**Machine Learning Project Proposal**

**Group Members**

Tehreem Fatima - BSDSF22A014

Hamna - BSDSF22A031

Muqadsa Qudoos - BSDSF22A039

**Project Title**

YouTube Video Popularity Predictor using Video Metadata & NLP

**Problem Statement**

In a world flooded with video content, content creators and marketers often struggle to predict which videos will perform well. This project aims to use machine learning to predict a video's potential popularity (views/likes) based on metadata like title, description, tags, and video duration before it's even uploaded.

**Objectives**

* Use regression/classification to estimate a video’s popularity score.
* Apply NLP techniques to extract features from textual metadata (title, description, tags).
* Identify which metadata features contribute most to performance.

**Proposed Methodology**

* **Data Collection**: Kaggle dataset (YouTube Trending Video dataset)
* **Data Preprocessing**:
* Clean and tokenize text fields.
* Encode categorical variables (category, channel).
* Normalize numerical fields (duration, views, likes).
* **Feature Engineering**:
* TF-IDF or Word2Vec for text.
* Category and length as input features.
* **Modeling**:
* Regression (if predicting views/likes).
* Classification (e.g., high/medium/low popularity).
* Try models like Random Forest, XGBoost, or Logistic Regression.
* **Evaluation**: Use MAE/RMSE (regression) or accuracy/F1 (classification).

**Dataset**

* **Kaggle:** YouTube Trending Videos Dataset
* **Features include:** title, views, likes, tags, category, publish time etc.

**Machine Learning Project Proposal**

**Expected Outcomes**

* A model that can predict whether a video will perform well.
* Insights into which metadata elements affect performance the most.
* Visualization of popularity clusters based on features.