CS 573 Final Project Proposal: Predicting Adverse Drug Reactions

Nicholas Rosenorn
nrosenor@purdue.edu
Purdue Univeristy
West Lafayette, Indiana, USA

Stephan Zapodeanu
szapodea@purdue.edu
Purdue Univeristy
West Lafayette, Indiana, USA

Jeret McCoy
mccoy86@purdue.edu
Purdue Univeristy
West Lafayette, Indiana, USA

Hieu Tran
tran335@purdue.edu
Purdue Univeristy
West Lafayette, Indiana, USA

Sreeram Nagappa
snagapp@purdue.edu
Purdue Univeristy
West Lafayette, Indiana, USA

ABSTRACT

Pharmacovigilance, the science of monitoring and analyzing adverse drug reactions (ADRs), plays a critical role in ensuring patient safety and the effectiveness of pharmaceutical products. This project uses data mining and machine learning techniques to enhance ADR prediction and contribute to drug safety assessment. Leveraging the comprehensive FDA Adverse Event Reporting System (FAERS) database, our project focuses on the development of predictive models for ADRs associated with specific drugs.

The project starts with data acquisition and preprocessing, involving the extraction of structured and unstructured information from FAERS reports. Exploratory data analysis (EDA) offers valuable insights into the distribution of ADRs, drugs, and patient demographics, setting the stage for feature engineering and model development. Various machine learning algorithms are considered, including logistic regression, decision trees, random forests, and support vector machines, with extensive hyper-parameter tuning to optimize model performance.

Beyond binary classification, this project may explore additional problem statements such as ADR severity prediction, temporal analysis of ADR trends, and geospatial variations in ADR reporting. The predictive models aim to provide not only early warnings of potential ADRs but also insights into the severity, temporal patterns, and geographic distributions of adverse events. Interpretability techniques shed light on the critical factors influencing ADR predictions.

Our project represents a comprehensive effort to leverage FAERS data for proactive drug safety assessment and offers insights into the broader field of pharmacovigilance. The results and methodologies presented herein contribute to the ongoing efforts to improve patient safety, guide regulatory decisions, and enhance the overall understanding of ADRs in the pharmaceutical landscape.

CCS CONCEPTS

• **Applied computing → Consumer health**.

KEYWORDS

Pharmacovigilance, Adverse Drug Reactions (ADRs), FAERS Database, Data Mining, Machine Learning, ADR Prediction, Drug Safety, Patient Safety, Drug Surveillance

1 PROJECT MOTIVATION

Pharmacovigilance stands as a cornerstone in ensuring the safety and efficacy of pharmaceutical products. In an era characterized by modern medication and increasing patient populations, the timely identification and assessment of ADRs are important to overall public health and ensuring patient saftey. The FDA Adverse Event Reporting System (FAERS) is a repository of real-world ADR reports, making it a resource for advancing drug safety by applying data mining techniques to its datasets. This project is motivated by the significance of leveraging FAERS data to innovate and enhance ADR prediction methodologies.

Understanding the Challenge: The pharmacovigilance landscape has evolved as cutting-edge medicine has progressed - this emergence has naturally presented new challenges. With millions of ADR reports added to FAERS annually, the volume of data necessitates advanced analytical techniques to sift through the noise and uncover meaningful patterns. Additionally, the emergence of complex medications and personalized medicine underscores the need for innovative approaches that consider diverse patient demographics and therapeutic contexts.

The Interest in ADR Severity and Temporal Trends: While binary classification of ADRs remains a fundamental goal, there is a growing interest in understanding the severity of ADRs and their temporal patterns. Predicting not only the occurrence but also the severity of ADRs is essential for prioritizing resources and interventions. Furthermore, temporal analysis provides insights into changing ADR landscapes, offering the potential to detect emerging risks early.

Unlocking Geographic Insights: Another opportunity for innovation lies in the exploration of geospatial variations in ADR reporting. Understanding how ADRs vary by region can shed light on regional prescribing practices, environmental factors, and genetic influences. This geographic perspective can inform targeted interventions and regulatory decisions.

Interpretable Models for Safer Medications: Advancements in machine learning and data mining offer the promise of more accurate ADR prediction models. However, the interpretability of these models is crucial for building trust and facilitating decision-making. Innovative approaches to model interpretability can provide insights into the factors driving ADR predictions, fostering a deeper understanding of drug safety.

Contributing to Patient Safety: Ultimately, the motivation behind this project is driven by the desire to contribute to patient safety. By enhancing the ability to predict, assess, and understand

ADRs, we aim to provide healthcare professionals, regulators, and pharmaceutical companies with valuable tools and insights. These advancements have the potential to lead to safer medications, improved patient care, and more informed decisions about drug use.

In summary, this project finds its motivation in the need to use FAERS data repository to address the challenges of pharmacovigilance. By exploring data mining applications in ADR prediction, severity assessment, temporal analysis, and geospatial insights, we hope to find meaningful drug safety insights, ultimately benefiting the health and well-being of patients worldwide.

2 PROJECT MILESTONES AND TIMELINE

The execution of the Adverse Drug Reaction (ADR) Prediction project using the FDA Adverse Event Reporting System (FAERS) data will be driven by structuring our project plan into key milestones. Each milestone represents a phase of the project that will guide us from project initiation to dissemination. Here is an overview of our project milestones:

(1) **Literature Review and Background Research**
We will conduct a literature review on pharmacovigilance, ADR prediction, and related methodologies. The team will become familiar with the FAERS database and relevant data mining techniques. Completed by September 29, 2023.

(2) **Data Collection and Preprocessing**
We will collect data from the FAERS dataset, and perform data preprocessing tasks, including handling missing values, standardizing data formats, and selecting relevant columns. The outcome will be a clean and well-structured FAERS dataset ready for analysis. Completed by October 6, 2023.

(3) **Exploratory Data Analysis (EDA)**
We will explore the FAERS data to understand its distribution, trends, and patterns. We will also visualize ADRs, drugs, and patient demographics in hopes of finding key insights to inform feature engineering and model design. Completed by October 13, 2023.

(4) **Feature Engineering and Model Design**
We will define relevant features for ADR prediction, considering factors like drug characteristics, patient demographics, and temporal information. We will design machine learning models for ADR prediction, severity assessment, and temporal analysis and finalize a feature-engineered dataset and well-defined model architectures. Completed by October 13, 2023.

(5) **Model Implementation and Hyperparameter Tuning**
We will implement machine learning models using Python libraries and frameworks, and conduct hyperparameter tuning to optimize model performance. Here we will have trained models with optimized hyperparameters. Completed by October 27, 2023.

(6) **Model Evaluation and Interpretability**
We will evaluate the models using appropriate metrics, including accuracy, precision, recall, F1-score, and ROC AUC. To understand feature importance, we will implement model interpretability techniques and conclude with performance metrics, confusion matrices, and insights into model decisions. Completed by October 27, 2023.

(7) **Exploration of Additional Problem Statements**
We will explore alternative problem statements beyond binary classification, such as severity prediction, temporal analysis, and geospatial variations, developing prototypes for these tasks. We will deliver preliminary results and prototypes for additional problem statements. Completed by November 10, 2023.

(8) **Final Model Evaluation and Integration**
We will perform a final evaluation of the ADR prediction models, considering the entire dataset including integrating insights from additional problem statements where applicable. We will collect final model performance metrics and a consolidated set of findings. Completed by November 17, 2023.

(9) **Documentation and Report Writing**
We will prepare a project report documenting the entire process, including data preprocessing, model development, and results interpretation. In the report, we will meet the expectations set by the course project description handout. Completed by November 24, 2023.

(10) **Project Presentation and Submission**
We will present the project findings and insights to the course instructor and peers, and submit the final project report and code. We will the guidelines set by the course project description handout. Completed by December 1, 2023.

3 EVALUATION PLAN

When considering an evaluation plan for this project, we have three primary goals we intend to consider:

(1) **ADR Prediction:** The primary goal of the project is to develop accurate predictive models for adverse drug reactions (ADRs) using the FDA Adverse Event Reporting System (FAERS) data. We aim to identify potential ADRs associated with specific drugs and assess their performance.
(2) **Additional Problem Statements:** In addition to binary ADR prediction, we aim to explore alternative problem statements such as ADR severity prediction, temporal analysis of ADR trends, and geospatial variations in ADR reporting.
(3) **Interpretability:** We strive to enhance model interpretability to provide insights into the factors driving ADR predictions and to build trust in the models.

With these goals driving our project, we also consider specific evaluation criteria:

(1) **Model Performance Metrics:** In the binary ADR prediction case, we will measure the success of ADR prediction model using standard classification metrics such as accuracy, precision, recall, F1-score, and ROC AUC. These metrics will assess the models' ability to correctly classify ADRs and non-ADRs. For the additional problem statements outside of binary classification, we will define evaluation metrics specific to each task. For ADR severity prediction, for example, we may use mean squared error (MSE) or a similar metric to assess the accuracy of severity predictions.
(2) **Cross-Validation and Testing:** We will conduct cross-validation to assess model performance on multiple subsets of the

FAERS dataset, ensuring robustness and generalizability. Final model evaluation will be performed on a separate test dataset that has not been used during model training or hyperparameter tuning.

(3) **Comparison with Baselines:** Where applicable, we will compare our models' performance with baseline models or traditional pharmacovigilance methods to demonstrate the added value of our approach.

(4) **Documentation and Reporting:** The project's documentation and report will be an important component of evaluation. Clear visualizations, along with well-documented methodologies, will provide transparency and support the assessment of our work.

The achievement of project goals will be measured through a combination of quantitative and qualitative assessments. Successful achievement will be indicated by the following:

- High model performance metrics for ADR prediction.
- Consistent and meaningful results for additional problem statements, supported by relevant evaluation metrics.
- Clear and informative model interpretability results, demonstrating the ability to understand and explain ADR predictions.