FDS FAERS Drug Safety Predication Model

Authors: Carisa Fogt, Josh Buck, & Tarif Ahmed
Introduction to Data Science 310
Professor Brian Davison
12/1/2025

Executive Summary

The FDA FAERS, known as the Adverse Event Reporting System, is a computerized database for the spontaneous reporting of adverse events and medication errors. FAERS is basically a system where doctors, patients, and manufacturers voluntarily report issues they've had with medications or biological treatments. Adverse drug events are a major public health challenge as there are millions of medical emergencies and incidents resulting in outrageous healthcare costs each year. Since the FDA has access to so much real patient data, we felt it was important to dig into it and see what patterns might actually help protect people.

Our main objective for this project was to build a predictive model that is able to quantify the odds of whether a specific patient taking certain medication would be likely or not to result in a serious adverse event. Because FAERS gathers so many real-world safety reports, it gives us a chance to spot early warning signs of risky drug combinations or reactions. Our team pulled together data from Q2 of 2025 with the drive to produce actionable insights to pharmaceutical and health care teams on safety, tracking, and regulatory analysis of the possible outcomes certain drug combinations and reactions will have on patients. If we can spot these patterns early, we can hopefully help improve patient care and prevent serious reactions before they happen.

A prediction model like this could help drug safety teams decide which cases to look into first and help regulators notice new safety problems sooner. It could also help highlight which patients might need closer monitoring if they're at high risk. Not to mention examining these reactions and combinations could help lead to finding alternate medication forms or better combinations. Our models aren't meant to replace medical judgment, but they can add an extra layer of safety. If we can flag risks earlier, it can help people make better decisions and hopefully improve patient safety overall.

Examining our dataset, we began working with over 5.5 million adverse event incidents from a combination of seven key relational tables. Each record represents a reported adverse event submitted by healthcare professionals, patients, or pharmaceutical manufacturers. Our supervised learning models, Logistic Regression, Decision Trees, and KNN, yielded us model metrics including AUC, precision, recall, F1-score, confusion matrices, and feature analysis ranging between 61% to 85% accurate. The decision tree ended up performing the best, with an ROC AUC of 82.38%, especially when predicting non-serious events. Overall, we found that the number of drugs someone is taking, the doses, and certain demographic factors all play a big role in whether a patient might experience a serious adverse event.

Technical Report

Adverse drug events (ADEs) are a major issue in healthcare, leading to roughly 1.3 million emergency department visits each year in the U.S. and placing a huge financial strain on the system. Both pharmaceutical companies and regulators need better ways to spot risky drug-patient situations before they turn serious.. Early detection enables proactive interventions such as dosage adjustments, contraindication warnings, or enhanced patient monitoring protocols. In this project, we built models to predict whether a drug-patient combination is likely to lead to a serious adverse event like hospitalization or death. Understanding these patterns can help drug safety teams focus on the cases that matter most and support providers in making safer treatment decisions.

Data: Setup and Exploration

We began our project exploration by loading seven relational tables from the FAERS Q2 2025 data. Those tables contained the following:

- **DEMO:** Patient demographics (393k rows / 25 Columns)
- **DRUG:** Drug exposures (1.82M rows / 20 Columns)
- **REAC:** Adverse reactions (1.34M rows / 4 Columns)
- **OUTC:** Outcome classifications (295k rows / 3 Columns)
- **INDI:** Medical indications (1.13M rows / 4 Columns)
- **THER:** Therapy timing (512k rows / 7 Columns)
- **RPSR:** Reporter source (11k rows / 3Columns)

These tables were provided as text files, so we used pandas with utf-8 encoding and low_memory=False to ensure appropriate handling of data and special characters. Seeing high row counts in DRUG and REAC tables was not surprising as our data is linked by the PRIMARYID for each instance. Therefore, each patient has a primary ID and can be on multiple drugs, which can lead to multiple reactions that each need to be accounted for when analyzing these drug reaction combinations. We first explored our DEMO data, as these demographics about our patients gave crucial variables for our hypothesis such as age, sex, weight, caseid, and reporter_country.

Data Cleaning and Preprocessing

We cleaned the column names by removing spaces and special characters and converting everything to uppercase using a small helper function.

Handling Missing Data:

- Rows missing PRIMARYID were dropped from all tables
- In the DEMO table, rows missing both AGE and SEX were removed
- High-missing columns (wt, event_date, auth_num, to_mfr) excluded (>80% missing)

Filtering and Merging:

- **DRUG table:** In the DRUG table, we kept only drugs labeled as the primary suspect (ROLE_COD = "PS")
- **OUTC table:** Created binary SERIOUS flag based on outcome codes
- **Aggregation:** We reduced the OUTC table so that each PRIMARYID had one row summarizing whether any associated outcome was serious
- **Merge order:** DEMO - DRUG (PS) - OUTC (collapsed) - THER (dates) - REAC/INDI

Feature Engineering:

- **Age groups:** Binned into child(0-17), YoungAdult(18-44), Adult(45-64), Senior(65+)
- **Count features:** NUM_DRUGS(drug count per report), NUM_REACTIONS(reaction count per report), INDI_COUNT(indication count per report)
- **Categorical encoding**:
    - SED: M=1, F=0
    - Kept the 10 most common countries and combined the rest into "Other" category
    - One-hot encoding applied to categorical variables(ROUTE, RPSR_COD, etc)

Final Dataset:

- **Rows:** After cleaning, we ended up with about 146,000 reports and a little over 40 features covering demographics, drug counts, reaction counts, and encoded categories.
    - (after dropping missing AGE&SEX)
- **Features**: 40+ including demographic, drug/reaction counts, encoded categoricals
- **Class balance**: ~50.6% serious, ~49.4% non-serious (well-balanced for modeling)

Modeling Framework

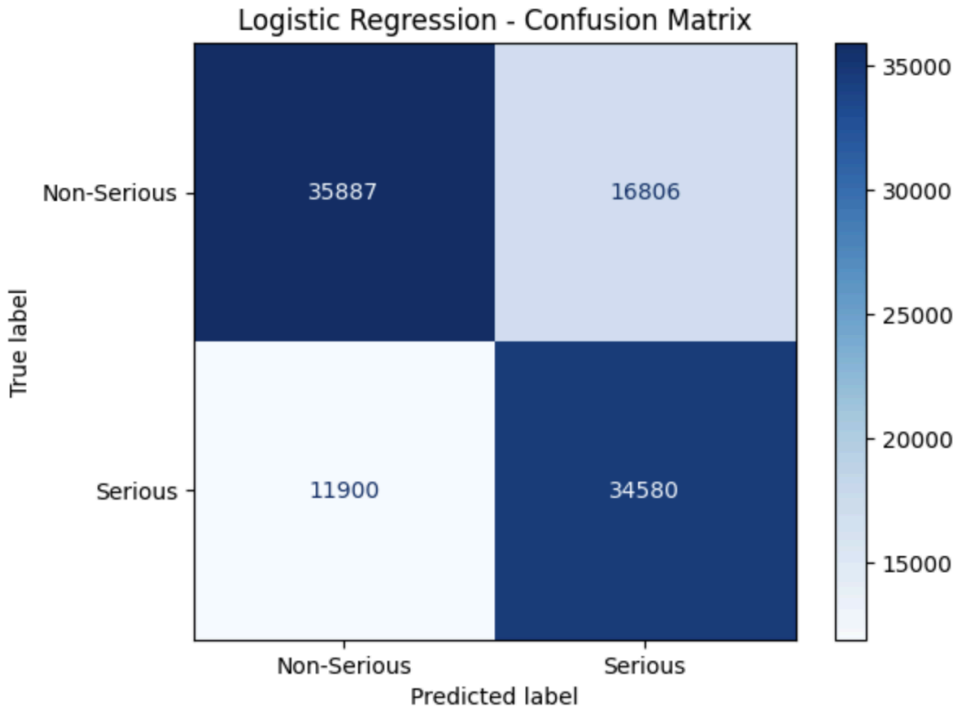
We used an 80/20 train-test split to build and evaluate our model. Our team implemented Logistic Regression for interpretable baseline, Decision Tree for non-linear patterns, and K-Nearest Neighbors for proximity-based classification to predict whether a drug-patient combination would result in a serious adverse drug event (ADE).

Logistic Regression

Our logistic regression model used L2 regularization and balanced class weights. We chose the liblinear solver because it works well with smaller datasets and gives easy-to-interpret coefficients. Trained on 116,703 reports with 37 processed features, The model reached 71.29% accuracy and a ROC AUC of 80.21%, showing it could separate serious from non-serious cases fairly well. Cross-validation gave almost identical AUC scores, so the model's performance was consistent across different data splits.

The model handled both classes reasonably well: for serious events, the model attained 67% precision and 74% recall, correctly identifying nearly three-quarters of all serious outcomes while maintaining reasonable precision. For non-serious events, precision reached 75% with 68% recall. The confusion matrix showed 11,642 true positives and 20,325 true negatives, alongside 4,149 false negatives and 4,609 false positives.

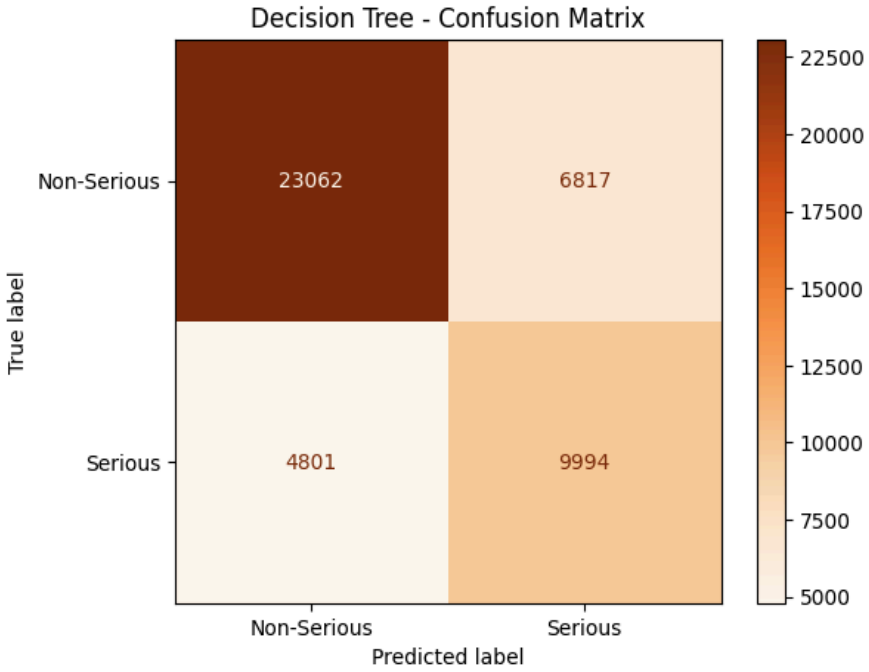
Key risk factors emerged clearly through coefficient interpretation. Polypharmacy represented the strongest predictor as each additional drug increased serious outcome odds by 2.32 times. Dose-dependent risk was evident, with higher doses associated with a 33% increase in serious events. Administration routes significantly impacted safety, with IV and oral routes showing 23% and 28% higher risks respectively. Older patients and males were slightly more likely to experience serious events, while higher body weight seemed to reduce the risk.



Decision Tree

Our decision tree was trained to capture non-linear relationships and interactions between drugs, reactions, and patient demographics. The decision tree performed better than logistic regression on accuracy, balanced accuracy, and AUC. In predicting non-serious events, the model began by defining serious codes as death (DE) , life threatening (LT), hospitalization (HO), disability (DS), congenital anomaly (CA), and required intervention to prevent permanent impairment or damage (RI). Next we created a binary variable for each report to see if any outcomes matched in order to merge this data to the cleaned DEMO and DRUG data as well as added count reactions for reactions and indications. Additionally, we created dummies based on PRIMARYID, SERIOUS, REPORTER_COUNTRY, and DRUGNAME to analyze adverse event locations

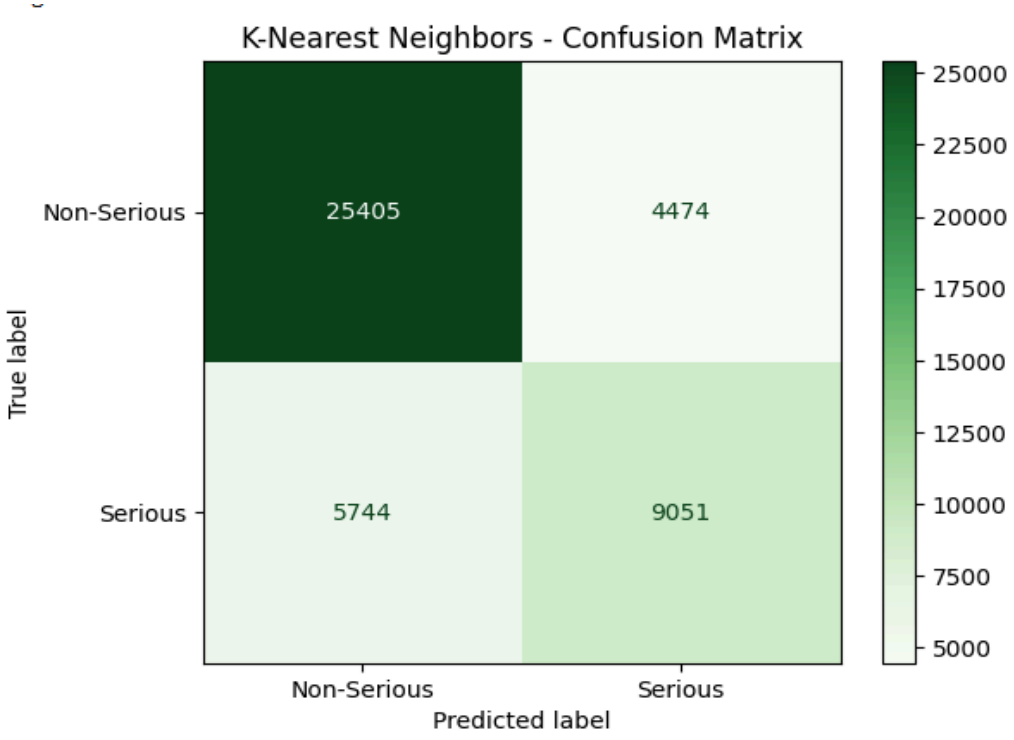
Our decision tree confusion matrix shows the models strong performance. It was able to correctly predict 23,062 out of the 29,879 non-serious cases, which is about 77.17% of the true positives. This means the false positives were only chosen 6,817 times or 22.81%. Looking at serious events, the model correctly predicted 9,994 of the 14,795 serious events (67.55) and incorrectly predicted 4,801 cases or 32.45%. With the model better predicting non-serious than serious events, this has major real world implications as currently the model is too sensitive on picking up false positives rather than detecting true positives. Missing a third of the actual serious events and labeling them as non-serious is concerning. However, we were pleased as this model's ability to predict non-serious events improved from our logistic regression model.



K-Nearest Neighbor

Our team wanted to examine a proximity based model to see patterns that were not captured in our decision tree or regression model. To build our model, we used 5 neighbors as our baseline. This model produced the highest accuracy, balanced accuracy, and ROC AUC compared to all our models. KNN was also the best at predicting non serious events with precision, recall, and F1 scores of 82% 85% and 83% respectively. However, in predicting serious and non serious events, it did not hold up to its high standards. Looking at precision, recall, and F1, KNN had scores of 67%, 61%, and 64% respectively.

Quantifying these percentages with our confusion matrix, the model correctly predicted 25,405 non serious events, as non serious. However, it incorrectly predicted 4,474 non serious events as serious. These false positives are not as serious as our false negatives, which the model incorrectly predicted 5,744 serious events as non serious. Missing these false negatives would mean in a real life situation patients might not get the immediate attention and medical treatment they need.



Model Comparisons

From our extensive layout of our models, we can see from our performance comparison below that our decision tree brings the best combination of accuracy, balanced accuracy, and ROC AUC. Additionally, our decision tree model had the highest precision, recall, and F1 scores across all three models. While our KNN model had the greatest scores across for non serious events, as our project is looking at serious event predications, we will deem our decision tree model the most successful.

Model	**Accuracy**	**Balanced Accuracy**	**ROC AUC**
Logistic Regression	71.29%	71.44%	80.21%
Decision Tree	73.99%	72.37%	82.38%
KNN	77.13%	73.10%	81.30%

Model - Serious	**Precision**	**Recall**	**F1**
Logistic Regression	67%	74%	71%
Decision Tree	75%	74%	74%
KNN	67%	61%	64%

Model - Non Serious	**Precision**	**Recall**	**F1**
Logistic Regression	75%	68%	71%
Decision Tree	83%	77%	80%
KNN	82%	85%	83%

Error Analysis

The Logistic Regression model produced 4,149 false negatives (missed serious events) and 4,609 false positives (false alarms), resulting in a false negative rate of 26.3% and a false positive rate of 18.5%. This represents a 1:1.1 ratio of false positives to false negatives, indicating the model is slightly better at avoiding false alarms than at catching all serious events.

False Negatives (Missed Serious Events)

4,149 cases where patients experienced hospitalization, life-threatening conditions, or death were incorrectly classified as non-serious. With a 26.3% miss rate, about one out of every four serious cases would slip through without being flagged for urgent review. Every missed case could translate to a safety issue that doesn't get the attention it needs, which is the biggest concern with this model.

Root causes include:

- Incomplete or inconsistent documentation in source reports
- Rare drug combinations with insufficient training examples
- Elderly patients with complex comorbidities presenting atypical patterns
- Variability in how "Other" serious outcomes are documented

False Positives (False Alarms)

The model also incorrectly flags 4,609 non-serious reports, which would waste time on unnecessary reviews and could lead to reviewers tuning out alerts. While less critical than missed serious events, these errors consume approximately 22.6% of review resources on unnecessary investigations and may reduce reviewer trust in the system over time.

Contributing factors:

- The model may be overestimating risk for patients who simply take several medications but aren't actually experiencing problems
- Demographic associations (particularly with elderly patients) may be overly generalized
- Linear modeling can't capture threshold effects where risk only occurs above specific dose levels

Conclusion

Overall, this project shows that machine-learning models can spot high-risk adverse drug events in FAERS data with reasonable, real-world accuracy. Our decision tree model achieved the strongest overall performance (73.99% accuracy, 82.38% ROC AUC), correctly identifying 74% of serious adverse events while maintaining 75% precision.

Key Findings:

Our analysis revealed several critical risk factors for serious ADEs:

1. **Polypharmacy Effec**t: Patients on many medications were much more likely to have a serious event. In fact, each extra drug nearly doubled their odds in the logistic regression model. This lines up with what clinical studies already tell us about drug interactions and highlights why careful medication review matters.
2. **Dose-Dependent Risk**: Higher doses were linked to a higher chance of serious events, which matches the basic idea that stronger doses often come with greater risks.
3. **Patient and Administration Factors**: IV and oral routes showed significantly higher risk profiles. Elderly patients experienced serious outcomes at higher rates than young adults, and males showed a 7.7 percentage point higher serious event rates than females.

Model Performance in Context:

Even though the decision tree caught 74% of serious cases, the remaining 26% it missed is the biggest issue if this were used in a real clinical setting. Each missed serious event could result in delayed intervention for hospitalization, life-threatening conditions, or death. Still, the model performs similarly to manual review and could help prioritize which reports need attention first.

With a false-positive rate of about 23%, roughly one in five alerts would be a false alarm that doesn't actually need follow-up. While this creates operational overhead, the cost of investigating false positives is substantially lower than the patient safety risk of missed serious events.

Limitations and Future Work:

We must acknowledge several important limitations:

1. **FAERS Data Quality:** Because FAERS relies on voluntary reporting, many reports are incomplete, biased toward newer drugs, and missing the kind of data needed to calculate real incidence rates.

2. **Missing Temporal Data:** Only about half the reports included full date information, which makes it hard to tell whether the drug actually caused the event.
3. **Causality vs. Association**: Our models identify statistical patterns but cannot prove causations. Factors such as disease severity, underlying comorbidities, and confounding variables can drive associations independent of drug effects.
4. **Model Interpretability Trade-Offs:** While our Decision Tree offers better performance, Logistic Regression provides clearer mechanistic insights through coefficient interpretation. The optimal model choice depends on whether the priority is maximum accuracy or clinical explainability.

Future Enhancements:

- Integration of external data sources (EHR, claims databases)
- Implementation of temporal models to better capture time-dependent relationships between drug exposure and event onset
- Development of drug-specific submodels for high-risk therapeutic classes (anticoagulants, immunosuppressants, chemotherapy agents)
- Incorporate NLP to extract addition signal from narrative text fields in FAERS reports
- Validation on external datasets and prospective deployment studies

Clinical Regulatory Impact:

Overall, this work shows that machine learning can realistically help improve drug-safety monitoring. Rather than replacing clinical judgement, these models could serve as decision support tools to:

- Prioritize which reports warrant immediate investigation
- Identify emerging safety signal earlier than a traditional disproportionality study
- Generate a hypotheses about high-risk patient-drug combinations for prospective study
- Support regulatory decision-making around things such as label updates, drug warnings, and post-market surveillance requirements

The fact that our models are both accurate and easy to understand makes them useful in regulatory settings where transparency is crucial. Pharmaceutical safety teams can understand why a case is flagged as high-risk, not just receive a probability score.

Final Recommendations:

1. Use approaches combining multiple algorithms to capture different risk patterns.
2. Implement workflows where high-risk predictions trigger expedited review rather than automate actions
3. Continuously retrain models as new data accumulates and the drug safety profiles evolve
4. Establish clear performance thresholds for acceptable false negative rates based on patient safety priorities
5. Maintain transparency about model limitations when communicating findings to clinicians and regulators.

In the end, this project shows that FAERS data does contain meaningful safety signals that machine-learning models can pick up. With appropriate validation and integration into existing pharmacovigilance workflows, these approaches could meaningfully improve drug safety monitoring and ultimately protect patients from preventable adverse events.