Technical Report: FDA Adverse Event Prediction System on (FAERS) Data

Abstract

The need for an effective system to predict and mitigate adverse drug events is critical for enhancing medication safety and supporting regulatory decision-making. This project leverages the FDA's Adverse Event Reporting System (FAERS) dataset to develop a predictive model for adverse drug events. Preprocessing strategies like hierarchical missing value imputation, advanced feature engineering, TF-IDF vectorization for textual features, and statistical validation using Chi-Square and Kruskal-Wallis tests were performed . Multiple machine learning models, including Logistic Regression, Random Forest, and XGBoost, were trained. Random Forest Classifier achieved best performance giving an accuracy of 82 percent after hyperparameter tuning. Feature importance analysis further refined the predictive framework by identifying critical risk factors. This model offers significant potential for proactive safety monitoring and real-time risk assessment, paving the way for future research in identifying high-risk drug combinations and mitigating adverse medication outcomes.

Introduction

The pharmaceutical and healthcare industries rely critically on comprehensive safety monitoring systems to track and analyze potential adverse events associated with medications. The Food and Drug Administration (FDA) plays a pivotal role in this ecosystem through its Adverse Event Reporting System (FAERS), a sophisticated database that collects, manages, and analyzes reports of medication-related adverse events and errors. This research project aims to develop a comprehensive analytical approach to understanding and predicting patient outcomes based on the rich and complex dataset provided by FAERS.

Business Problem Statement

The fundamental challenge addressed in this research is extracting meaningful insights from the complex and multidimensional adverse event reporting data. With 9% of reported cases involving patient deaths, there is an urgent need to develop predictive models that can help healthcare professionals and regulatory bodies identify potential risk factors and improve patient safety. The project seeks to transform raw adverse event data into actionable intelligence by leveraging advanced data preprocessing, feature engineering, and machine learning techniques.

Dataset Description

The FAERS dataset is a comprehensive collection of adverse event reports compiled from multiple sources. The original dataset was structured across seven distinct tables: Demographic, Drugs, Indications, Outcome, Reaction, Report Source, and Therapy. These tables were merged using primary identifiers (PrimaryID and CaseID) to create a unified dataset for analysis.

The dataset has been included in the zip file. We chose to work on the Q3 data. Data can also be downloaded from the FDA website

The dataset encompasses a wide range of features, including:

- Patient Demographics: Age, sex, occupation, and country of occurrence

- Drug Information: Drug name, active ingredients, administration route, dosage, and frequency

- Reporting Details: Report date, submission mode, and manufacturer information

- Outcome Variables: Categorized outcomes such as death, hospitalization, disability, and other serious medical events

Data Preprocessing and Feature Engineering

The preprocessing approach was methodical and sophisticated, addressing multiple challenges in data preparation:

Missing Value Imputation

A hierarchical imputation strategy was employed to handle missing values. The approach prioritized filling missing values using measures of central tendency like median and mode of respective subgroup, and subsequently rows with nulls dropped and the entire case which the row belonged to was also dropped to maintain data integrity.

Feature Engineering Techniques

Several advanced feature engineering techniques were implemented:

1.Dose Information Extraction from dose_vbm column

A custom function was developed to extract and standardize medication dosage information. This included parsing dose amount, identifying units (mg, ml, tablet), and recognizing dosing frequencies using regex expressions.

2. Dosage Standardization

A comprehensive conversion dictionary was created to standardize medication amounts to milligrams, handling various units and case-insensitive variations. This ensured consistent quantitative analysis across different medication formulations.

3. Route Categorization

Administration routes were grouped into broader, more meaningful categories. A categorization function mapped specific routes to high-level categories like "Oral and Related Routes" or "Intravenous Routes", facilitating more insightful analysis.

4. Age and Date Processing

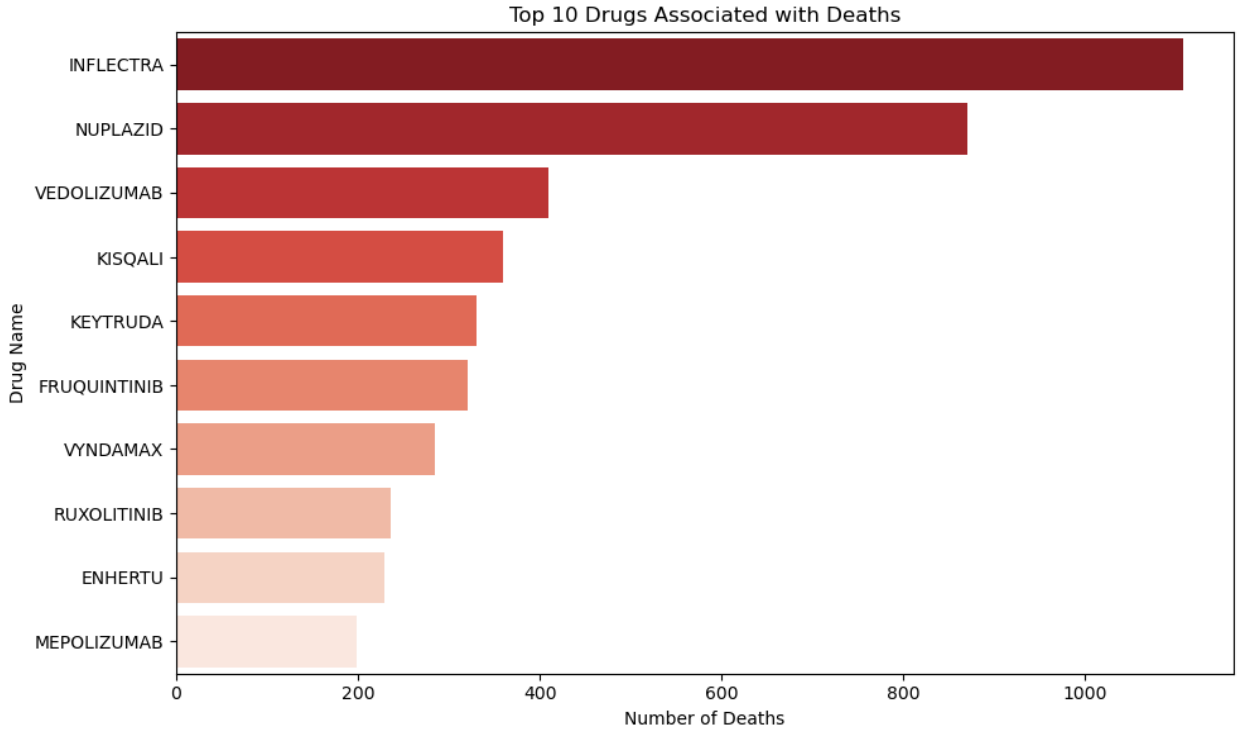
Age was mentioned in years, months, weeks, days. All of them were converted to years. Age outliers were handled by dividing extreme values by 365 or 52 based on the range of the value ensuring more reliable age-related analysis. Report dates were transformed into day and month columns to enable temporal insights.

5. Text Processing

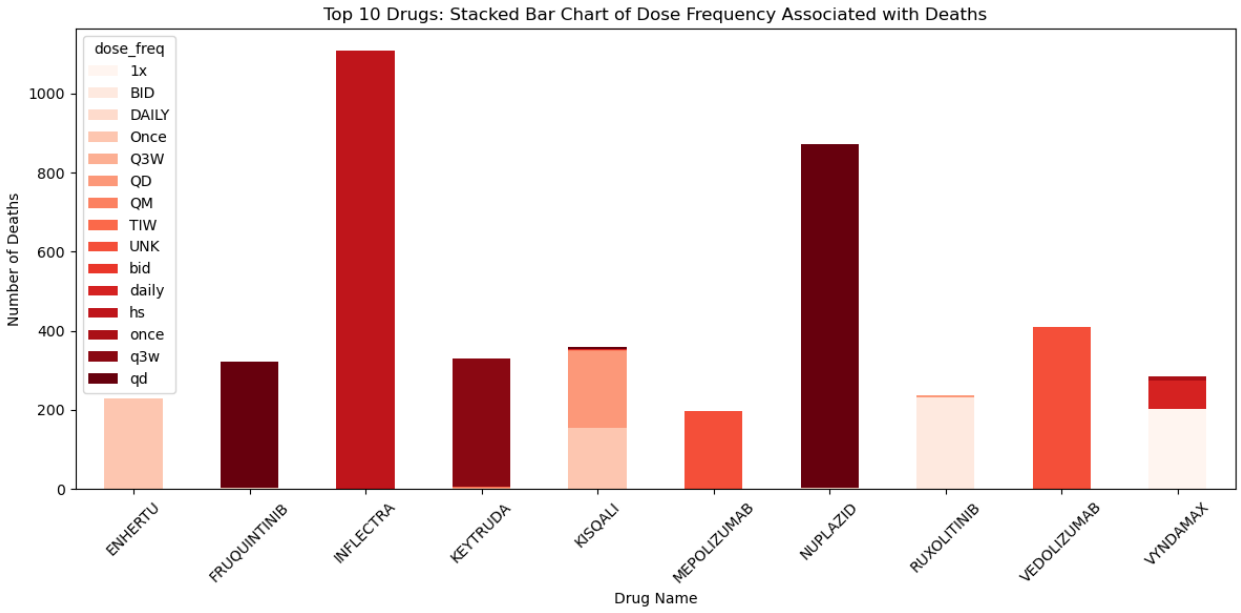
Preferred medical terms were converted into TF-IDF (Term Frequency-Inverse Document Frequency) vectors, enabling sophisticated text-based feature representation. This was performed to check the correlation using Kruskal-Wallis Test as chi-square test was not rejecting the null hypothesis.

Exploratory Data Analysis (EDA):

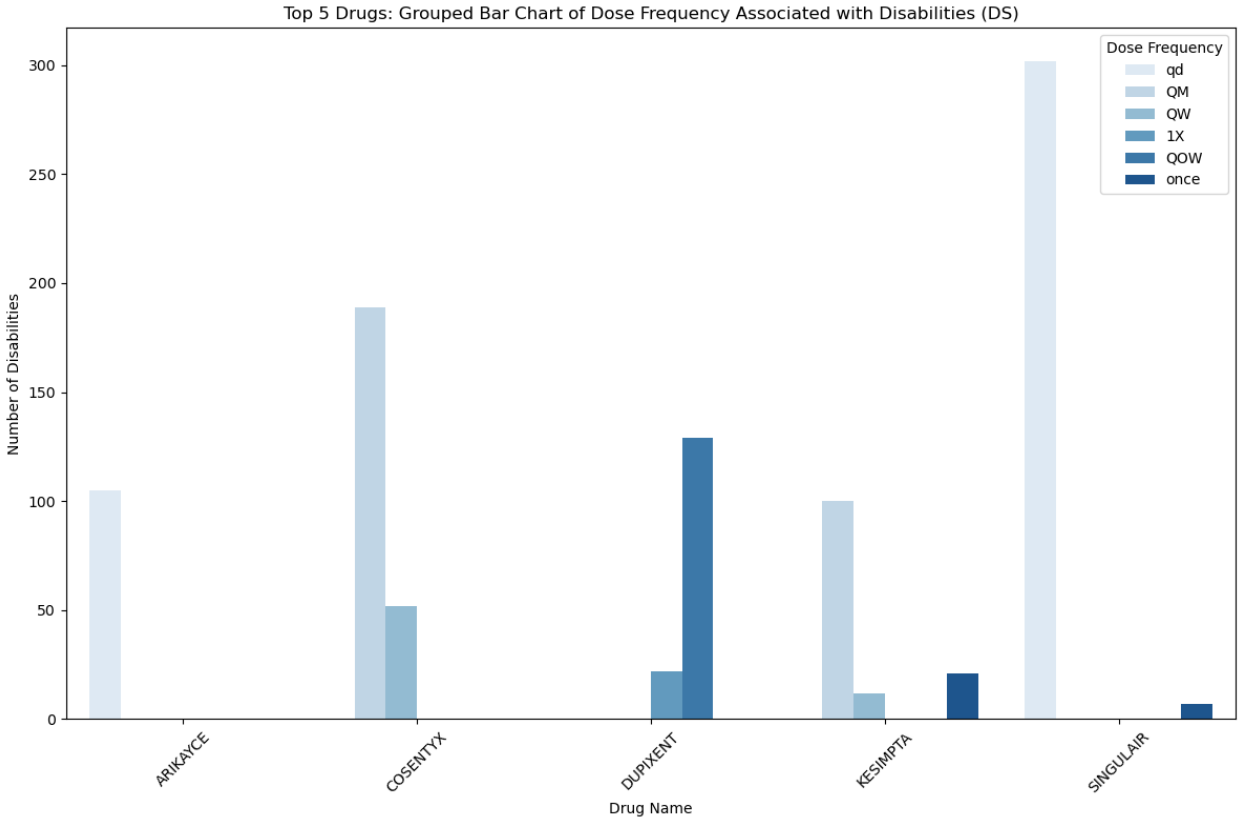
1. Underdstanding drugs causing maximum deaths



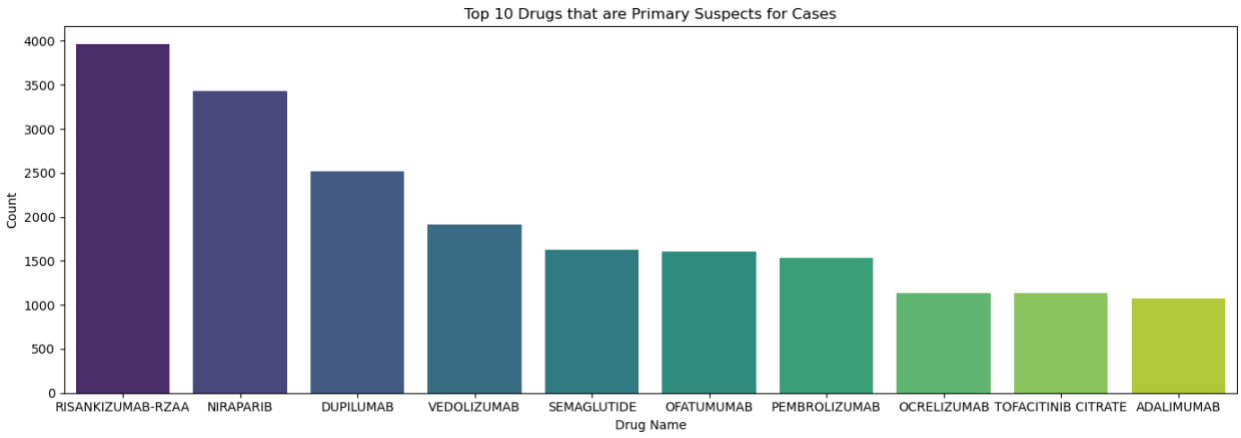
2. The plot explains the dependence of adverse effect on drug frequency for deaths



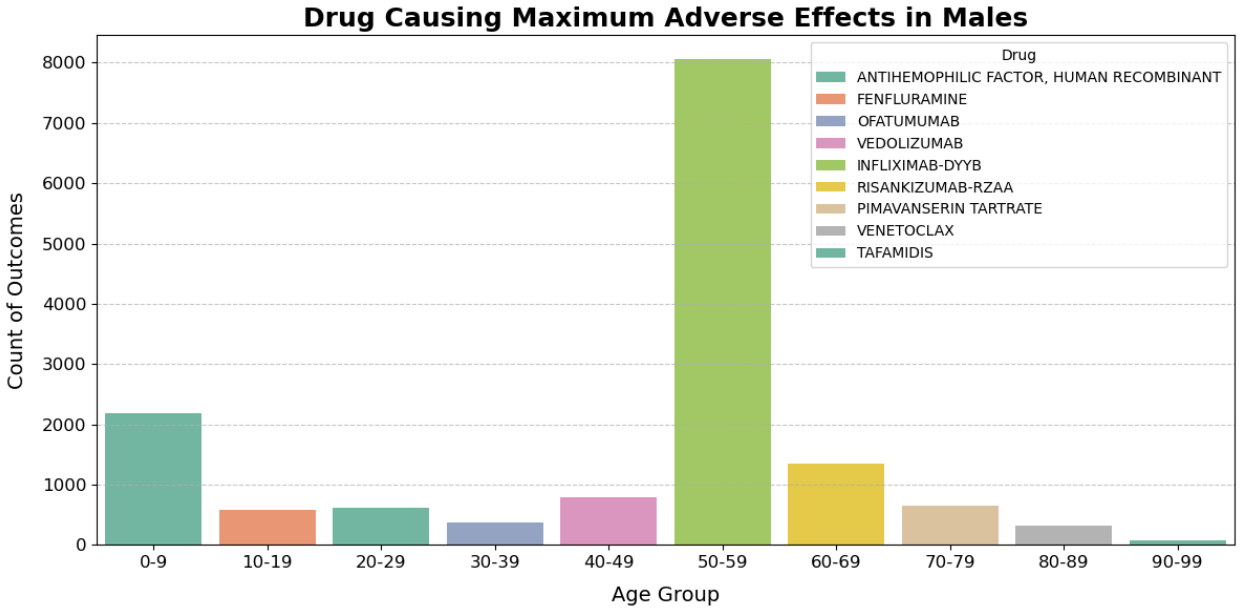
3. The plot explains the dependence of adverse effect on drug frequency for disabilities

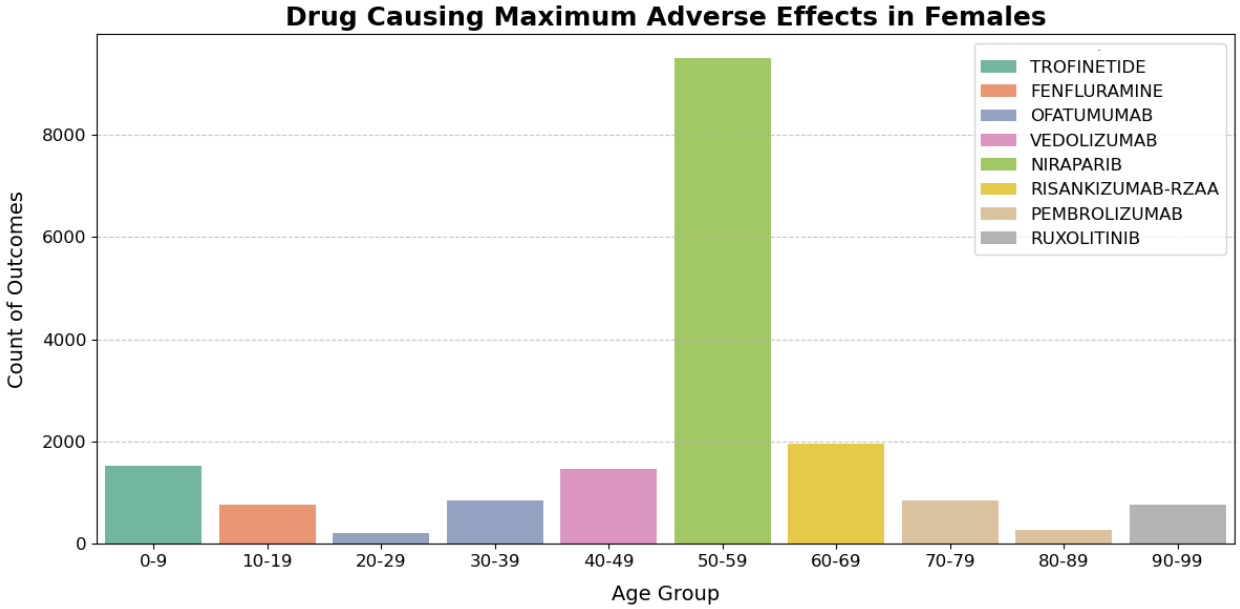


4. Drugs which are the primary cause of adverse effects



5. The plots explain how adverse effects differ among different age groups and gender





Statistical Processes: Validating Data Relationships

Power Analysis

Before diving into complex statistical tests, a power analysis was conducted to determine the appropriate sample size. Power analysis helps researchers understand how many data points are needed to detect statistically significant effects with a desired level of confidence.

The analysis determined that 87 samples would provide reliable results for chi-square testing. This is like calibrating a scientific instrument to ensure it can detect meaningful signals amidst noise.

Chi-Square Tests

Chi-square tests were used to examine relationships between categorical variables. The tests help determine whether observed differences between categories are statistically significant or merely due to random chance.

Variables Failing to Reject the Null Hypothesis (no significant difference):

- Electronic submission type
- Report code
- Reporting month
- Sex
- Month
- Broad Category Route
- Preferred Medical Term
- Route

Variables Rejecting the Null Hypothesis (showing significant differences):

- Manufacturer sender
- Occurrence country
- Drug role code
- Product active ingredient
- Dechallenge status
- Dose frequency

These results suggest that factors like manufacturer, geographical location, and specific drug characteristics have meaningful variations in adverse event patterns.

Kruskal-Wallis Test

For numerical variables, the Kruskal-Wallis test was employed. This non-parametric test determines whether statistically significant differences exist among multiple groups.

Variables Failing to Reject the Null Hypothesis:

- Day of report (no significant variation)

Variables Rejecting the Null Hypothesis:

- Dose amount
- Patient age
- Processed medical term vectors

This indicates significant variations in dose amounts, patient ages, and the semantic content of medical terms across different event categories.

Encoding Strategies: Translating Data for Machine Learning

Encoding is a critical preprocessing step that transforms categorical and text data into numerical formats that machine learning algorithms can understand. Think of it like translating different languages into a common mathematical vocabulary that computers can process.

1. Frequency Encoding and Standard Scaling

Frequency Encoding is a sophisticated technique used for categorical variables with high cardinality (many unique values). In this project, it was applied to:

- Manufacturer sender
- Occurrence country
- Product Active Ingredient
- Primary ID

How it works: Instead of creating numerous binary columns (like one-hot encoding), frequency encoding replaces each category with the frequency of that category in the column. This method captures more nuanced information compared to simpler encoding techniques.

2. One-Hot Encoding

Used specifically for the 'Dechal' (dechallenge) column, one-hot encoding creates binary columns for each unique category. If 'Dechal' has multiple possible values, each value gets its own column with 0/1 representation. This allows the model to treat each category independently without implying any ordinal relationship.

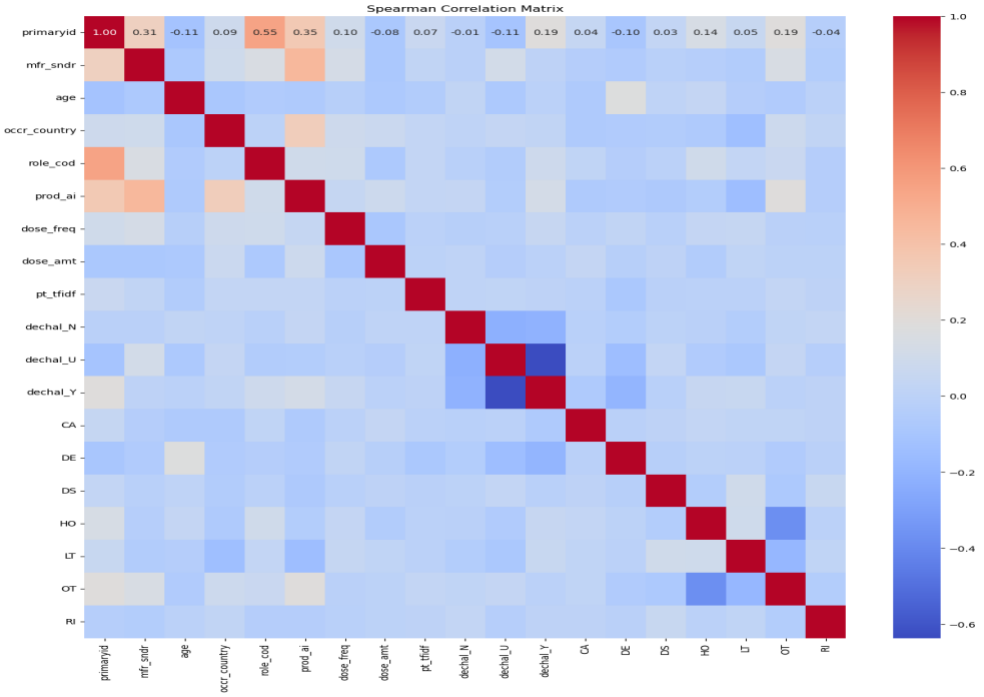
3. Ordinal Encoding

Applied to the 'role_cod' column, ordinal encoding assigns a unique integer to each category. This is useful when there's a natural progression or hierarchy in the categories. For drug roles, this might represent increasing levels of clinical significance.

4. Standard Scaling

Used for numerical features like age, dose amount standard scaling transforms these features to have a mean of 0 and standard deviation of 1. This is crucial because machine learning algorithms can be sensitive to the scale of input features. By standardizing, you ensure that features with larger magnitudes don't disproportionately influence the model. We used it for age, dose amount columns

Correlation Heat Map: The correlation heat map provides valuable insights into the relationships between various features in the dataset. Some key observations:



1. **Feature Correlations:**
    - The heat map visualizes the Spearman correlation coefficients between different features, ranging from -1 to 1.
    - Features with strong positive correlations (indicated by darker red cells) suggest that they tend to vary together in the same direction.
    - Features with strong negative correlations (indicated by darker blue cells) suggest that they tend to vary in opposite directions.
2. **Identified Relationships:**
    - The heat map reveals several interesting relationships, such as the strong positive correlation between 'role_cod' (drug role code) and 'prod_ai' (product active ingredient).
    - It also highlights the negative correlation between 'dechal_N' (dechallenge status) and 'dechal_Y' (dechallenge status), indicating that these two features are inversely related.
    - Other notable relationships can be observed between features like 'age', 'occ_country', 'dose_freq', and 'pt_tfidf' (processed medical term vectors).
    - Observed a strong correlation between role code and primary id.

Reason for not dropping the primary id column:

In our project, the primaryid/caseid is not just a number to uniqely identify rows. It helps us in identifying different drugs, side effects of drugs, outcomes for a particular case which is expressed in several rows. Our main aim was to train a model which can identify drug combinations that can cause adverse effects.

Our dataset is not Independently and Identically Distributed. Rows of each case are associated with each other. Ideally to predict the type of adverse effects, all the drugs belonging to a case should be provided to the model for prediction.

Reason why we did not combine all the rows of a case into a single row:

Few cases had more than 50 rows due to the number of drugs or preferred medical terminology (side effects) of drugs. To combine all such drug combinations and preffered medical terminology into a single row and later into numerical encoded columns will lead to a lot of information loss. We cannot clearly explain which drug causes what type of side effect(pt) and what role the drug played in causing the adverse effect. This would also reduce the size of the dataset to 6000 rows.

Therefore, we decided to use the primary id column in our initial model development to tell the model which rows belong to the same case. For this we performed frequnency encoding of the primaryid. There might be cases which have a same counts of rows but we felt this was the best encoding technique to work with.

Modeling Approaches:

With the Primaryid Column

1. **Logistic Regression:**

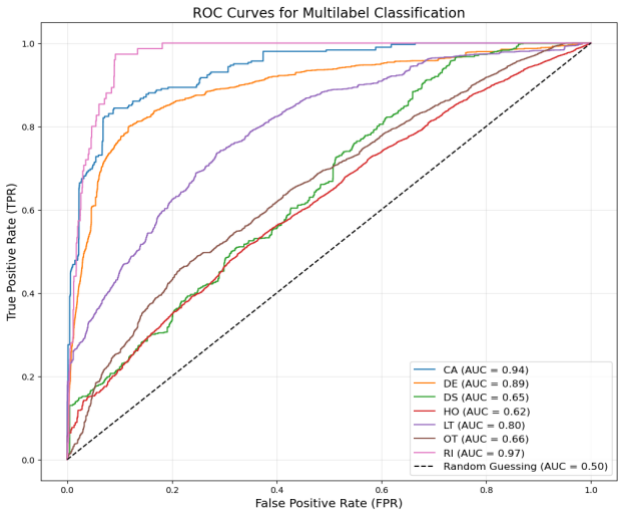
    - The Logistic Regression model achieved an accuracy of 0.4711 on the training data and 0.4733 on the test data.
    - The model's classification report shows the precision, recall, f1-score, and support for various target classes (CA, DE, DS, HO, LT, OT, RI).
    - The ROC (Receiver Operating Characteristic) curves for the multi-label classification task are also provided, highlighting the model's performance for each target class.

```
Accuracy: 0.4733

Classification Report:
              precision    recall  f1-score   support

          CA       1.00      0.01      0.03       301
          DE       0.64      0.39      0.49      3858
          DS       0.00      0.00      0.00      1108
          HO       0.58      0.36      0.44     20910
          LT       1.00      0.15      0.26      2979
          OT       0.81      1.00      0.89     36972
          RI       0.00      0.00      0.00        75

   micro avg       0.76      0.70      0.73     66203
   macro avg       0.58      0.27      0.30     66203
weighted avg       0.72      0.70      0.68     66203
 samples avg       0.78      0.74      0.73     66203
```



2. **Random Forest Classifier:**

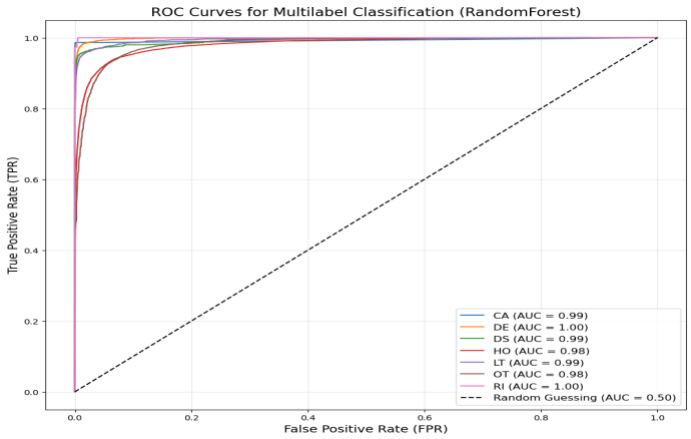
    - The Random Forest Classifier achieved an accuracy of 0.9023.
    - The classification report provides similar metrics as the Logistic Regression model, including precision, recall, f1-score, and support for the target classes.
    - The ROC curves for the multi-label classification using the Random Forest Classifier are also shown, allowing for a comparison with the Logistic Regression Model.

```
Accuracy: 0.9023

Classification Report:
              precision    recall  f1-score   support

          CA       1.00      0.89      0.94       301
          DE       0.97      0.93      0.95      3858
          DS       0.97      0.86      0.91      1108
          HO       0.94      0.91      0.93     20910
          LT       0.98      0.86      0.91      2979
          OT       0.96      0.98      0.97     36972
          RI       1.00      0.81      0.90        75

   micro avg       0.96      0.95      0.95     66203
   macro avg       0.97      0.89      0.93     66203
weighted avg       0.96      0.95      0.95     66203
 samples avg       0.95      0.95      0.94     66203
```



We later dropped the primaryid column, making the assumption now that the dataset is independtly and identically distributed. We train the model on individual independent rows. This model will work because for each row, the drug specifications and side effects are clearly mentioned. Although the model cannot capture the drug combinations which can case adverse effects. It can predict which drugs are prone to adverse effects.

After dropping the primary id, we trained the models again

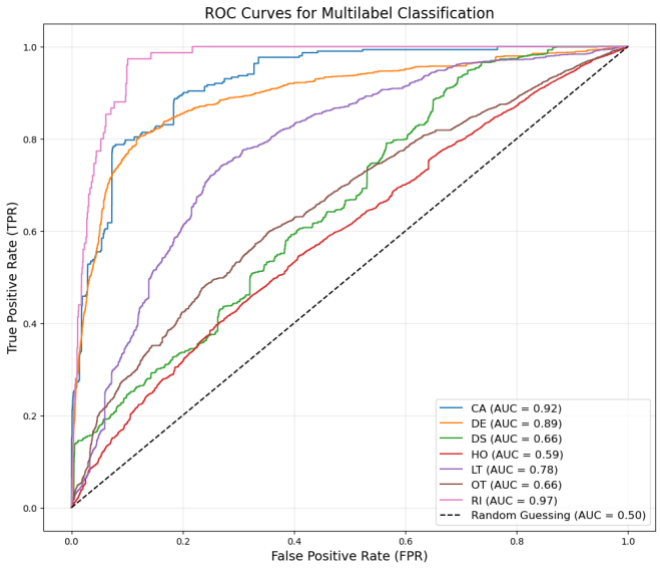
1. **Logistic Regression:**
    - The Logistic Regression model achieved an accuracy of 0.4419 on the training data and 0.4459 on the test data.

```
Accuracy: 0.4459

Classification Report:
              precision    recall  f1-score   support

          CA       1.00      0.01      0.03       301
          DE       0.64      0.40      0.49      3858
          DS       0.00      0.00      0.00      1108
          HO       0.56      0.35      0.43     20910
          LT       0.00      0.00      0.00      2979
          OT       0.80      1.00      0.89     36972
          RI       0.00      0.00      0.00        75

   micro avg       0.74      0.69      0.72     66203
   macro avg       0.43      0.25      0.26     66203
weighted avg       0.67      0.69      0.66     66203
 samples avg       0.78      0.74      0.72     66203
```



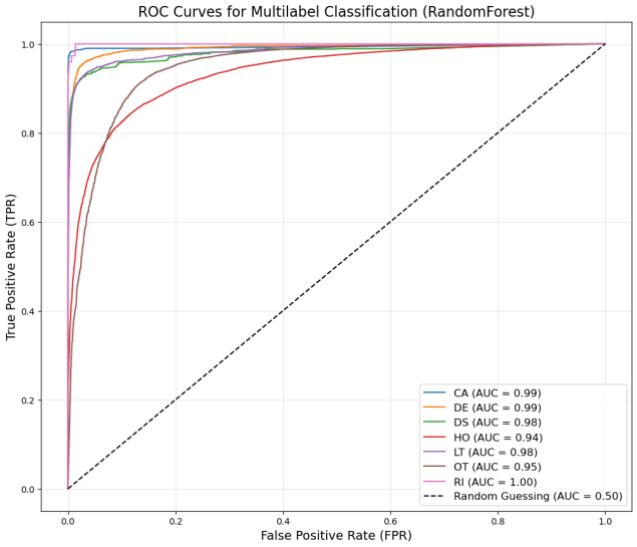
2. **Random Forest Classifier:**
    - The Random Forest Classifier achieved an accuracy of 0.8131

```
Accuracy: 0.8131

Classification Report:
              precision    recall  f1-score   support

          CA       0.98      0.87      0.92       301
          DE       0.92      0.86      0.89      3858
          DS       0.91      0.76      0.83      1108
          HO       0.87      0.82      0.85     20910
          LT       0.93      0.80      0.86      2979
          OT       0.94      0.97      0.95     36972
          RI       0.98      0.76      0.86        75

   micro avg       0.92      0.91      0.91     66203
   macro avg       0.93      0.83      0.88     66203
weighted avg       0.92      0.91      0.91     66203
 samples avg       0.91      0.91      0.90     66203
```



3. XGBoost Classifier:

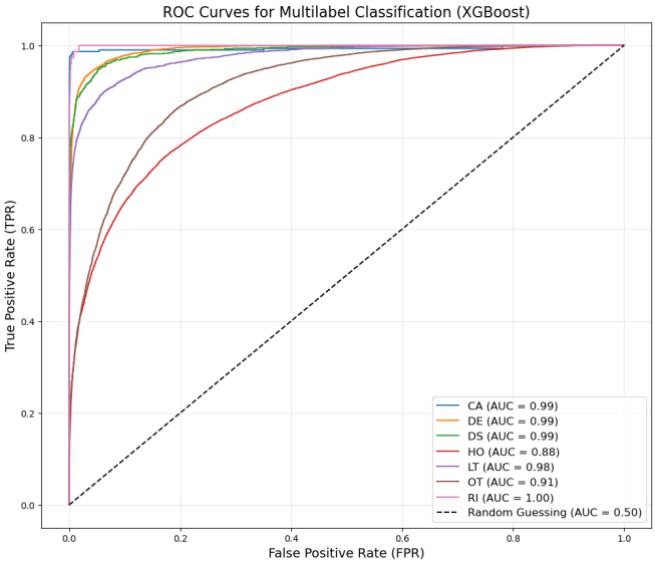
- The XGboost Classifier achieved an accuracy of 0.7133

```
Accuracy: 0.7133

Classification Report:
              precision    recall  f1-score   support

          CA       0.99      0.92      0.95       301
          DE       0.92      0.82      0.87      3858
          DS       0.98      0.67      0.80      1108
          HO       0.82      0.70      0.75     20910
          LT       0.94      0.65      0.77      2979
          OT       0.89      0.97      0.93     36972
          RI       0.94      0.80      0.86        75

   micro avg       0.88      0.86      0.87     66203
   macro avg       0.93      0.79      0.85     66203
weighted avg       0.88      0.86      0.86     66203
 samples avg       0.87      0.87      0.85     66203
```



4. Hyperparameter Tuning using Grid Search:

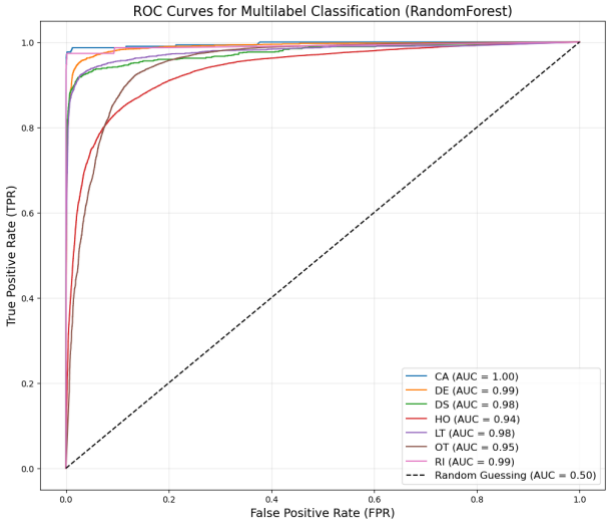
- Hyperparameter tuning process using Grid Search, a technique for systematically exploring a specified parameter space to find the optimal combination of hyperparameters was performed.
- The classification report and ROC curves are provided for the optimized Random Forest Classifier, which achieved an accuracy of 0.8207.

- The hyperparameter tuning likely involved adjusting parameters such as the number of trees, maximum depth, minimum samples per split, and other relevant hyperparameters to improve the model's performance.

Accuracy: 0.8207

Classification Report:

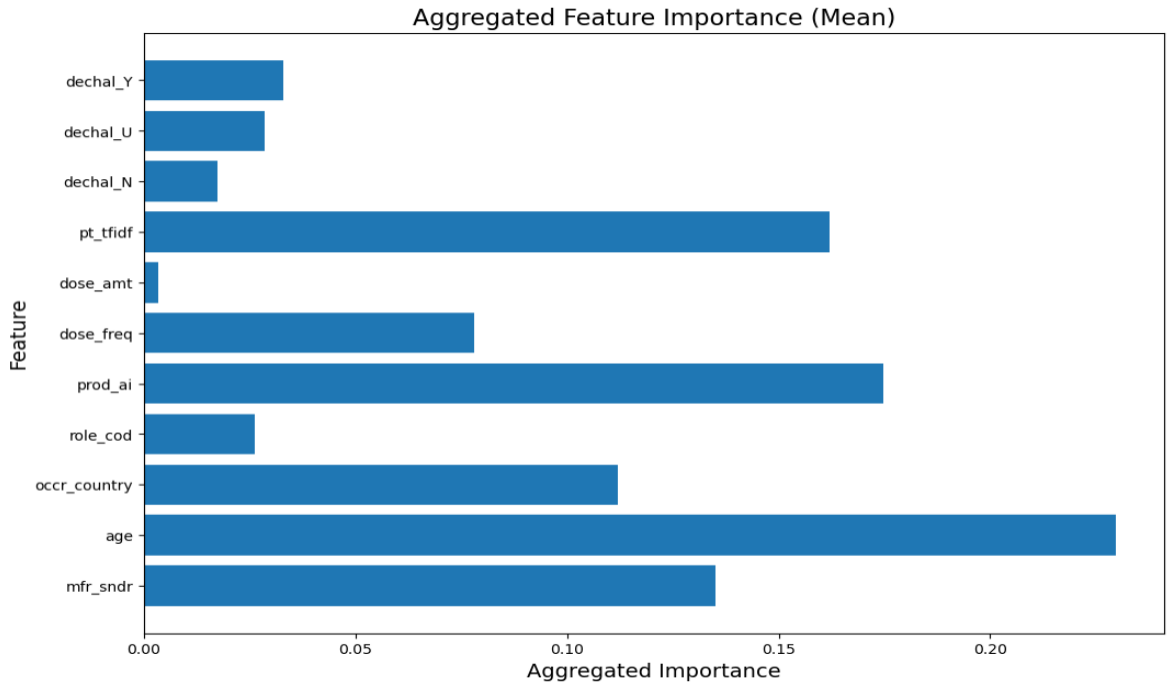
	precision	recall	f1-score	support
0	0.98	0.87	0.92	301
1	0.92	0.86	0.89	3858
2	0.91	0.75	0.83	1108
3	0.88	0.83	0.85	20910
4	0.93	0.80	0.86	2979
5	0.94	0.97	0.95	36972
6	0.98	0.76	0.86	75
micro avg	0.92	0.91	0.91	66203
macro avg	0.93	0.84	0.88	66203
weighted avg	0.92	0.91	0.91	66203
samples avg	0.92	0.91	0.90	66203



We chose the above model to be our final predictive model.

Feature Importance:

- Calculated Feature Importance of the features in X_train to select the most important features in training the model.
- Dropped the Dechal, dose amount, role code columns as their average featire importance score was less than 0.05



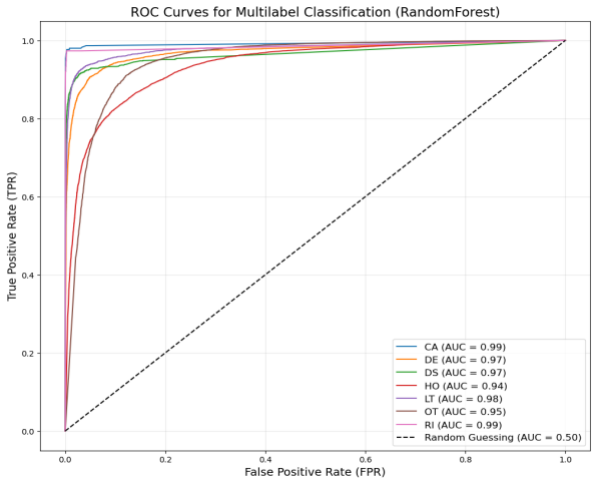
Random Forest Classifier:

- The Random Forest Classifier achieved an accuracy of 0.8060 (less than the previous model)
- After Grid Search, the classifier gave an accuracy of 0.8112

Accuracy: 0.8112

Classification Report:

	precision	recall	f1-score	support
0	0.96	0.88	0.92	301
1	0.90	0.74	0.81	3858
2	0.93	0.76	0.83	1108
3	0.87	0.82	0.85	20910
4	0.92	0.80	0.85	2979
5	0.94	0.96	0.95	36972
6	0.98	0.76	0.86	75
micro avg	0.92	0.89	0.91	66203
macro avg	0.93	0.82	0.87	66203
weighted avg	0.92	0.89	0.90	66203
samples avg	0.92	0.90	0.90	66203



Conclusion and Future Scope:

This comprehensive analysis of the FAERS dataset demonstrates the potential of advanced data science techniques in enhancing pharmaceutical safety monitoring. By transforming complex, multi-dimensional adverse event data into actionable insights, the research provides a robust framework for predictive risk assessment.

The methodological approach combining feature engineering, careful preprocessing, and advanced machine learning techniques offers a powerful toolkit for regulatory bodies and healthcare professionals to proactively identify and mitigate medication-related risks.

Future research could focus on expanding the model's predictive capabilities in identifying drug combinations which are fatal and developing real-time risk assessment frameworks that could potentially save lives by identifying adverse event patterns earlier in the medication lifecycle.