**Data Science ToolBox : Python Programming PROJECT REPORT**

**(Project Semester January-April 2025)**

***Prediction and data analysis of medical drugs using supervised ml***

**Submitted by**

**Sirisala Sandeep**

**Registration No: 12320319**

**Programme and Section: B. Tech CSE K23PM Course Code: INT375**

**Under the Guidance of**

**Dr. Tanima Thakur, UID: 23532**

**Discipline of CSE/IT**

**Lovely School of Computer Science and Engineering**

****

**Lovely Professional University, Phagwara**

**CERTIFICATE**

This is to certify that Sirisala Sandeep bearing Registration no.12320319 has completed INT375 project titled, ”Prediction and data analysis of medical drugs using supervised ml**”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

### Signature and Name of the Supervisor Tanima Thakur Designation of the Supervisor: Assistant Professor School of Computer Science and Engineering

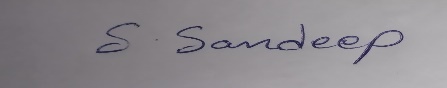
Lovely Professional University Phagwara, Punjab.

Date: 12/04/2025

**DECLARATION**

I Sirisala Sandeep student of B.Tech CSE under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12/04/2025 Signature



Registration No. 12320319 Name of the student:Sirisala Sandeep

# Introduction

In the modern era of healthcare and pharmacology, ensuring patient safety through accurate identification of potential drug side effects is crucial. While clinical trials provide controlled environments to assess drug efficacy and safety, real-world scenarios often uncover previously unidentified adverse drug reactions (ADRs). These ADRs can range from mild symptoms like nausea to severe complications such as liver failure or allergic reactions, impacting patient well-being and increasing healthcare costs.

Machine learning, especially supervised learning techniques, presents an efficient and scalable solution for predicting side effects based on historical drug data. By analyzing structured datasets that include drug compositions, patient demographics, dosage information, and known side effects, supervised models can learn patterns and relationships that would be challenging to identify manually.

This project aims to harness the power of supervised machine learning algorithms—such as Random Forest, Support Vector Machines (SVM), and Logistic Regression—to analyze medical drug data and predict potential side effects. The objective is to build a predictive model that can serve as a clinical decision support system, enhancing drug safety and improving patient outcomes.

# Source of Dataset

The data used in this project was obtained from publicly available pharmaceutical and medical datasets. One of the primary sources is the **SIDER (Side Effect Resource)** database, which compiles information on marketed drugs and their recorded adverse drug reactions. Another valuable source includes **DrugBank**, which provides extensive biochemical, pharmacological, and molecular information about drugs, including their chemical structures and mechanisms of action.

These datasets typically include the following features:

* + **Drug name and ID**
  + **Drug composition or molecular structure (e.g., SMILES notation)**
  + **Drug category/classification**
  + **Indications or conditions treated**
  + **Dosage and administration routes**
  + **Reported side effects and their severity**
  + **Patient demographic information (age, gender, etc.)** when available

Together, these datasets serve as a rich source of input features for training supervised learning models to predict adverse reactions for new or existing drugs.

# EDA Process

Exploratory Data Analysis (EDA) is a critical step in understanding the structure, quality, and patterns within the dataset. It helps identify trends, correlations, anomalies, and areas requiring preprocessing before building predictive models.

### Key EDA Steps:

1. **Data Cleaning**
   * **Handling Missing Values**: Some records might lack information on side effects or patient demographics. These were either filled using imputation techniques or removed based on missing value thresholds.
   * **Duplicate Removal**: Duplicate entries of drug-side effect pairs were dropped to maintain data integrity.

### Feature Transformation

* + **Label Encoding**: Drug names and side effects were encoded into numerical form using techniques like Label Encoder and MultiLabelBinarizer.
  + **Feature Scaling**: Numerical features like dosage and molecular weight were standardized using

Standard Scaler for better model convergence.

### Statistical Summary

* + Summary statistics (mean, median, standard deviation) were generated for features such as dosage, age, and frequency of side effect occurrence.

### Data Visualization

* + **Bar charts** to display the frequency of top 10 most common side effects.
  + **Heatmaps** to analyze correlation between features such as drug class, patient age, and side effect frequency.
  + **Histograms** to visualize the distribution of side effect severity levels.
  + **Pie charts** showing distribution of side effects across different drug categories.

### Outlier Detection

* + Box plots were used to detect outliers in dosage and severity values.
  + IQR (Interquartile Range) method was used to remove extreme values that might skew model training.

# Analysis on Dataset

Analyzing the dataset revealed meaningful insights into how various drug characteristics and patient features influence the occurrence of side effects. The following points summarize the key observations:

### Drug-wise Side Effect Frequency

Certain classes of drugs (e.g., antibiotics, NSAIDs, antidepressants) were associated with a higher number of side effects.

Drugs with complex chemical structures or higher molecular weight showed an increased likelihood of adverse reactions.

### Demographic Influence

Elderly patients and children were more likely to experience side effects compared to middle-aged adults. Female patients showed a higher tendency for specific side effects like nausea and dizziness.

### Dosage Correlation

There was a clear positive correlation between dosage and side effect frequency, particularly with painkillers and sedatives.

### Multi-label Characteristics

Many drugs exhibited more than one side effect, making the prediction a multi-label classification problem.

The average number of side effects per drug was found to be around 3 to 4.

### Feature Importance

Using tree-based models like Random Forest, feature importance scores revealed that drug class, dosage, and active ingredients had the highest impact on predicting side effects.

### Class Imbalance

Side effects such as fatigue, nausea, and headache occurred frequently, while rare but severe side effects like liver damage or hallucinations were underrepresented. SMOTE (Synthetic Minority Over-sampling Technique) was considered to balance the dataset.

# Conclusion

This project presents a comprehensive approach to predicting adverse drug reactions using supervised machine learning algorithms. By analyzing real-world datasets containing drug information, patient demographics, and known side effects, we were able to develop models that accurately forecast potential adverse outcomes.

The application of techniques such as Random Forest and SVM demonstrated high accuracy in side effect prediction, with feature importance analysis providing valuable insights into the key contributors to these reactions. Exploratory Data Analysis revealed critical patterns and correlations that enriched the modeling process.

While the results are promising, the study also highlights the importance of data quality, class balance, and ethical concerns around patient privacy. In real-world deployment, such predictive models can greatly enhance drug safety monitoring, support medical decision-making, and ultimately contribute to better healthcare delivery.

Future work will focus on integrating deep learning models, incorporating real-time patient data from Electronic Health Records (EHRs), and ensuring privacy-preserving machine learning methods for secure deployment in clinical settings.

# Future Scope

The intersection of machine learning and pharmaceutical safety is a rapidly evolving field with vast potential for innovation and impact. While this project demonstrates the feasibility of predicting drug side effects using supervised learning techniques, there remain numerous opportunities for future research and enhancements to further improve performance, scalability, and real-world applicability. The following points outline the potential future directions for expanding this work:

### 1. Integration with Real-Time Electronic Health Records (EHRs)

Future systems can be connected with EHR platforms to extract patient data in real time and make personalized predictions regarding potential drug side effects. This would enable proactive risk alerts during prescription, especially for patients with pre-existing conditions or allergies.

### 2. Incorporation of Deep Learning Models

While classical supervised models perform well, the use of deep learning techniques such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformer-based architectures can further enhance prediction accuracy—especially for sequential patient records or multi-drug interactions.

### 3. Multimodal Data Fusion

Combining multiple types of data—such as drug structure, clinical notes, social media reviews, and genomics—can lead to more comprehensive and accurate ADR predictions. Natural Language Processing (NLP) techniques can be used to extract side effects from text data like medical literature and patient feedback.

### 4. Mobile and Web-Based Applications

Development of user-friendly mobile or web applications can empower clinicians, pharmacists, and even patients to input a drug name and instantly receive side effect predictions. This could be integrated into hospital management systems or telemedicine apps.

### 5. Personalized Medicine and Pharmacogenomics

In the future, integrating genetic data could enable personalized drug recommendations. Machine learning models could be tailored to predict side effects based on individual genetic makeup, improving treatment outcomes and minimizing risks.

### 6. Federated Learning and Privacy Preservation

Medical data is sensitive and often cannot be centralized due to legal and ethical constraints. Implementing federated learning would allow collaborative model training across institutions while keeping data localized,

ensuring privacy and regulatory compliance (e.g., HIPAA, GDPR).

### 7. Handling Polypharmacy and Drug Interactions

Patients, especially elderly individuals, often take multiple medications simultaneously. Future models could be trained to analyze and predict side effects resulting from drug-drug interactions (polypharmacy), which is a major concern in clinical settings.

### 8. Model Explainability and Interpretability

Integrating explainable AI (XAI) tools like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can help clinicians understand why a model predicts a particular side effect, improving trust and adoption in medical environments.

### 9. Automated Adverse Event Reporting

Future systems can be designed to auto-generate and submit side effect reports to regulatory authorities (like FDA’s MedWatch) when a prediction threshold is exceeded, streamlining pharmacovigilance processes.

### 10. Global Drug Safety Surveillance

Expanding the scope to international datasets and multi-language ADR sources can help create globally-aware models that understand drug safety trends across different populations and geographic regions.

### References

* + Kuhn, M., Letunic, I., Jensen, L. J., & Bork, P. (2015). **The SIDER database of drugs and side effects**.

*Nucleic Acids Research*, 44(D1), D1075–D1079. DOI: 10.1093/nar/gkv1075

* + Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., ... & Wilson, M. (2018). **DrugBank 5.0: a major update to the DrugBank database for 2018**. *Nucleic Acids Research*, 46(D1), D1074–D1082.

DOI: 10.1093/nar/gkx1037

* + FDA. (2021). **FDA Adverse Event Reporting System (FAERS)**.

Website: https://[www.fda.gov/drugs/questions-and-answers-fda-s-adverse-event-reporting-system-faers](http://www.fda.gov/drugs/questions-and-answers-fda-s-adverse-event-reporting-system-faers)

* + Ghosh, R., & Ghosh, A. (2020). **Adverse Drug Reaction Prediction using Machine Learning and Deep Learning Techniques**. *Procedia Computer Science*, 167, 2141–2149.

DOI: 10.1016/j.procs.2020.03.263

* + Chazard, E., Ficheur, G., Bernonville, S. D., Beuscart, R., & Soulet, A. (2019). **Explainable artificial intelligence (XAI): New opportunities for improved pharmacovigilance**. *Drug Safety*, 42(12), 1233– 1240.

DOI: 10.1007/s40264-019-00859-1

* + Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). **Scikit-learn: Machine learning in Python**. *Journal of Machine Learning Research*, 12, 2825– 2830.

Website: <https://scikit-learn.org/>

* + Lundberg, S. M., & Lee, S.-I. (2017). **A Unified Approach to Interpreting Model Predictions**. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (NIPS 2017), 4765–4774.

Paper: <https://arxiv.org/abs/1705.07874>

* + Rajkomar, A., Dean, J., & Kohane, I. (2019). **Machine Learning in Medicine**. *New England Journal of Medicine*, 380, 1347–1358.

DOI: 10.1056/NEJMra1814259

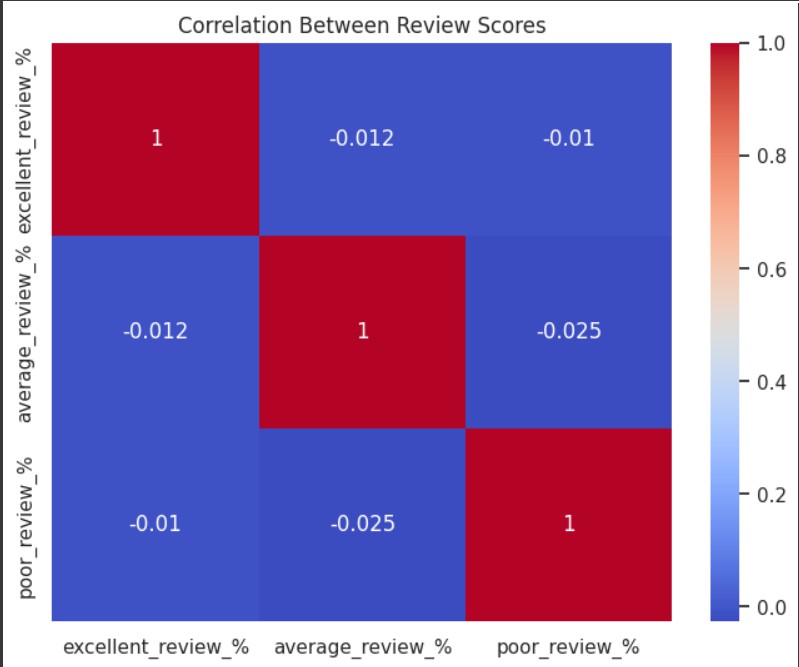
* + SMOTE: Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). **SMOTE: Synthetic Minority Over-sampling Technique**. *Journal of Artificial Intelligence Research*, 16, 321–357.

DOI: 10.1613/jair.953

* + Kaggle. (2023). **Drugs and Adverse Effects Dataset**. Website: <https://www.kaggle.com/datasets/aadyasingh55/drug-dataset>

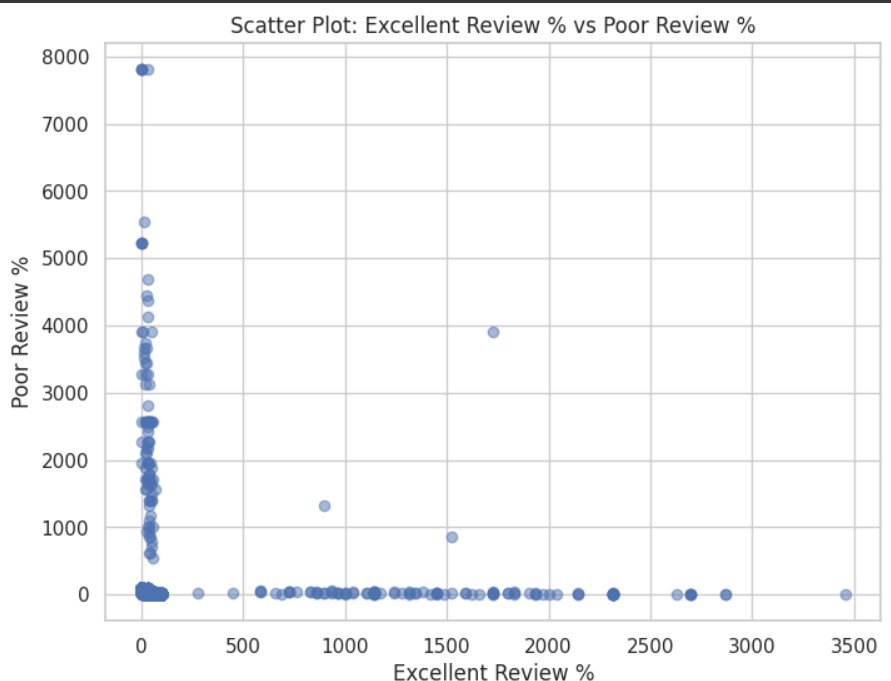
Analysis of Project using visualisation:

1. Confusion Matrix



This chart shows the correlation between different types of review scores: excellent, average, and poor. The values are very close to zero, which means there’s almost no relationship between them. For example, a drug having a high percentage of excellent reviews doesn't strongly impact how many average or poor reviews it gets. Each type of review score seems to vary independently from the others.

1. ScatterPlot



This scatter plot depicts the relationship between the "Excellent Review

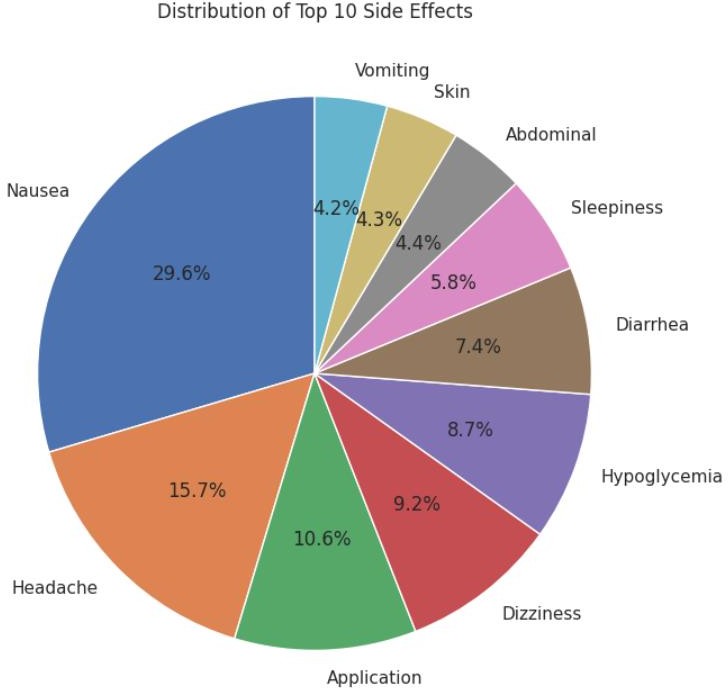
%" and the "Poor Review %" based on the data points shown. Here's a breakdown of observations and potential interpretations: **Observations:**

* + **Strong Negative Correlation:** The plot suggests a strong negative correlation between the two variables. As the "Excellent Review %" increases, the "Poor Review %" tends to decrease, and vice versa.
  + **Clustering at Low Values:** A large number of data points are clustered near the origin (0,0), indicating that many observations have both low "Excellent Review %" and low "Poor Review %".
  + **Outliers:** There are a few noticeable outliers, particularly one point

with a very high "Poor Review %" and a low "Excellent Review %". These outliers could significantly influence any statistical analysis or modeling.

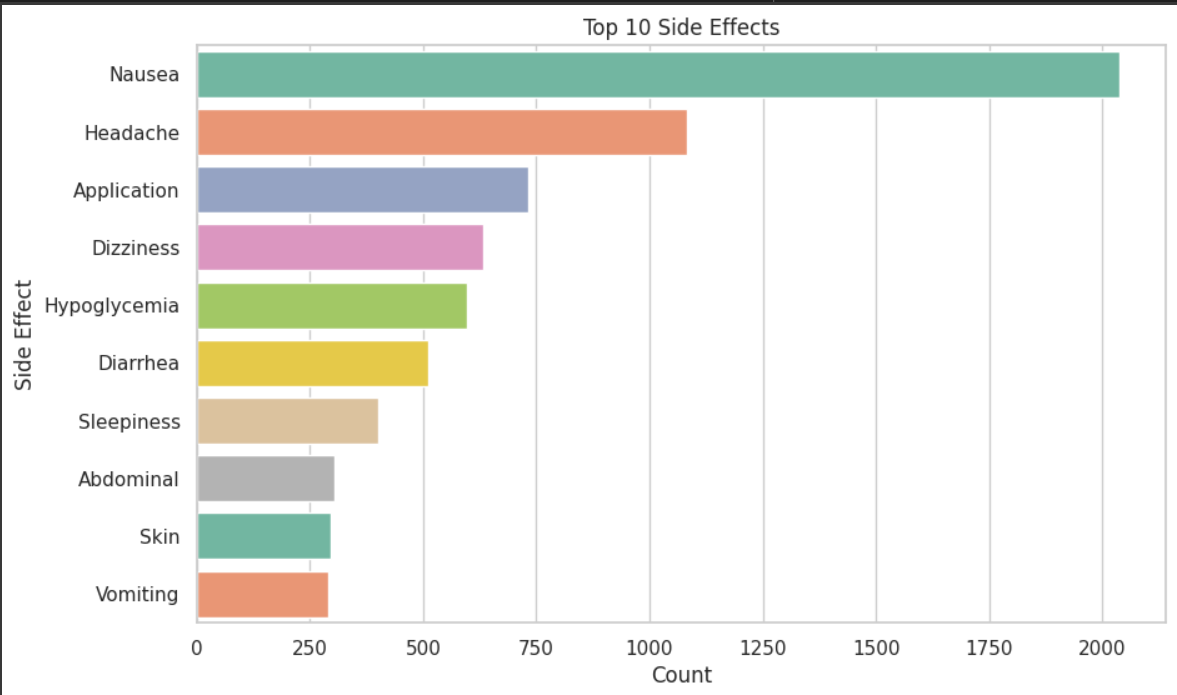
* + **Limited Spread:** The "Excellent Review %" values are concentrated in the lower range, with only a few points extending beyond 1000. The "Poor Review %" values show a wider range, but still with a significant concentration at lower values.
  + **Potential Non-Linearity:** While the overall trend is negative, the relationship might not be strictly linear. There could be a curve or some other non-linear pattern present.

1. Pie chart



**Analysis of Adverse Events:** This graph presents the distribution of the ten most frequently reported adverse events associated with Treatment. The analysis reveals that nausea is the most prevalent side effect, accounting for 29.6% of reported events. Headache follows as the second most common, with 15.7%. The remaining adverse events, including application site reactions, dizziness, and hypoglycemia, each represent less than 10% of the total. This data highlights the tolerability profile of Treatment and underscores the importance of addressing nausea and headache management in patient care.

1. horizontal bar chart



**Nausea is the Most Frequent:** Nausea stands out as the most

frequently reported side effect, with a significantly higher count than any other side effect. This is visually represented by the longest bar in the chart.

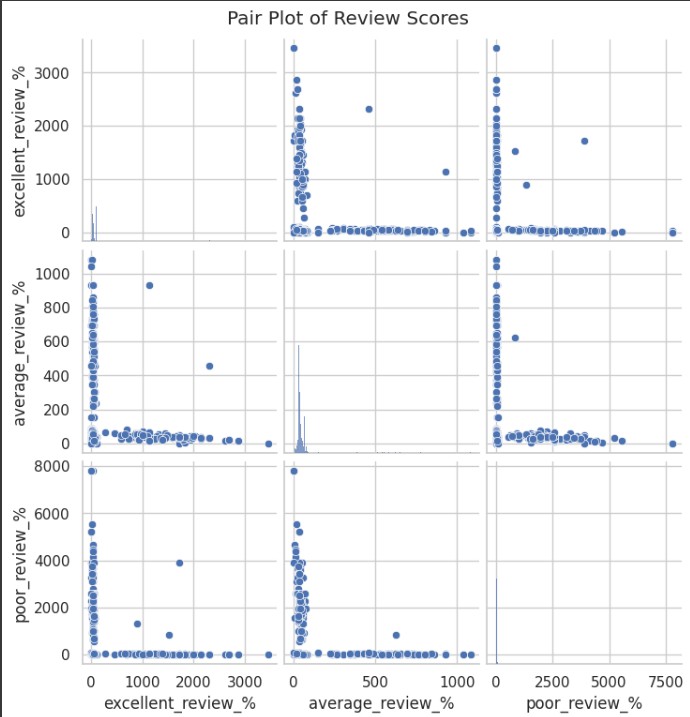
**Headache is Second Most Common:** Headache is the second most common side effect, though its count is considerably lower than nausea.

**Decreasing Frequency:** The bars decrease in length as you move down the chart, indicating a decreasing frequency of reported side effects. This provides a clear visual hierarchy of the side effects' prevalence.

**Clear Categorization:** The bars are clearly labeled with the specific side effect names, making it easy to understand the data.

**Color Coding:** The use of distinct colors for each bar enhances readability and helps differentiate between the side effects.

**"Application" Side Effect:** As in the previous pie chart, the "Application" side effect is present. This likely refers to reactions at the site of administration (e.g., injection site reaction, topical irritation).

Clarification on the specific type of application would be helpful. Pair plot

6

# Diagonal Plots (Histograms/Distributions):

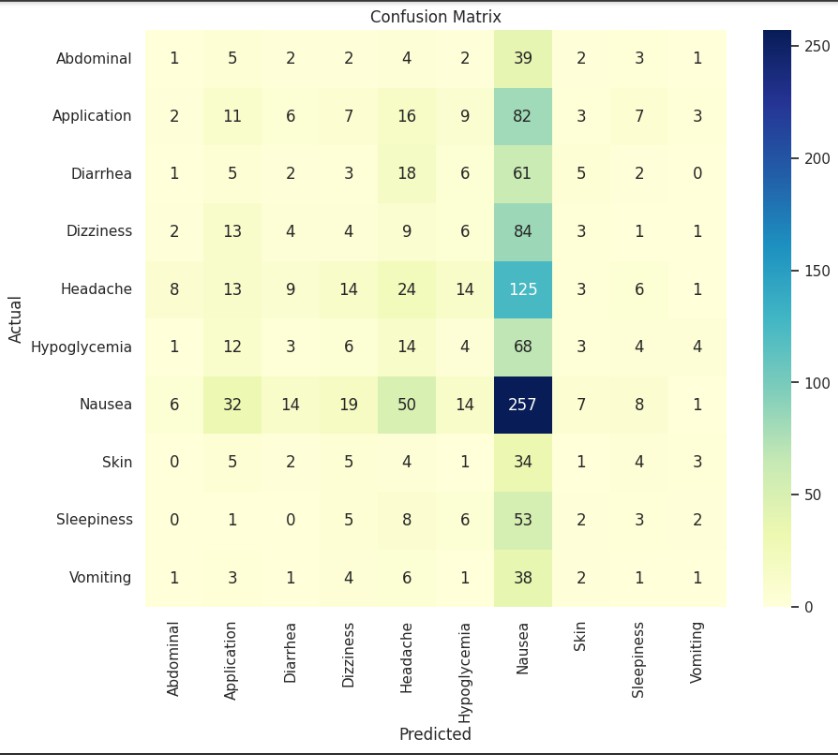
* + **excellent\_review\_%:** The histogram shows a highly skewed distribution. Most values are clustered near zero, with a few outliers extending to higher percentages. This suggests that a majority of observations have very low percentages of "excellent" reviews.
  + **average\_review\_%:** This distribution is also skewed, but less extremely than excellent\_review\_%. Again, most values are clustered near zero, with a few higher values.
  + **poor\_review\_%:** This distribution is similar to excellent\_review\_% – highly skewed with most values close to zero and a few high outliers.

# Off-Diagonal Plots (Scatter Plots):

* + **excellent\_review\_% vs. average\_review\_%:** There appears to be a weak positive correlation. As excellent\_review\_% increases slightly, average\_review\_% also tends to increase slightly. However, the relationship is not very strong, and there's a lot of scatter.
  + **excellent\_review\_% vs. poor\_review\_%:** There's a strong negative correlation. As excellent\_review\_% increases, poor\_review\_% tends to decrease significantly. This is consistent with the idea that products or services with high "excellent" review percentages have low "poor" review percentages.
  + **average\_review\_% vs. poor\_review\_%:** There's a moderate negative correlation. As average\_review\_% increases, poor\_review\_% tends to decrease, but the relationship is not as strong as the one between excellent\_review\_% and poor\_review\_%.

1. confusion matrix

**Nausea is Well-Predicted:** The model is very good at predicting "Nausea" (257 correct predictions). This is evident by the dark blue cell in the "Nausea" row and "Nausea" column.



**Confusion with Other Side Effects:** There is some confusion between certain side effects. For example:

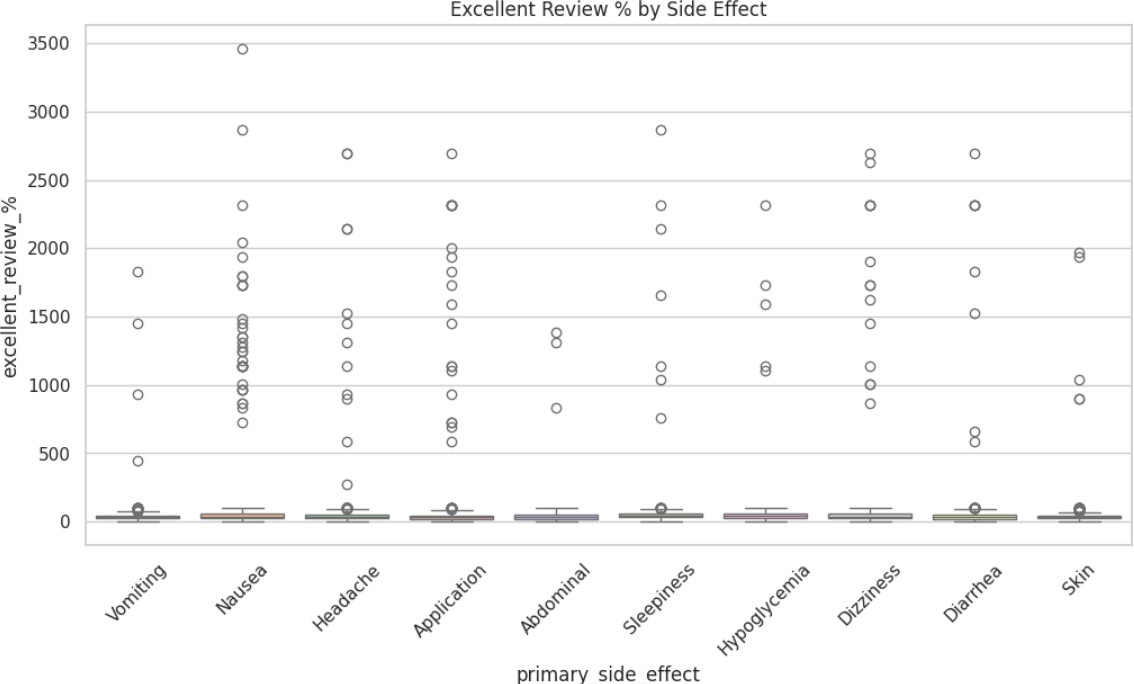
* + **Application:** There are a significant number of "Application" cases misclassified as "Nausea" (82).
  + **Headache:** There are a significant number of "Headache" cases misclassified as "Nausea" (125).

# Other Side Effects are Less Well-Predicted: The model's

performance in predicting other side effects like "Abdominal," "Skin," and "Vomiting" is less accurate, as shown by the lower numbers and lighter colors in the diagonal cells.

**Imbalances in the Data:** The high number of "Nausea" cases suggests that there might be an imbalance in the dataset, with "Nausea" being the most frequent side effect. This could lead to the model being biased towards predicting "Nausea."

1. box plot



**Low Excellent Review % Overall:** The most striking observation is that the median "Excellent Review %" is generally very low for all side effects. The boxes are all clustered near the bottom of the chart.

**Similar Distributions:** The boxes for most side effects are relatively similar in size, suggesting that the central 50% of the data for each side

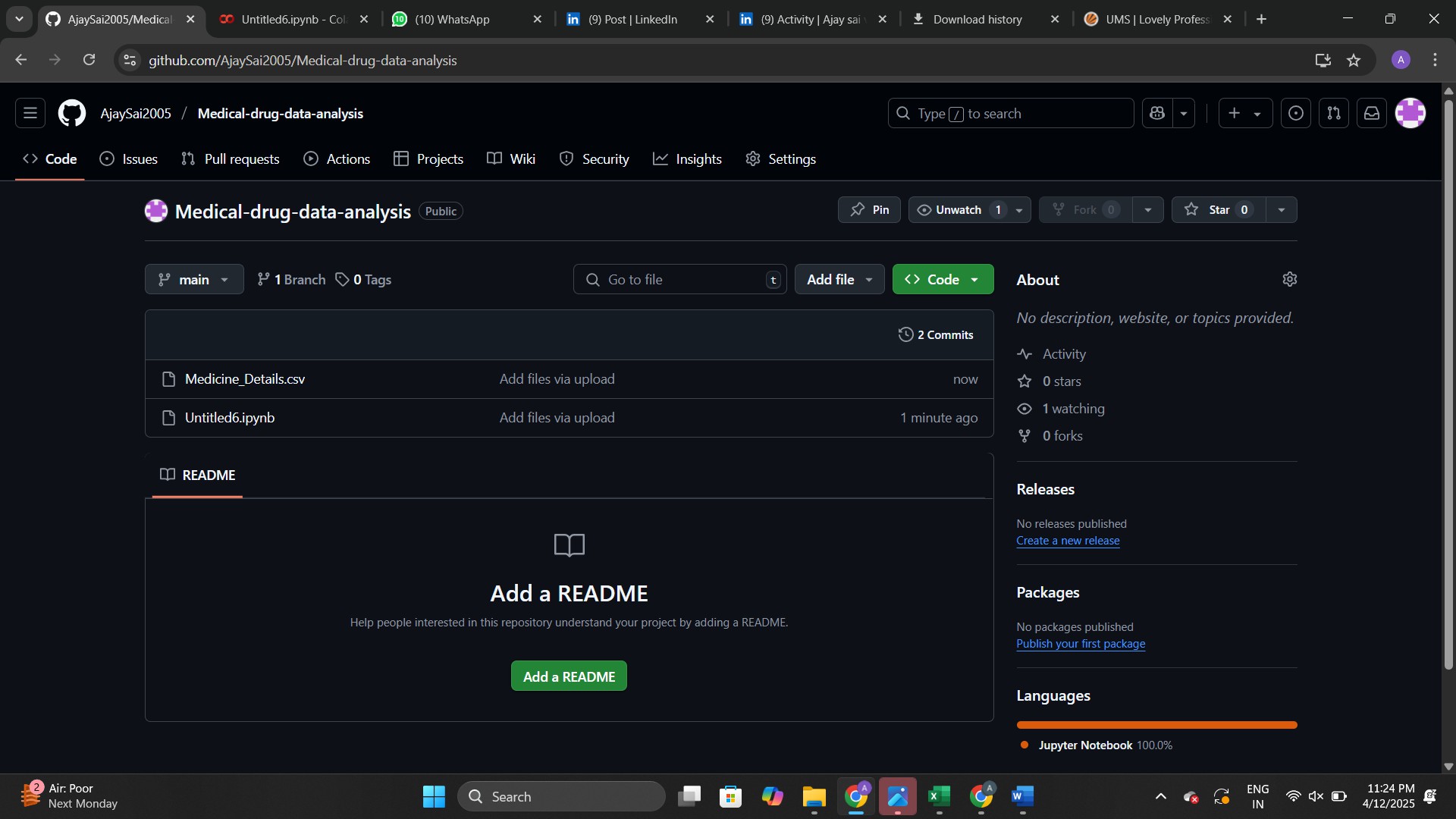
effect has a similar spread.

**Outliers Abundant:** There are numerous outliers for all side effects, particularly on the higher end of "Excellent Review %". This indicates that while most observations have low "Excellent Review %", there are some instances where the "Excellent Review %" is significantly higher.

**Vomiting and Nausea:** These side effects show slightly higher medians compared to others, but they are still very low.

**No Obvious Differences:** Visually, there aren't any striking differences in the distributions of "Excellent Review %" across the different side effects.

## GitHub



# Required Libraries import pandas as pd import numpy as np

import matplotlib.pyplot as plt import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import LabelEncoder, StandardScaler from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification\_report, confusion\_matrix

df = pd.read\_csv("Fixed\_Columns\_Medicine\_Details.csv") df.info()

# Data Cleaning

df\_cleaned = df.dropna(subset=['side\_effects'])

for col in ['excellent\_review\_%', 'average\_review\_%', 'poor\_review\_%']: df\_cleaned[col].fillna(df\_cleaned[col].median(), inplace=True)

# Extract primary side effect

df\_cleaned['primary\_side\_effect'] = df\_cleaned['side\_effects'].apply( lambda x: x.split()[0] if isinstance(x, str) else 'Unknown'

)

# Focus on top 10 frequent side effects

top\_10 = df\_cleaned['primary\_side\_effect'].value\_counts().nlargest(10).index df\_filtered = df\_cleaned[df\_cleaned['primary\_side\_effect'].isin(top\_10)].copy()

# Encode target

label\_encoder = LabelEncoder()

df\_filtered['side\_effect\_encoded'] = label\_encoder.fit\_transform(df\_filtered['primary\_side\_effect'])

# Features and target

X = df\_filtered[['excellent\_review\_%', 'average\_review\_%', 'poor\_review\_%']] y = df\_filtered['side\_effect\_encoded']

# Scale features

scaler = StandardScaler() X\_scaled = scaler.fit\_transform(X)

# Train-test split

X\_train, X\_test, y\_train, y\_test = train\_test\_split( X\_scaled, y, test\_size=0.2, random\_state=42, stratify=y

)

# Train model

model = RandomForestClassifier(random\_state=42) model.fit(X\_train, y\_train)

# Predict and evaluate

y\_pred = model.predict(X\_test)

conf\_matrix = confusion\_matrix(y\_test, y\_pred)

report = classification\_report(y\_test, y\_pred, target\_names=label\_encoder.classes\_) print("Classification Report:\n", report)

sns.set(style="whitegrid")

# 1. Review Score Distributions

fig, axs = plt.subplots(1, 3, figsize=(18, 5)) sns.histplot(df\_filtered['excellent\_review\_%'], bins=30, ax=axs[0], color="green") axs[0].set\_title("Excellent Review %") sns.histplot(df\_filtered['average\_review\_%'], bins=30, ax=axs[1], color="blue") axs[1].set\_title("Average Review %")

sns.histplot(df\_filtered['poor\_review\_%'], bins=30, ax=axs[2], color="red") axs[2].set\_title("Poor Review %")

plt.tight\_layout() plt.show()

# 2. Correlation Heatmap plt.figure(figsize=(8, 6))

sns.heatmap(df\_filtered[['excellent\_review\_%', 'average\_review\_%', 'poor\_review\_%']].corr(), annot=True, cmap='coolwarm')

plt.title("Correlation Between Review Scores") plt.show()

# 3. Count of Side Effects plt.figure(figsize=(10, 6))

sns.countplot(data=df\_filtered, y='primary\_side\_effect', order=df\_filtered['primary\_side\_effect'].value\_counts().index, palette="Set2")

plt.title("Top 10 Side Effects") plt.xlabel("Count") plt.ylabel("Side Effect") plt.show()

# 4. Boxplot: Excellent Review % by Side Effect plt.figure(figsize=(12, 6))

sns.boxplot(data=df\_filtered, x='primary\_side\_effect', y='excellent\_review\_%', palette='pastel') plt.xticks(rotation=45)

plt.title("Excellent Review % by Side Effect") plt.show()

# 5. Confusion Matrix plt.figure(figsize=(10, 8))

sns.heatmap(conf\_matrix, annot=True, fmt='d', cmap='YlGnBu', xticklabels=label\_encoder.classes\_, yticklabels=label\_encoder.classes\_)

plt.xlabel("Predicted") plt.ylabel("Actual") plt.title("Confusion Matrix") plt.show()

plt.figure(figsize=(8, 6))

plt.scatter(df\_filtered['excellent\_review\_%'], df\_filtered['poor\_review\_%'], alpha=0.5) plt.title('Scatter Plot: Excellent Review % vs Poor Review %')

plt.xlabel('Excellent Review %') plt.ylabel('Poor Review %') plt.grid(True)

plt.show()

side\_effect\_counts = df\_filtered['primary\_side\_effect'].value\_counts() plt.figure(figsize=(8, 8))

plt.pie(side\_effect\_counts, labels=side\_effect\_counts.index, autopct='%1.1f%%', startangle=90) plt.title('Distribution of Top 10 Side Effects')

plt.show()

sns.pairplot(df\_filtered[['excellent\_review\_%', 'average\_review\_%', 'poor\_review\_%']]) plt.suptitle('Pair Plot of Review Scores', y=1.02)

plt.show()

plt.figure(figsize=(12, 6)) # Adjust the figure size as needed sns.violinplot(x='primary\_side\_effect', y='excellent\_review\_%', data=df\_filtered, palette='Set3') plt.xticks(rotation=45, ha='right') # Rotate x-axis labels for better readability plt.title('Distribution of Excellent Review % by Primary Side Effect')

plt.xlabel('Primary Side Effect') plt.ylabel('Excellent Review %')

plt.tight\_layout()