**Duplicate Data Detection using Machine Learning**

Abstract:

This project is designed to detect duplicate or nearly identical records in large datasets using Machine Learning techniques. It utilizes TF-IDF vectorization and cosine similarity to measure the textual similarity between different records. The system identifies redundant entries, which helps in data cleaning, deduplication, and improving data quality.

The project is particularly useful in industries dealing with customer records, financial transactions, and identity management, where duplicate data can lead to data inconsistencies, errors, and redundancy in decision-making. By eliminating duplicate entries, organizations can ensure data integrity, enhance efficiency, and optimize storage.

Prerequisites:

* Python (for scripting)
* Pandas (for data manipulation)
* Scikit-learn (for feature extraction and similarity measurement)
* TfidfVectorizer (to convert text into numerical features)
* Cosine Similarity (to measure similarity between records)
* CSV File Handling (for loading and saving datasets)