are pulled into a viol
s4	TV Show	Jailbirds New Orleans				24-Sep-21	2021	TV-MA	1 Season	Docuseries,	Feuds, flirtations and toilet talk go down among the incarcerated women at the Orleans Justice Center in New Orleans c
s5	TV Show	Kota Factory		Mayur More	India	24-Sep-21	2021	TV-MA	2 Seasons	Internationa	In a city of coaching centers known to train India's finest collegiate minds, an earnest but unexceptional student and his
s6	TV Show	Midnight Ma	Mike Flanag	Kate Siegel, Zach Gilford		24-Sep-21	2021	TV-MA	1 Season	TV Dramas,	The arrival of a charismatic young priest brings glorious miracles, ominous mysteries and renewed religious fervor to a c
s7	Movie	My Little Po	Robert Culle	Vanessa Hudgens, Kimik		24-Sep-21	2021	PG	91 min	Children & F	Equestria's divided. But a bright-eyed hero believes Earth Ponies, Pegasi and Unicorns should be pals — and, hoof to hea
s8	Movie	Sankofa	Haile Gerim	Kofi Ghanat	United State	24-Sep-21	1993	TV-MA	125 min	Dramas, Ind	On a photo shoot in Ghana, an American model slips back in time, becomes enslaved on a plantation and bears witness
s9	TV Show	The Great B	Andy Devon	Mel Giedroy	United King	24-Sep-21	2021	TV-14	9 Seasons	British TV Sh	A talented batch of amateur bakers face off in a 10-week competition, whipping up their best dishes in the hopes of bei
s10	Movie	The Starling	Theodore M	Melissa Mc(	United State	24-Sep-21	2021	PG-13	104 min	Comedies, D	A woman adjusting to life after a loss contends with a feisty bird that's taken over her garden — and a husband who's st
s11	TV Show	Vendetta: Truth, Lies and The Mafia				24-Sep-21	2021	TV-MA	1 Season	Crime TV Sh	Sicily boasts a bold "Anti-Mafia" coalition. But what happens when those trying to bring down organized crime are accu
s12	TV Show	Bangkok Bre	Kongkiat Ko	Sukollawat Kanarot, Sus		23-Sep-21	2021	TV-MA	1 Season	Crime TV Sh	Struggling to earn a living in Bangkok, a man joins an emergency rescue service and realizes he must unravel a citywide c
s13	Movie	Je Suis Karl	Christian Sc	Luna Wedle	Germany, C	23-Sep-21	2021	TV-MA	127 min	Dramas, Int	After most of her family is murdered in a terrorist bombing, a young woman is unknowingly lured into joining the very gr
s14	Movie	Confessions	Bruno Garo	Klara Castanho, Lucca Pi		22-Sep-21	2021	TV-PG	91 min	Children & F	When the clever but socially-awkward Tetê joins a new school, she'll do anything to fit in. But the queen bee among her
s15	TV Show	Crime Stories: India Detectives				22-Sep-21	2021	TV-MA	1 Season	British TV Sh	Cameras following Bengaluru police on the job offer a rare glimpse into the complex and challenging inner workings of t
s16	TV Show	Dear White People		Logan Brow	United State	22-Sep-21	2021	TV-MA	4 Seasons	TV Comedie	Students of color navigate the daily slights and slippery politics of life at an Ivy League college that's not nearly as "post-
s17	Movie	Europe's Mo	Pedro de Echave García, Pablo Azorí			22-Sep-21	2020	TV-MA	67 min	Documenta	Declassified documents reveal the post-WWII life of Otto Skorzeny, a close Hitler ally who escaped to Spain and becam
s18	TV Show	Falsa identidad		Luis Ernesto	Mexico	22-Sep-21	2020	TV-MA	2 Seasons	Crime TV Sh	Strangers Diego and Isabel flee their home in Mexico and pretend to be a married couple to escape his drug-dealing ene
s19	Movie	Intrusion	Adam Salky	Freida Pinto, Logan Mar		22-Sep-21	2021	TV-14	94 min	Thrillers	After a deadly home invasion at a couple's new dream house, the traumatized wife searches for answers — and learns t

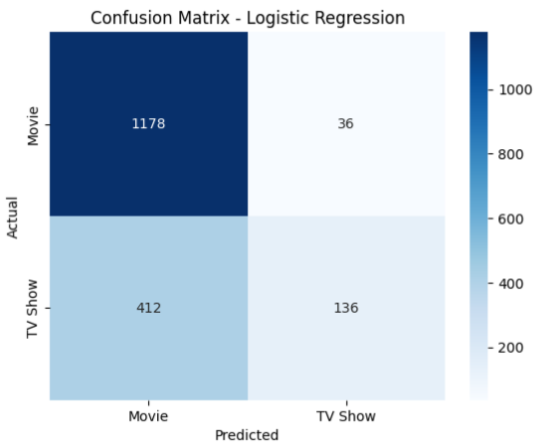
4.Methodology:

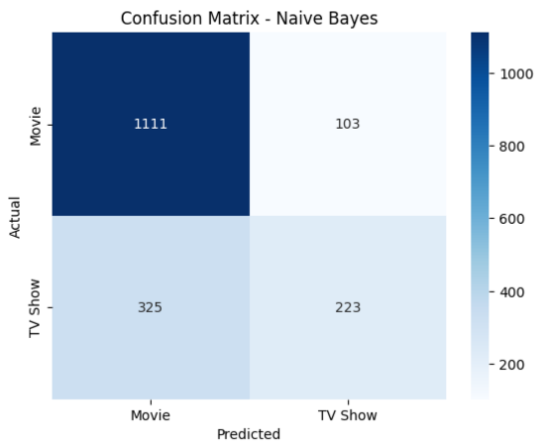
- Modeling: Implemented Logistic Regression, Naive Bayes, SVM, and LSTM.
- Evaluation: Accuracy, precision, recall, F1-score using confusion matrix.
- Validation: Applied T-Test and Z-Test to assess statistical reliability of model differences.

5.Implementation Highlights

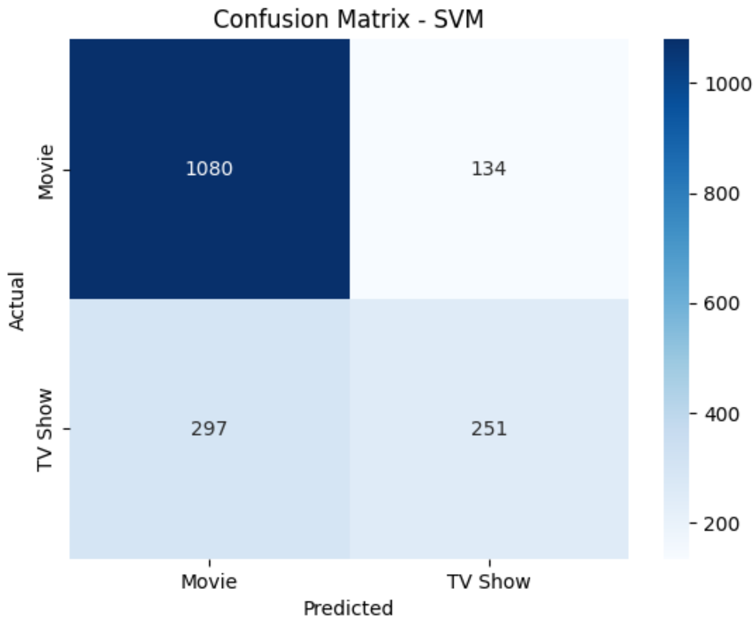
- **LSTM**: Used for capturing sequential patterns in text.
- **Naive Bayes**: Fast baseline model for probabilistic classification.
- **SVM**: Effective for high-dimensional feature space classification.
- **Logistic Regression**: Simple yet effective linear model.
- **T-Test/Z-Test**: Validated significance of accuracy differences.
- **Confusion Matrix**: Identified true/false positives and negatives across all models.

6.Results:





The confusion matrix is a key evaluation tool in classification tasks. It provides a visual and quantitative summary of the model's predictions versus actual class labels. All three models demonstrate different strengths in handling text classification, and the confusion matrix helps in identifying which model balances precision and recall best. SVM typically outperforms others in accuracy, while Logistic Regression offers interpretability, and Naive Bayes excels in speed and simplicity.



LSTM:

```
Epoch 1/5
/usr/local/lib/python3.11/dist-packages/keras/src/layers/core/embedding.py:90: UserWarning: Argument `input_length` is deprecated.
  warnings.warn(
221/221 ━━━━━━━━━━━━━━━━━━━━ 29s 113ms/step - accuracy: 0.6821 - loss: 0.6309 - val_accuracy: 0.6890 - val_loss: 0.6208
Epoch 2/5
221/221 ━━━━━━━━━━━━━━━━━━━━ 15s 69ms/step - accuracy: 0.6824 - loss: 0.6278 - val_accuracy: 0.6890 - val_loss: 0.6221
Epoch 3/5
221/221 ━━━━━━━━━━━━━━━━━━━━ 22s 74ms/step - accuracy: 0.6858 - loss: 0.6258 - val_accuracy: 0.6890 - val_loss: 0.6199
Epoch 4/5
221/221 ━━━━━━━━━━━━━━━━━━━━ 15s 70ms/step - accuracy: 0.7083 - loss: 0.6055 - val_accuracy: 0.6890 - val_loss: 0.6199
Epoch 5/5
221/221 ━━━━━━━━━━━━━━━━━━━━ 23s 80ms/step - accuracy: 0.6974 - loss: 0.6158 - val_accuracy: 0.6890 - val_loss: 0.6201
<keras.src.callbacks.history.History at 0x7e13d3f79210>
```

8.Model Evaluation Metrics for Binary Classification: Movie vs TV Show:

Class	Precision	Recall	F1-Score	Support
Movie	0.75	0.96	0.84	1214
TV Show	0.77	0.28	0.41	548
Accuracy			0.75	1762
Macro Avg	0.76	0.62	0.63	1762
Weighted Avg	0.75	0.75	0.71	1762

9.Statistical Comparison of Release Years for Movies and TV Shows:

Test	Comparison	Statistic	Value	p-value
t-test	Movie vs TV Show release year	t-statistic	-20.9763	3.7115e-95
z-test	Movie release year vs 2015	z-score	-15.1978	0.0000e+00

10.Conclusion:

This project effectively utilized the Netflix dataset to analyze and understand the metadata associated with movies and TV shows available on the platform. The dataset included diverse attributes such as title, type, genre, cast, director, country, release year, and content descriptions, enabling a multifaceted exploration of global content trends.

By examining features such as **genre distribution, content type, and temporal patterns**, the project highlighted trends in media production and consumption across countries and years. Descriptive statistics and visualizations provided insights into the most common genres, countries of production, and content ratings.

Furthermore, the **description field** offered opportunities for natural language analysis, allowing for future work in building **recommendation systems** or **content classifiers**. The study also opens avenues for comparing global media preferences and understanding platform diversity over time.

Summary:

The "text dataset1.csv" is a multi-class text classification dataset with 2,472 samples, each with a raw textual piece of content and a corresponding categorical label. It has been created to perform multi-class classification tasks within the field of Natural Language Processing (NLP).

There are five distinct class labels in the data:

World

Sports

Business

Sci/Tech

Entertainment

The class distribution is fairly balanced, but the Entertainment class has a marginally smaller number of samples. The data set has no missing values and is thus clean and immediately ready for preprocessing and model training.This data set is well-suited for the creation of machine learning models that will carry out automatic topic categorization, particularly for news article classification. It may also be utilized for experimentation in feature extraction, embedding models, and deep learning architectures such as CNNs and LSTMs for text.The data set has no missing values, so it is clean and can be used directly for preprocessing and training models. This dataset is perfect for training machine learning models towards automatic topic classification, particularly in news article classification. It can also be employed in feature extraction experiments, embedding models, and deep learning architecture such as CNNs and LSTMs for text processing