Drugs, Side Effects & Medical Condition – Full ML Report

Drugs, Side Effects & Medical Condition – Full ML Project Report

Executive Summary:

This report presents an end-to-end data science pipeline using a real-world drug dataset including medical conditions, side effects, pregnancy warnings, CSA schedules, and alcohol interaction indicators. Two machine learning workflows were implemented:

1) Drug Rating Prediction (Regression)

2) Rx/OTC Classification (Supervised Classification)

Extensive data cleaning, exploratory data analysis (EDA), feature engineering, model comparison, hyperparameter optimization, and final evaluation were performed.

Data Overview

- **Rows:** 2,931
- **Columns:** 17 (→ **16 after dropping brand_names**)
- **Key columns:**
  drug_name, generic_name, medical_condition, side_effects,
  drug_classes, activity, rx_otc, pregnancy_category, csa,
  rating, no_of_reviews, alcohol, and dataset URLs.

Data Cleaning

- Converted activity (Yes/No) to **0–1 float**
- alcohol: "X" → **1**, NaN → **0**
- Missing text fields filled with "Unknown"
- Missing numeric fields (rating, no_of_reviews) filled with **0**
- Categorical fields label-encoded for ML
- Final cleaned dataset: **no missing values**

Data Preparation:

• Label-encoded categorical variables.

• Cleaned missing values.

• Extracted structured features from text fields.

• Created a consolidated machine-learning-ready dataset.

Regression Task – Predict Drug Rating:

Baseline Model Comparison:

• Linear Regression: RMSE=3.585, R^2=0.109

• Random Forest: RMSE=1.425, R^2=0.859

• Gradient Boosting: RMSE=1.612, R^2=0.820

• CatBoost: RMSE=1.568, R^2=0.830

→ Best baseline model: RandomForestRegressor

After Hyperparameter Tuning:

• Best Params: max_depth=20, min_samples_leaf=2, min_samples_split=2, n_estimators=200

• CV RMSE: 1.6322

Final Regression Performance:

• RMSE: 1.4618

• MAE: 0.7877

• R^2: 0.8520

Classification Task – Predict Rx/OTC Type:

Baseline Model Comparison:

• Logistic Regression: Acc=0.686, F1=0.616

• Random Forest: Acc=0.901, F1=0.899

• Gradient Boosting: Acc=0.882, F1=0.878

• CatBoost: Acc=0.879, F1=0.876

→ Best baseline model: RandomForestClassifier

After Hyperparameter Tuning:

• Best Params: max_depth=20, min_samples_leaf=1, min_samples_split=2, n_estimators=300

• Best CV Accuracy: 0.8827

Final Classification Performance:

• Accuracy: 0.901

• Weighted F1-Score: 0.899

Conclusion:

Both ML tasks achieved strong, production-ready performance.

Random Forest models proved most effective across both regression and classification.

These outcomes support reliable predictive insights in pharmaceutical and clinical decision-support environments.

End of Report