**Project Title:**

**“Advanced Drug Review Analysis: Predicting Ratings and Understanding Side Effects”**

**Objective:**

To analyze a comprehensive drug review dataset to:

* Understand drug effectiveness through rating distribution.
* Identify top drugs for common medical conditions.
* Analyze frequently reported side effects.
* Predict rating categories using machine learning models.
* Discover patterns using clustering techniques.

**Dataset Overview:**

* **Total records**: [Enter total count]
* **Key columns**:
  + drug\_name, medical\_condition, side\_effects, rating, no\_of\_reviews, pregnancy\_category, rating\_category

**Exploratory Data Analysis (EDA):**

**🔹 1. Drug Ratings Distribution**

* Most drug ratings fall in the 6–8 range.
* The histogram revealed a slight right-skew — indicating more favorable drug experiences.

Visual: Histogram of Drug Ratings

**🔹 2. Top Drugs for Common Conditions**

* Used bar charts to identify most frequently prescribed drugs for:
  + **Hypertension**
  + **Diabetes**
  + **Pain Relief**

Visual: Bar Chart of Top Drugs for Each Condition

**🔹 3. Top 10 Side Effects**

Most reported side effects:

* Nausea
* Headache
* Dizziness
* Fatigue
* Diarrhea

*Visual: Bar Chart or Value Counts*

**Machine Learning – Rating Prediction**

**Objective:**

To predict if a drug’s rating falls into **Low**, **Medium**, or **High** category based on attributes like condition, side effects, and pregnancy category.

**Method:**

* Used **One-Hot Encoding** for categorical variables.
* Classification Models:
  + Logistic Regression
  + Random Forest
  + Support Vector Machine (optional)
* Evaluation Metric: Accuracy & Classification Report

**Results:**

* Best Model: Random Forest
* Accuracy: ~XX%
* Insights:
  + Pregnancy category and side effects showed strong correlation with high/low ratings.

Visual: Confusion Matrix and Classification Report Table

**Unsupervised Learning – Clustering**

**Objective:**

To identify drug groups based on similarity in:

* Ratings
* Number of Reviews

**Algorithm:**

* K-Means Clustering
* Scaled features: rating, no\_of\_reviews

**Output:**

* Identified 3 clusters:
  + Cluster 0: High-rated drugs with many reviews.
  + Cluster 1: Medium-rated drugs with moderate reviews.
  + Cluster 2: Low-rated or niche drugs.

Visual: Scatter Plot with K-Means Clusters

**Conclusion:**

* **Most drugs receive favorable reviews**, but many lie in the moderate satisfaction range.
* **Top conditions like hypertension and diabetes have diverse drug options**.
* **Nausea and headache are the most common side effects**.
* **Machine learning can effectively predict drug rating categories** using side effects and condition data.
* **Clustering uncovered hidden drug groupings** based on review activity.

**Future Enhancements:**

* Include **sentiment analysis** on review text.
* Add **dosage and duration** fields to improve prediction accuracy.
* Integrate real-time drug review scraping for live model updates.