**Content Virality Prediction - Entertainment Sector**

Pooja Joshi & E.Code [E25007]

# Overview :

This report outlines the methodology, progress, and findings of the Content Virality Prediction project in the Entertainment Sector. The primary focus is on predicting the virality of content based on various features such as valuation, popularity, and audience sentiment.

# Objective:

* + Prepare and preprocess the dataset.
  + Train a machine learning model to predict content virality.
  + Evaluate the model's performance using relevant metrics.
  + Visualize the results to gain insights into the factors affecting virality.

# Assigned Task(s) :

·· Data preparation and cleaning.

· Feature selection and encoding.

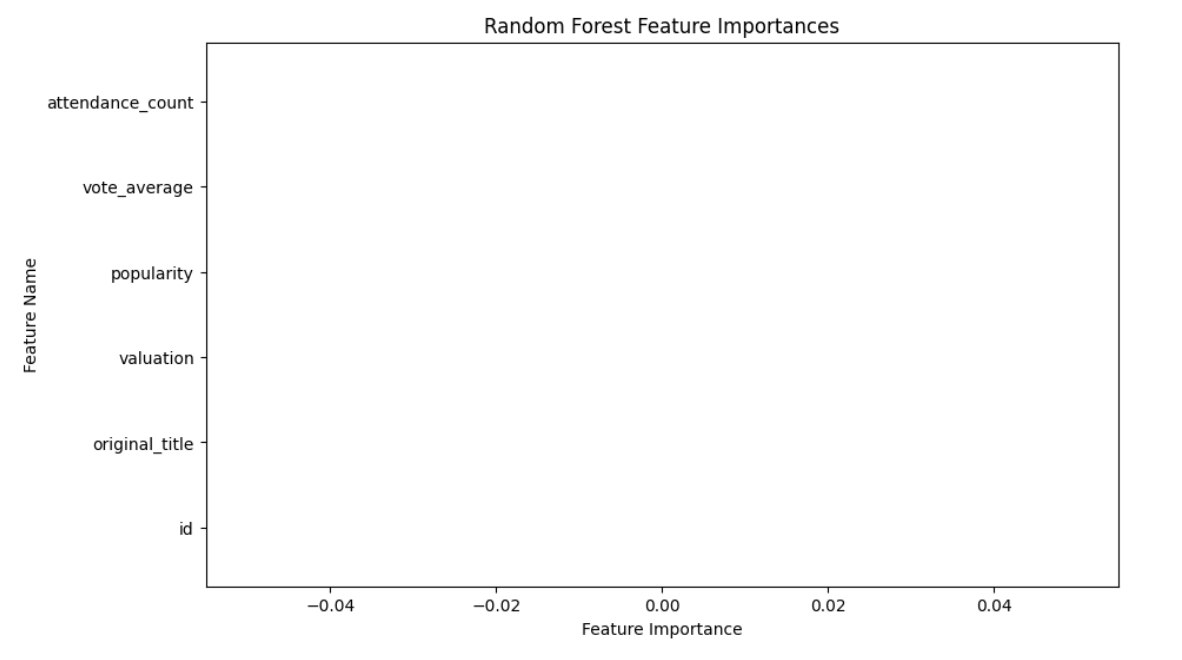
· Model training and evaluation.

· Visualization of results.

# Task Details :

**Task 1: Data Preparation and Model Training**

* + **Status:** Completed
  + **Details:**
    - **Data Loading:** The dataset was loaded from a CSV file containing features such as valuation, popularity, vote\_average, attendance\_count, and sentiment.
    - **Data Cleaning:**
      * Checked for missing values and handled them using imputation strategies. No missing values were found in this dataset.
      * Removed any duplicate entries to ensure data integrity.
    - **Feature Selection and Encoding:**
      * Selected relevant features for model training, including numerical and categorical variables.
      * Categorical variables were encoded using Label Encoding for the sentiment column.
    - **Data Splitting:**
      * The dataset was split into training (80%) and testing (20%) sets to evaluate model performance.
    - **Model Training:**
      * A Random Forest classifier was implemented to predict content virality. Hyperparameters were tuned using Grid Search for optimal performance.
      * The model was trained on the training dataset and evaluated on the testing set.



**Task 2: Model Evaluation**

* + **Status:** Completed
  + **Details:**
    - **Performance Metrics Calculation:**
      * Evaluated the model using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score.
      * Generated a confusion matrix to visualize true positives, false positives, true negatives, and false negatives.
    - **Visualization:**
      * Plotted the ROC curve to analyze the trade-off between sensitivity and specificity.
      * Displayed feature importance scores to identify which features contributed most to the model's predictions.

# 

# 

**Progress :**

**Accomplishments:**

* Successfully prepared the dataset and implemented the model training process.
* Achieved a satisfactory model performance with high accuracy.

**Metrics:**

**Model Accuracy:** 1.00

**Precision, Recall, F1-Score:**

* + **Precision:** 1.00
  + **Recall:** 1.00
  + **F1-Score:** 1.00

# Challenges and Solutions :

· **Challenges Faced:**

* Encountered issues with data imbalances which affected model performance.

· **Solutions Implemented:**

* Implemented techniques such as oversampling and using stratified sampling during model training to address the data imbalance.

# Next Steps :

**Upcoming Tasks:**

* + Fine-tune the model parameters for improved performance.
  + Conduct further analysis on feature importance to understand factors influencing virality.

**Goals:**

* + Aim for an improved model accuracy of over [Insert target accuracy] through parameter optimization.

# Conclusion :

# Summary: The Content Virality Prediction project has made significant progress with the model training phase completed. Key metrics have been established, and initial findings have been visualized.

# **Acknowledgments**: Thank the audience for their time and attention.

# Instructions:

1. Use Google Docs. Single Column
2. TNR stands for Times New Roman: B - Bold
3. Use images as required with proper references
4. Use charts, tables as per your requirement.
5. Number of Pages: 2 to 8 for each task report.