**Tik tok dataset**

**Exploratory Data Analysis (EDA):**

* Analyze the distribution of claim\_status (this is likely your target variable)
* Compare claim\_status across different verified\_status and author\_ban\_status categories
* What is the distribution of video durations, and is there a typical length for videos on the platform?
* How do the engagement metrics (views, likes, shares, downloads, comments) correlate with each other? Are there any strong relationships?
* What percentage of videos in the dataset are from verified users, and how does this compare to the percentage of videos containing claims?
* What is the average engagement rate (e.g., likes per view) across all videos, and how does this vary between videos with and without claims?
* Is there a relationship between the length of the video transcription text and the number of views or other engagement metrics?
* Calculate engagement rates (e.g., likes/views, comments/views)
* Analyze the relationship between verified\_status and claim\_status
* Investigate if author\_ban\_status correlates with likelihood of making claims
* Examine if video duration correlates with likelihood of containing a claim
* Create a correlation matrix to identify relationships between numeric variables

**Data Preprocessing**

**Handling missing values**

* Identify missing values (NaN) in the dataset.
* Decide on a strategy to handle them (e.g., filling with mean/median/mode, forward/backward filling, or dropping rows/columns).

**Duplicates**

* Check for and remove any duplicate rows to ensure that each entry is unique.

**Outliers**

* Identify and handle outliers in numerical columns like views, likes, comments, etc., using methods such as the IQR method or Z-score.

**Binarize Categorical Variables**

* Convert categorical variables like verified\_status, author\_ban\_status, etc., into binary or dummy variables if needed for modeling.

**Data Transformation**

* Prepare data for machine learning (encoding categorical variables, scaling numeric features)

**Modeling**

* Split data into training and testing sets
* Try different algorithms (e.g., Logistic Regression, Random Forest, Gradient Boosting)
* Evaluate models using appropriate metrics (accuracy, precision, recall, F1-score)