CAPSTONE PROJECT
Drug Satisfaction Prediction App

Abstract
This dataset represents a meticulously curated collection of pharmaceutical data, which aims to transcend the confines of simplistic analysis by providing a multifaceted view of drug performance metrics across 37 prevalent health conditions. It comprises 685 records, each detailing an exhaustive array of parameters including the drug name, its form, the condition it treats, and its designation as a prescription medication. Notably, the dataset is enriched with both qualitative and quantitative consumer feedback, incorporating ease of use and effectiveness ratings, as well as satisfaction scores and the number of reviews, each measured on a scale from 1 to 5.

Use Case
The primary use case of this dataset extends into the development of sophisticated predictive models that endeavor to estimate drug satisfaction levels, an endeavor that has profound implications for patient adherence, healthcare outcomes, and market positioning. The predictive pursuit is not merely academic but is structured to emulate an industry-like scenario where such models can inform strategic business decisions, patient-centric drug development, and targeted marketing approaches.

A key facet of the dataset's complexity lies in its amalgamation of subjective consumer opinions with objective data such as drug type and pricing. The dataset's depth is exemplified by its nuanced inclusion of both "On Label" and "Off Label" indications, various drug formulations, and a wide price range, reflecting the real-world complexities of pharmaceutical usage and healthcare economics. The variability and distribution of reviews highlight significant disparities in consumer engagement, adding another layer to the analytical challenge.

Dependencies
Streamlit, Pandas, NumPy, Matplotlib, Seaborn, joblib

Observation
The dataset, as downloaded from Kaggle, has been preprocessed and cleaned, ready for in-depth analysis and model development. Here are detailed observations based on the provided information:

1. Completeness of Data:
   - The dataset is fully populated with no missing values across all features. This indicates a meticulous preprocessing phase where either missing data was imputed or incomplete records were removed, ensuring a smooth workflow for subsequent data analysis tasks.

2. Data Types and Structures:
   - The dataset consists of a mix of categorical (`object`) and continuous (`float64`) variables, suitable for regression analysis, classification, and clustering.
   - Variables such as `Condition`, `Drug`, `Form`, `Indication`, and `Type` are categorical and can be used for grouping and stratification in analysis.
   - Numeric variables like `EaseOfUse`, `Effective`, `Price`, `Reviews`, and `Satisfaction` are floating-point numbers, which indicates that these metrics were likely derived from averages or calculations that allow for more nuanced interpretations than integers would.

3. Ease of Use:
   - The `EaseOfUse` variable has a mean of 3.92 and a standard deviation of 0.89, indicating that most drugs are rated fairly high on ease of use but with some variation, reflecting different user experiences.

4. Effectiveness:
   - The `Effective` score has a mean of 3.52 and a standard deviation of 0.95, suggesting a moderate perception of drug effectiveness with a slightly wider spread in ratings than `EaseOfUse`, pointing towards varied efficacy perceptions among consumers.

5. Pricing Data:
   - The `Price` variable has a large mean of $174.21 and an even larger standard deviation of $667.74, showing significant price variability, which could be due to different drug brands, dosages, or treatment courses. The max price of $10,362.19 suggests the presence of some very high-cost drugs, potentially specialty or orphan drugs.

6. Review Count:
   - The review counts vary widely (mean of 82.64, standard deviation of 273.28), indicating that some drugs are much more frequently reviewed than others. This could be indicative of market share, consumer preference, or prescription frequency.

7. Satisfaction Ratings:

- Satisfaction scores have a mean of 3.19 with a standard deviation of 1.03. This variable is critical for predictive modeling purposes, as it may be influenced by other factors such as effectiveness, price, and ease of use.

8. Distribution of Ratings:
- The distribution of `EaseOfUse`, `Effective`, and `Satisfaction` ratings show that the median values are close to or above the mean values, which may suggest a skewed distribution with outliers affecting the mean.

9. Range of Data:
- The minimum values for `EaseOfUse`, `Effective`, and `Satisfaction` are all 1.0, which represents the lowest rating score, while the maximum is 5.0, the highest possible score, demonstrating a full range of customer feedback.

10. Quantiles:
- The 25th, 50th (median), and 75th percentiles provide additional insights into the distribution. For example, the median price is significantly lower than the mean price, which indicates that the data is right-skewed with a few expensive drugs driving up the average.

11. Form and Indication:
- The `Form` and `Indication` columns suggest a diversity in the modes of drug delivery and the conditions they are prescribed for, adding another dimension to the analysis.

Given the cleaned and preprocessed state of the dataset, it is primed for exploratory data analysis to uncover patterns and relationships. The data can also be leveraged to build predictive models with the goal of anticipating drug satisfaction scores based on other attributes, which is a valuable asset for both healthcare providers and pharmaceutical companies aiming to understand and enhance patient experiences.

Training Model

I wrote a python code that is a sequence of steps that constitute a machine learning pipeline using a random forest regressor. The purpose of the pipeline is to predict the 'Satisfaction' of drugs based on various features in the dataset. A random forest regressor is chosen for its ability to handle complex datasets with nonlinear relationships without requiring extensive feature engineering. It's also robust to overfitting, especially with datasets having numerous features. The use of a pipeline encapsulates the preprocessing and modeling steps, ensuring consistency in the transformations applied to both training and test datasets, which is essential for

producing reliable predictions. OneHotEncoder is used for categorical variables since most machine learning models require numerical input. This method is preferable over label encoding for categorical variables that do not have an ordinal relationship. The dataset is split into training and test sets to validate the model's performance on unseen data, which is crucial for assessing its generalizability.

Results of Trained Model

The MSE value of approximately 0.5704 is relatively low, suggesting that the model's predictions are, on average, within 0.5704 satisfaction points of the actual satisfaction scores in the test set. Given that satisfaction is likely on a scale of 1 to 5, this error might be considered reasonably small, indicating a model that performs well.

Streamlit Code

Next, I wrote a Python script for a Streamlit web application designed to predict drug satisfaction scores based on user input and visualize the distribution of satisfaction scores in the dataset.

1. Predictive Functionality:
   The script loads a pre-trained machine learning model (my_model.pkl) using joblib. It creates a user interface with Streamlit, allowing users to input various features related to a drug, such as condition, drug name, form, indication, type, ease of use, effectiveness, price, and reviews. Upon clicking the "Predict Satisfaction" button, the input features are passed through the pre-trained model to predict the satisfaction score for the specified drug.
2. Data Visualization:
   The script also includes a section for data analysis, specifically visualizing the distribution of satisfaction scores in the dataset. It uses Matplotlib and Seaborn to create a histogram of satisfaction scores, providing users with insights into the distribution and spread of satisfaction ratings among drugs in the dataset.
3. Inferences:
   The inclusion of a machine learning model allows users to obtain predictions on drug satisfaction without requiring knowledge of machine learning or coding. The data visualization component enhances user understanding of the dataset by presenting a visual representation of the distribution of satisfaction scores.
4. Functionality:
   Users can select a drug's condition, name, form, indication, type, and provide ratings for ease of use, effectiveness, price, and number of reviews. Upon submission, the app predicts the satisfaction score for the specified drug using a pre-trained machine learning model. Additionally, users can explore the distribution of satisfaction scores through a histogram visualization.

References

https://www.kaggle.com/datasets/thedevastator/drug-performance-evaluation?select=Drug_clean.csv
https://www.nature.com/articles/s41598-023-31694-6
https://www.degruyter.com/document/doi/10.1515/nanoph-2017-0001/html
https://github.com/alejandro-ao/streamlit-cancer-predict/blob/main/app/main.py

License
