# Class Project: Final Submission 1 (Report)
# Balancium RX

Chehak Arora, Andrew Normandin, Marius Nwobi, David Rogers & Franco Valdes

## 1. Introduction

Problem Statement
The growing number of drugs has made predicting drug-drug interactions (DDIs) a critical challenge in ensuring safety and efficacy. Adverse DDIs can reduce efficacy, increase toxicity, and cause serious health complications, yet traditional detection methods are limited by the vast number of drug combinations and complex mechanisms. Machine learning offers a promising solution, leveraging large datasets to predict DDIs early in development, but current models face issues with data quality, completeness, and generalizability. Developing robust and interpretable models is essential to improving drug safety and clinical outcomes.

Objective: The goal of this project was to develop a machine-learning model to predict DDIs using molecular descriptors and chemical properties as predictors. By analyzing and comparing multiple models, we aimed to evaluate the performance of different algorithms in predicting DDIs accurately.

## 2. Data Collection and Preprocessing

Data Sources:
The primary datasets used in this project were:

- **Drug-Drug Interaction (DDI) Dataset:** A collection of drug-drug interactions from the database DDinter involving alimentary tract and metabolism drugs, marking the level of interaction on a scale of minor, moderate, and major, mapped as 0, 1, and 2 in classification, respectively.
- **PubChem Data:** Contained molecular descriptors and chemical properties retrieved using CIDs retrieved from the PubChem database.
- **Drug-CID Mapping:** We used a mapping between drug names and PubChem CIDs, simplifying batch queries for data retrieval.

Data Cleaning and Preparation:
Data cleaning played a crucial role in this project. We identified duplicate entries, missing values, and imbalances within the dataset. Three versions of the dataset were created:

1. **Multiple CIDs per drug:** Retained multiple entries for drugs that had different PubChem CIDs.
2. **Single CID per drug:** Dropped duplicates and kept only one CID per drug to reduce redundancy and improve model efficiency.
3. **Balanced Class Multiple CIDs per drug:** Filtered down the majority class of moderate interactions to be more even with the number of minor and major interactions, while still keeping multiple CIDs per drug.

In addition, we used the Synthetic Minority Over-sampling Technique to handle class imbalance in the Single CID per drug dataset, ensuring that the models were trained on a balanced dataset for better performance.

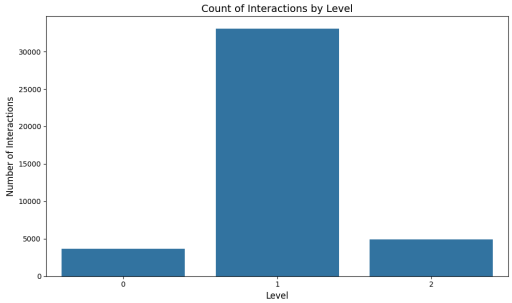



**Figure 1:** Bar showing the imbalance across "Minor," "Moderate," and "Major" interaction levels

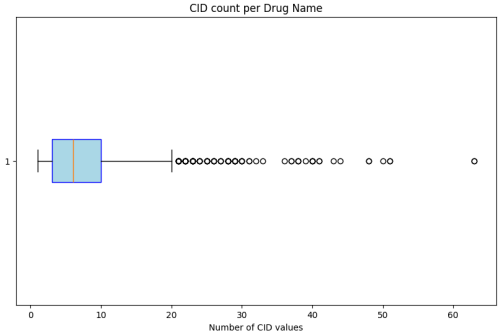



**Figure 2:** Box plot of the count of CID values per Drug Name.

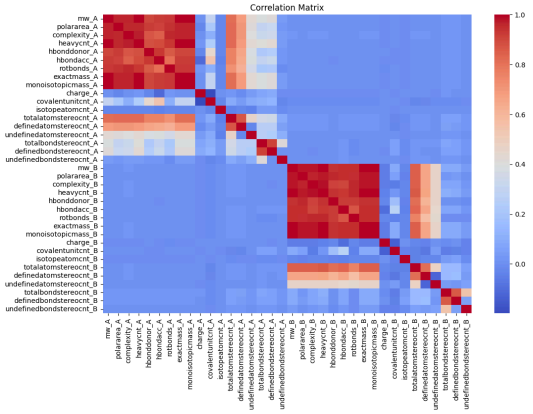



**Figure 3:** Correlation Matrix: Heatmap of feature correlations.

## 3. Exploratory Data Analysis (EDA)

EDA was performed to gain insights into the dataset. Key findings included:

- Class Distribution: The interaction levels were imbalanced, with the majority of interactions categorized as "Moderate" and fewer instances of "Minor" and “Major” interactions.
- Correlation Analysis: A high correlation between group A and group B drugs was observed, which might influence the models’ predictions.
- CID Distribution: We analyzed the number of unique CIDs per drug to ensure that the data was representative and well-distributed.

## 4. Model Training and Evaluation

- Multiple CIDs per drug: Obtained an accuracy score of .991 with Random Forest, .899 with XGBoost, and .866 with Light GBM. Due to the number of duplicate drugs, it is likely that the model was memorizing the interaction and thus able to obtain .991 accuracy with Random Forest.
- Single CID per drug: Obtained an accuracy of .886 with Random Forest, .901 with XGBoost, and .906 with LightGBM. All models struggled to predict the minority classes accurately.
- Balanced Class Multiple CIDs per drug: Obtained an accuracy score of .985 with Random Forest, .8765 with XGBoost, and .9823 with LightGBM. Also likely achieved such high accuracy due to memorization from the duplicate drugs.

**Feature Importance:** The following features consistently showed high importance across models:

- Polararea
- Complexity
- Monoisotopic Mass

## 5. Conclusion

- Machine learning models demonstrated strong predictive capabilities for DDIs, especially when using datasets with multiple CIDs per drug.
- The models leveraged key molecular descriptors, with polararea, complexity, and monoisotopic mass consistently ranking as top predictors.
- High accuracy in some cases was attributed to potential memorization due to duplicate entries in the dataset.

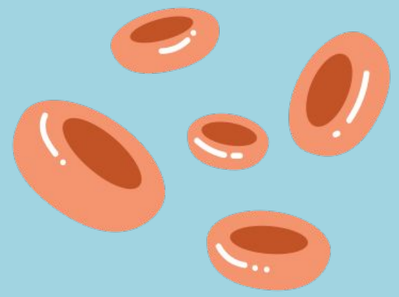
## 6. Expanding scope

- Refining Dataset
    - Addressing biases introduced by multiple CIDs per drug to improve model generalizability.
    - Investigating methods to consolidate or average molecular descriptors for drugs with multiple CIDs.
- Improving Model Robustness
    - Incorporating ensemble methods and hypermater tuning to improve performance in minority classes.
    - Extending the analysis to include a broader range of drug categories beyond alimentary tract and metabolism drugs.
- Adding Features:
    - Include pharmacokinetics or pharmacodynamics data to better understand drug behavior and to capture the underlying phenomena in the interaction mechanisms.
    - Add features related to biological targets, such as protein binding profiles and receptor affinities, to provide a biological context for DDIs.

## 7. Final Thoughts

By leveraging molecular descriptors and machine learning models, this project lays the groundwork for a data-driven approach to predict drug-drug interactions (DDIs), which has significant potential for enhancing drug safety in clinical practice. This project was not only a technical challenge but also an enriching experience that gave us insights into how data in real-life scenarios is messy and complex.

# Balancium RX - Drug-Drug Interaction Checker




Chehak Arora

Marius Nwobi

Franco Valdes

Andrew Normandin

David Rogers

## Problem:

- Pharmaceutical development is becoming more complex.
- Risk of harmful drug - drug interactions
  - Can reduce drug efficacy.
  - Can increase toxicity.
  - Can create unforeseen side effects.

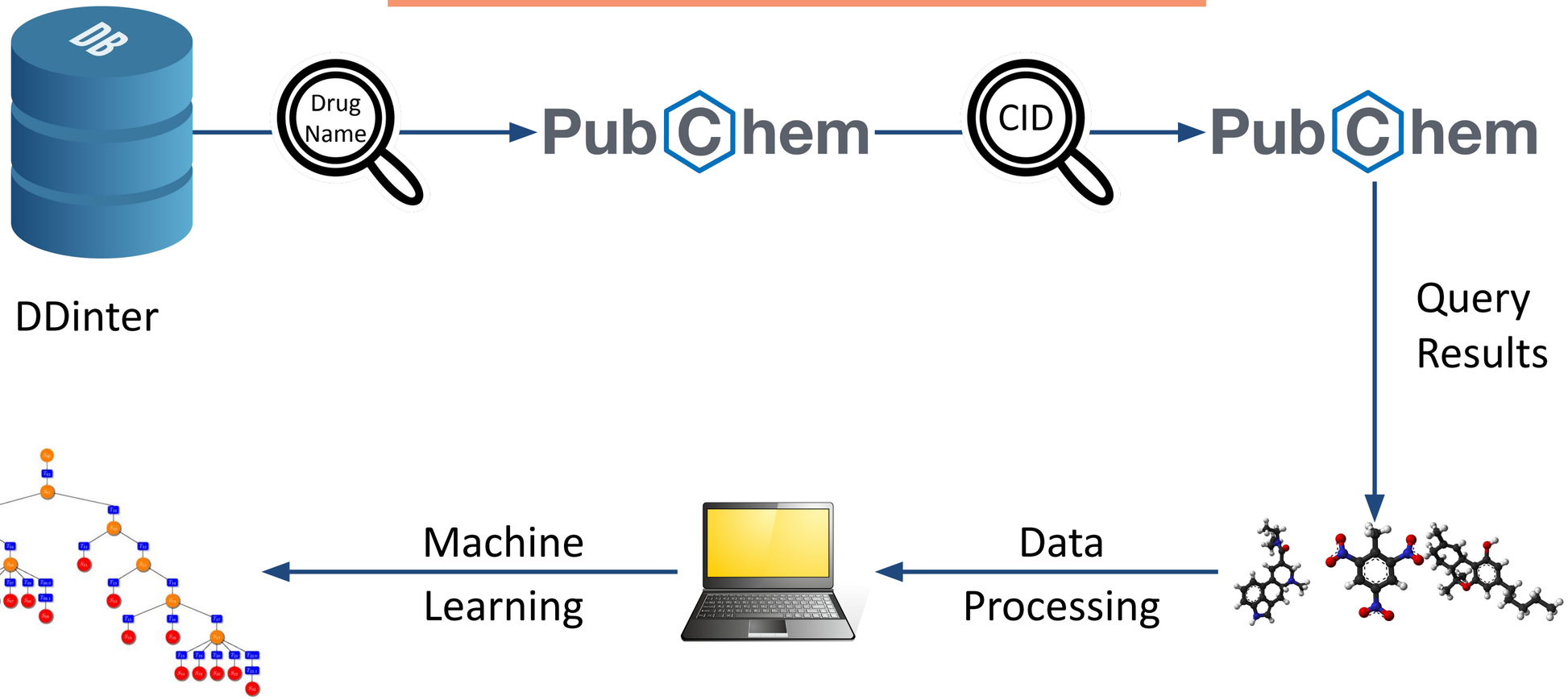
## Data Set Sources:

- **Original DDI Dataset:**
  - Interactions involving alimentary tract and metabolism drugs.
  - Drug names mapped to multiple PubChem CIDs (some drugs have more than one CID).
  - Molecular descriptors and chemical properties retrieved from PubChem.
- **One-CID-Per-Drug Dataset:**
  - Derived by filtering the original dataset to retain a single representative CID per drug.
  - Ensures one unique ID per drug for cleaner analysis and comparison.

## Final Data Sets:

- Both datasets include chemical properties as predictors.
- Interaction labels used as targets for machine learning models.

# Workflow Diagram



# Data Extraction & Cleaning

**1. Raw Data**: The initial dataset was obtained from DDInter, mapping drug-drug interactions (DDI). The data included drug names and associated interaction labels.

**Additional Data Sources**:

- **Drug CIDs**: A mapping of drug names to their respective PubChem Compound Identifiers (CIDs).
- **PubChem Data**: Chemical and molecular descriptors for drugs were retrieved using CIDs from PubChem

**2. Handling Multiple CIDs**

- **Issue**: Multiple CIDs exist for the same drug, each representing different molecular forms or chemical properties (e.g., Naltrexone with 13 CIDs).
- **Decision**: To leave multiple CIDs unprocessed due to time constraints and the lack of more granular data. This decision was made to allow for faster project progression.

# Data Extraction & Cleaning

**3. Data Augmentation via Multiple CIDs**

- **Augmentation**: Multiple CIDs per drug naturally expand the dataset by providing additional examples, especially for underrepresented drugs.
- **Data Balance**: This approach helps mitigate class imbalances by increasing the presence of drugs with fewer associated data points, enhancing model performance.

**4. Current Limitations**

- **Bias & Redundancy**: The dataset may have biases due to overrepresentation of drugs with more CIDs, and redundancy due to similar chemical properties across multiple CIDs.
- **Future Refinement**: The approach will be revisited if more data becomes available, considering potential methods like averaging descriptors or dimensionality reduction.
- **Lack of Relevant Features:** We are relying only on molecular descriptors so far to train the machine learning models. This is not optimal, since we could also leverage pharmacokinetic and pharmacodynamic information as well that is very relevant for understanding the underlying phenomena.

# Data Set Challenges

**Multiple CID Entries per Drug Name:**

- Each drug name in the PubChem database can have multiple CID entries.
- This can result in row duplication when analyzing drug-drug interactions.

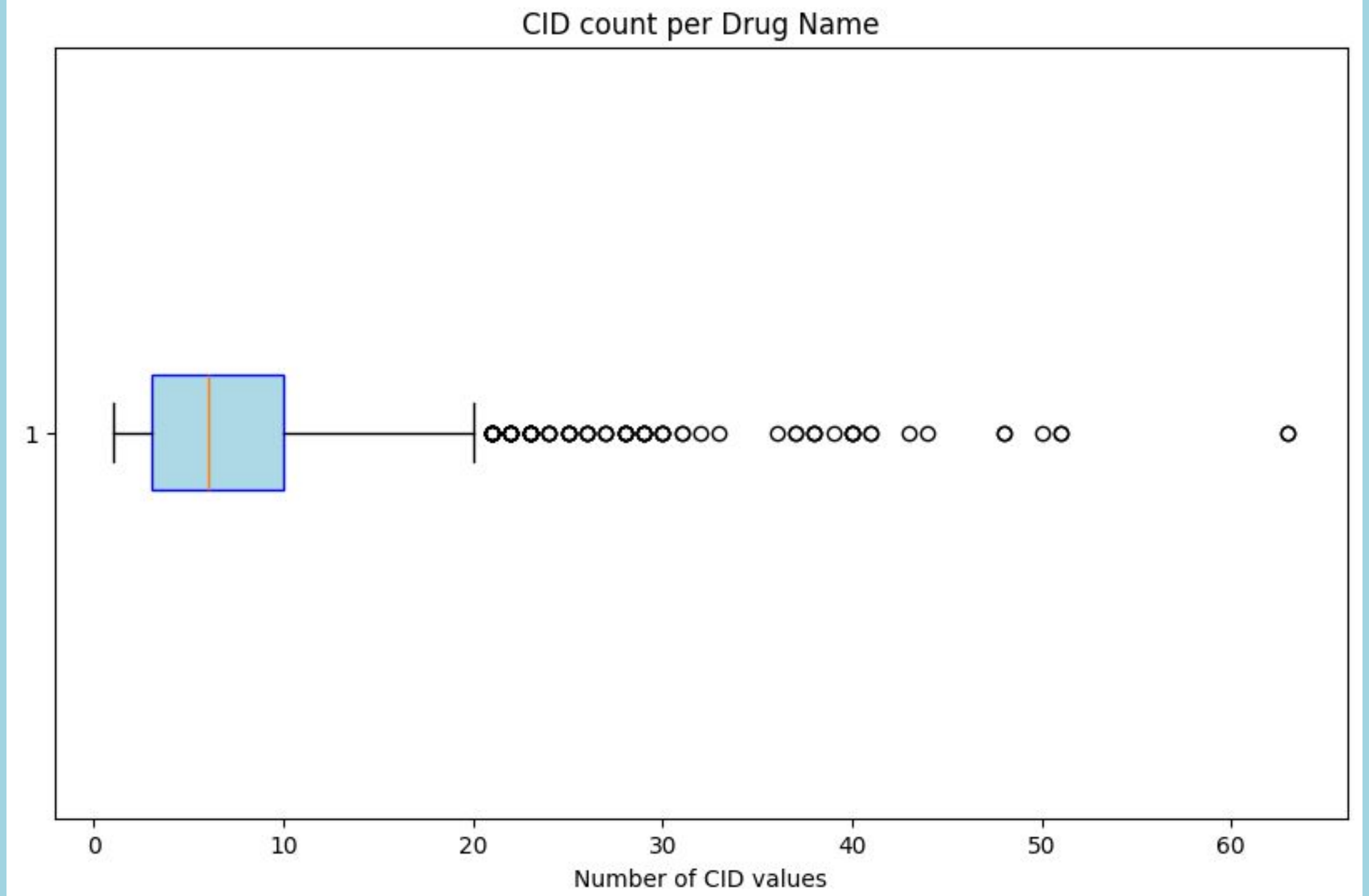
**Class Imbalance for the Level Label:**

- The Level label exhibits significant class imbalance, which could affect model performance and require appropriate handling during training.

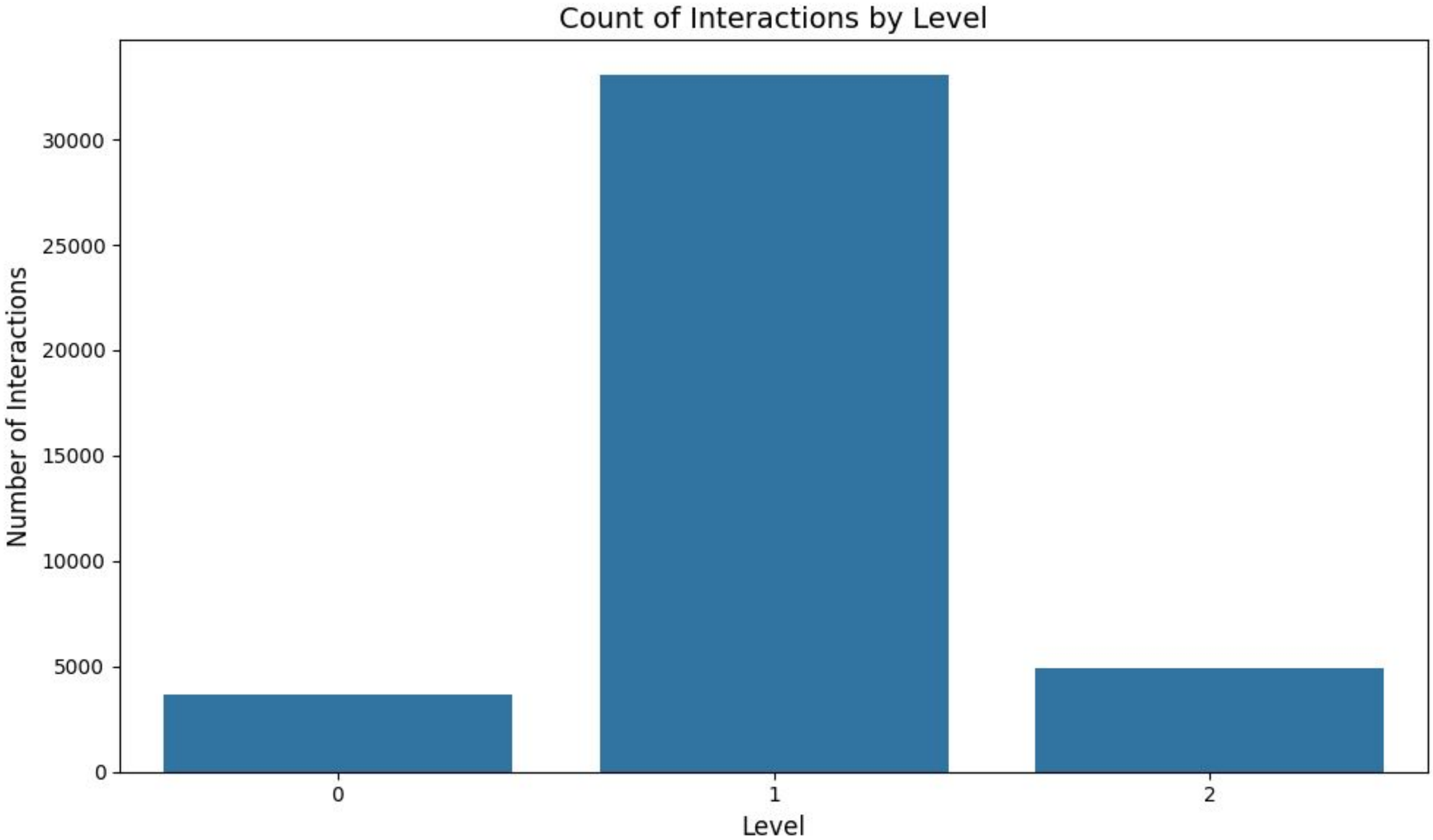
**Limited Features for Analysis:**

- Currently focusing on molecular descriptors, which may not fully capture the underlying phenomena in drug-drug interactions.
- Additional features or domain-specific data may be needed for better predictive accuracy.

# There are many unique CID values per drug name, this can bias the training process

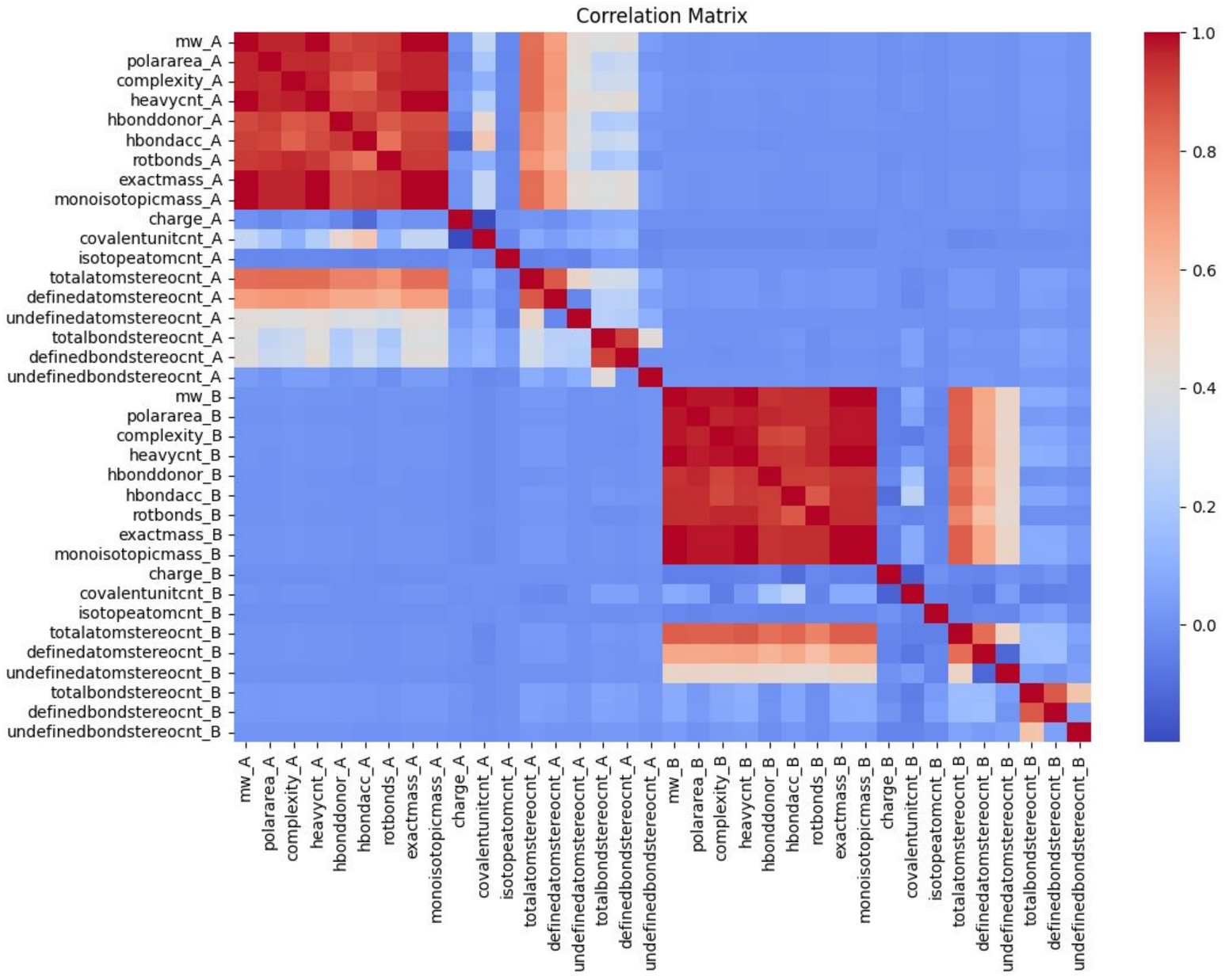


There is a class imbalance, having Level 1 ‘Minor’ Interaction as the most represented value

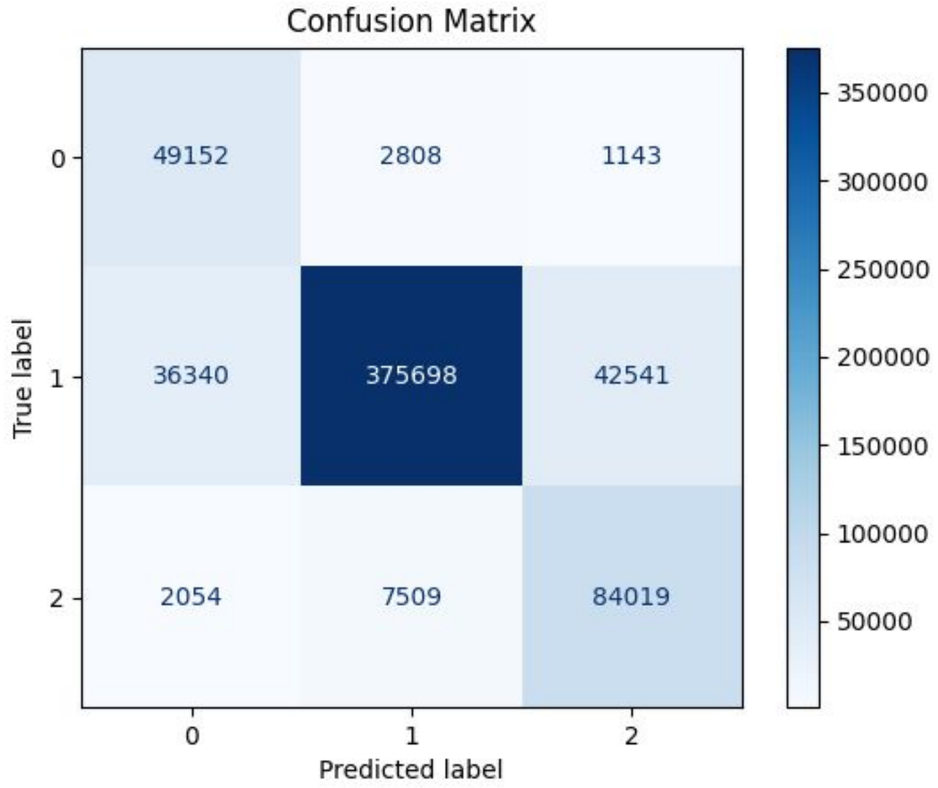


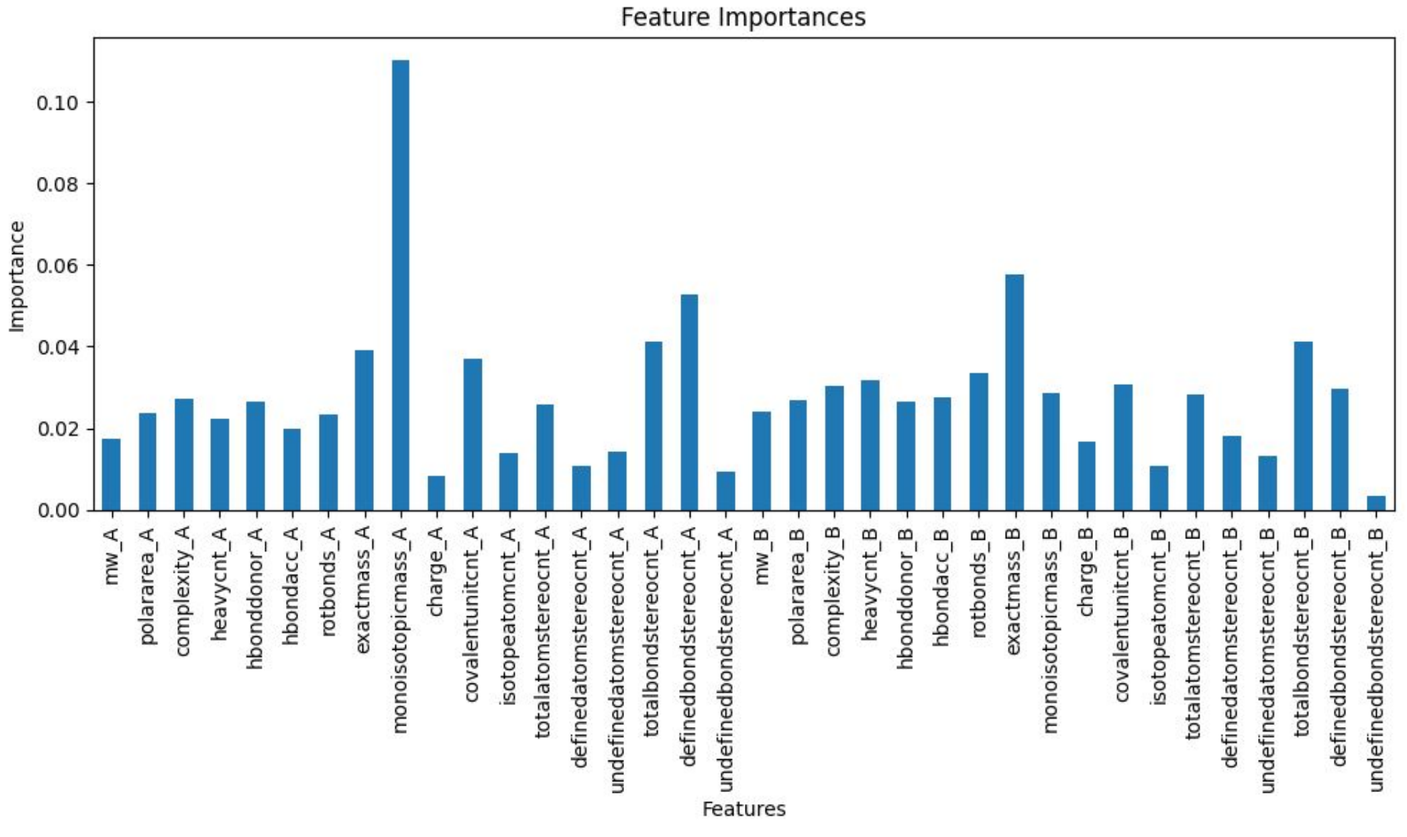
There is heavy multicollinearity between the columns for each drug,

but not between drugs

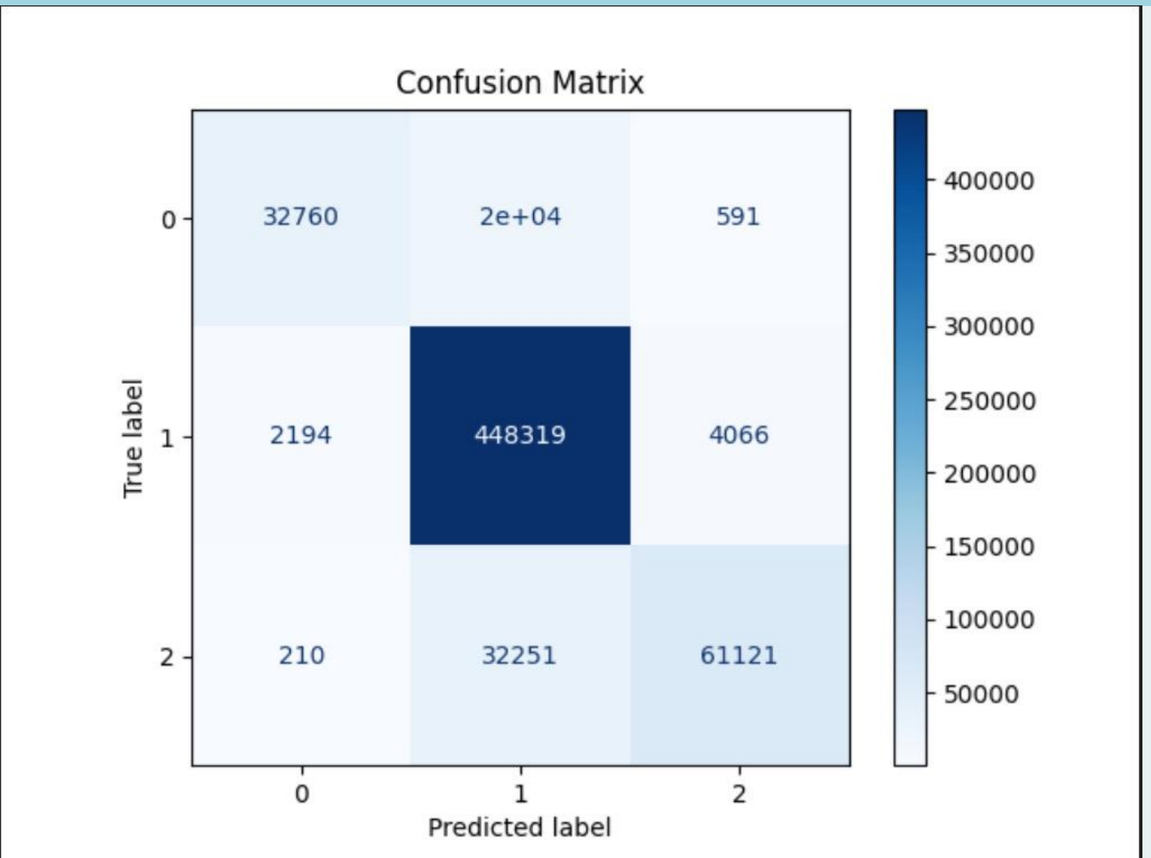


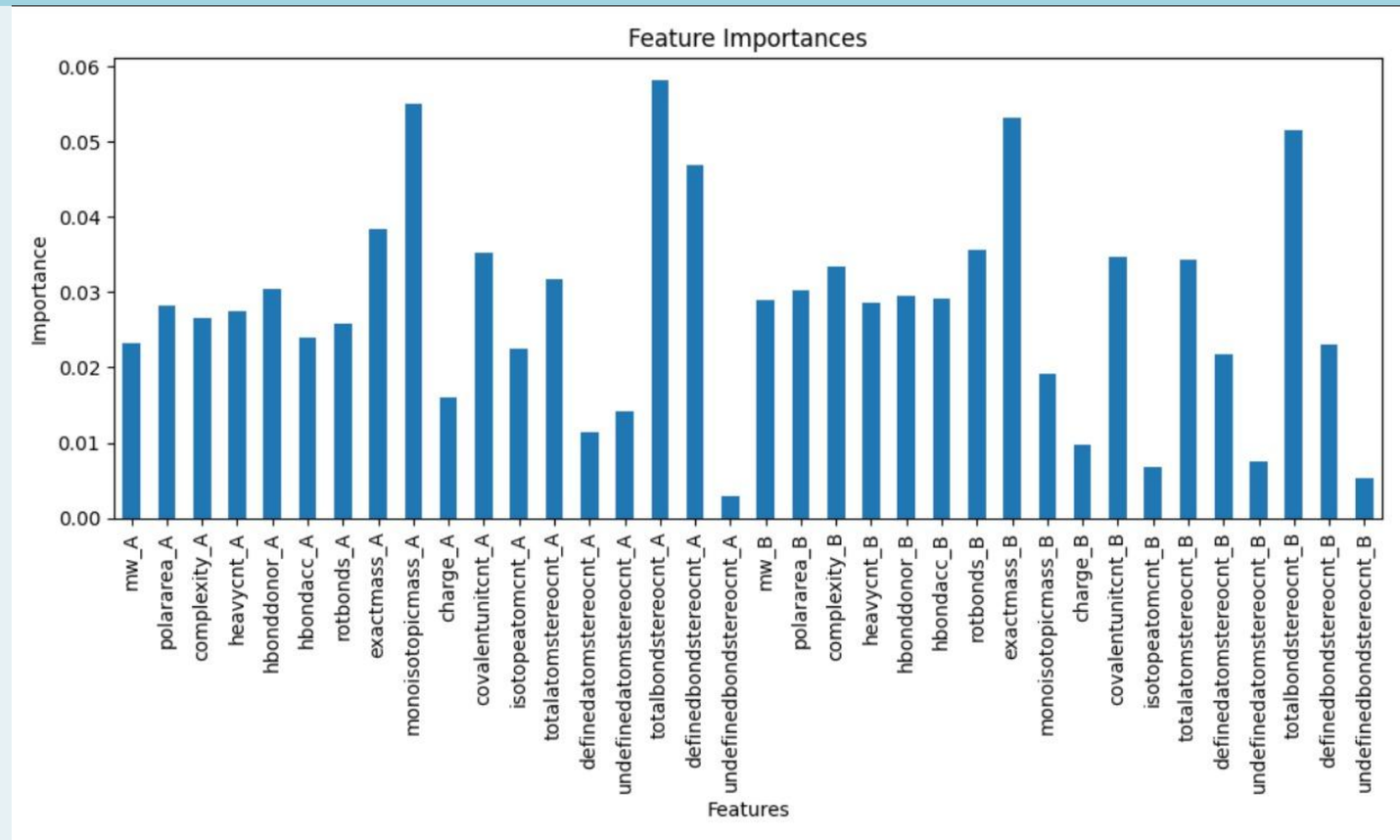
# XGBoost with Weights



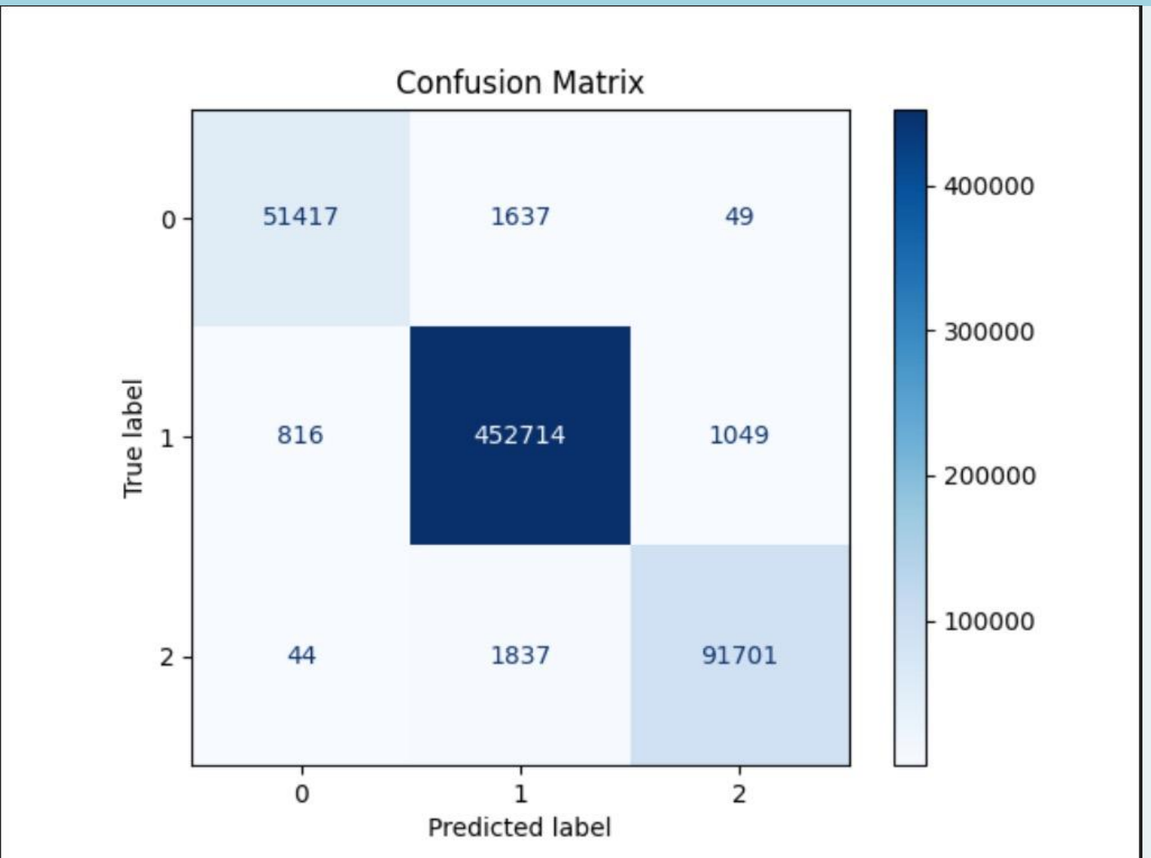



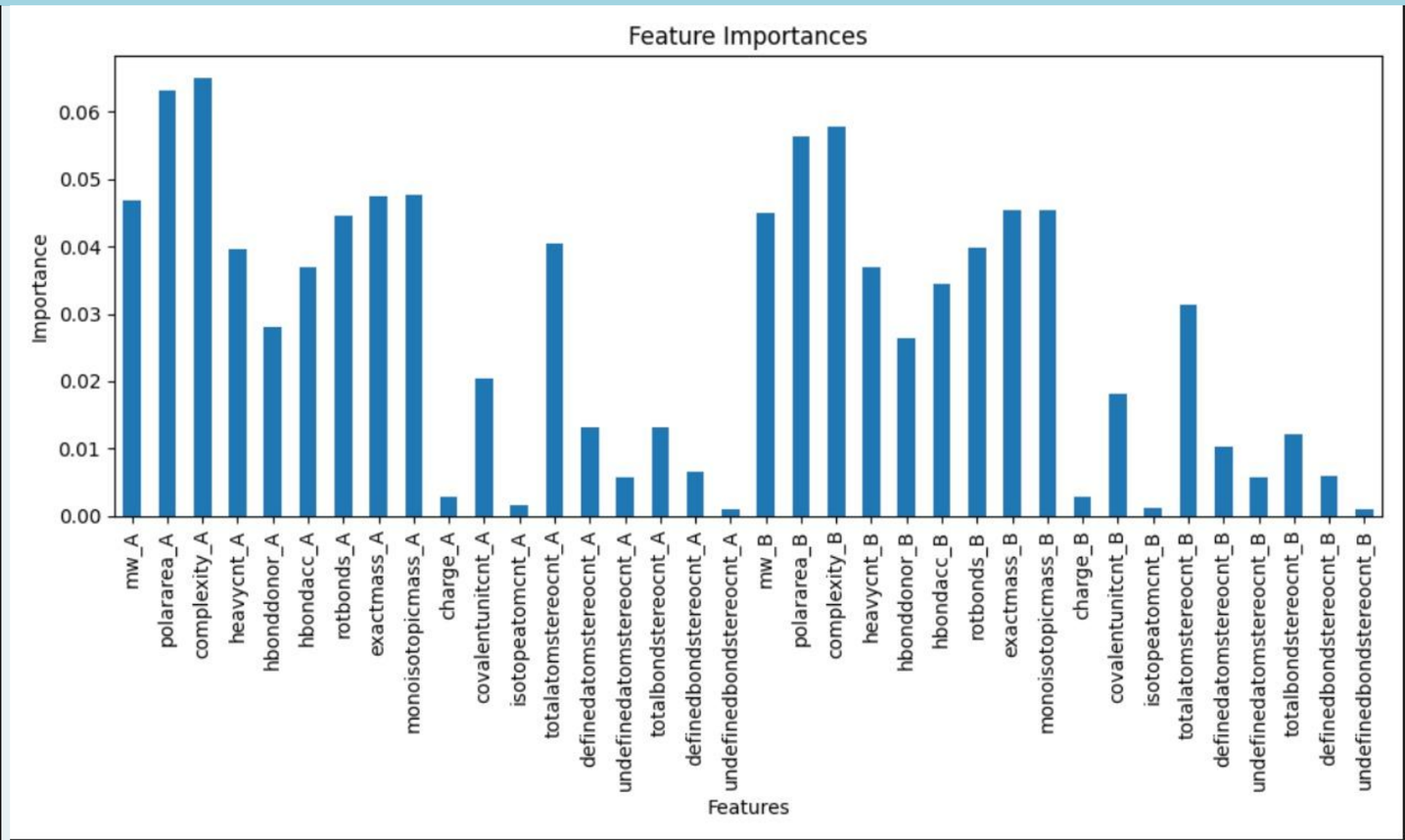
# XGBoost without Weights



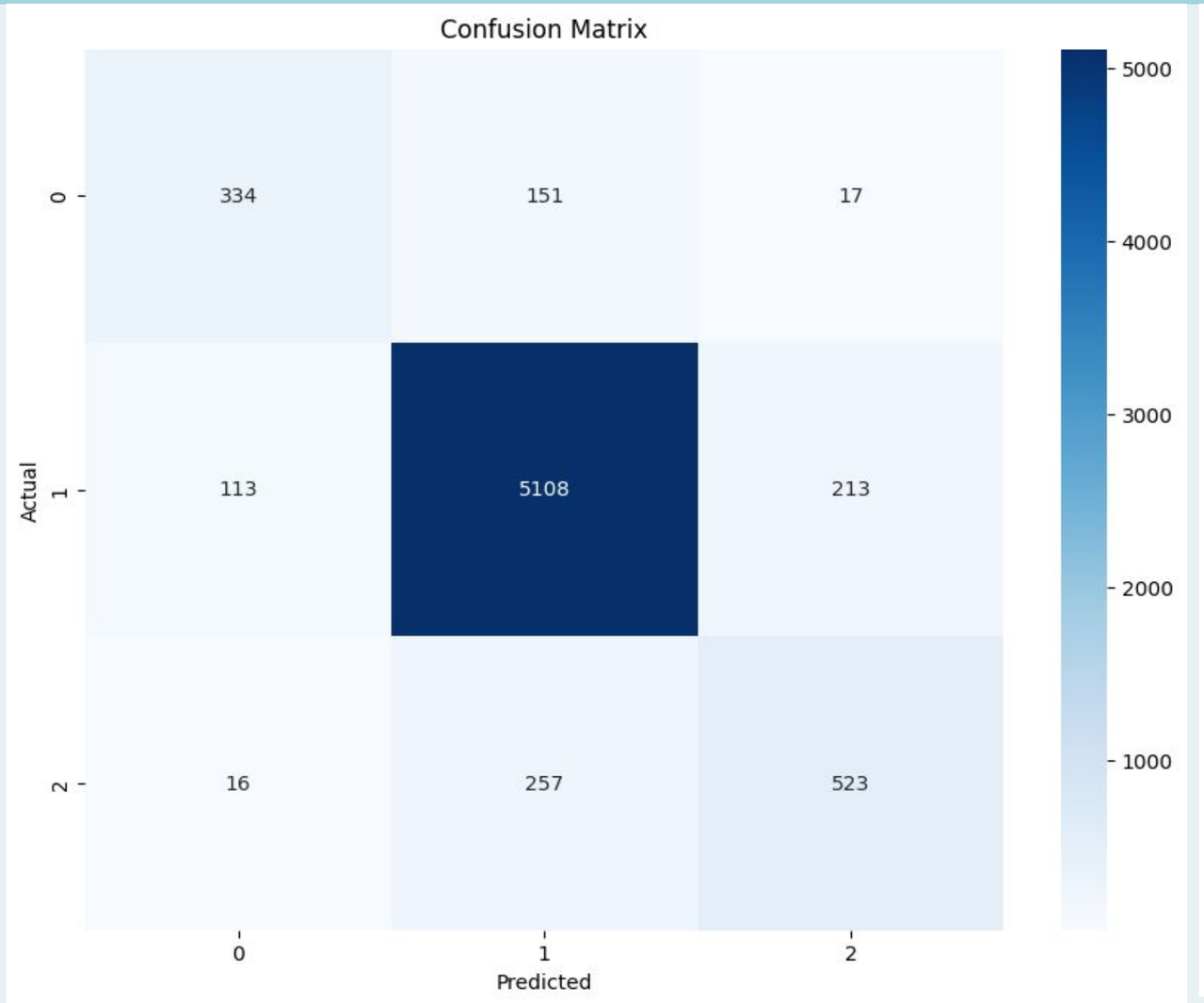



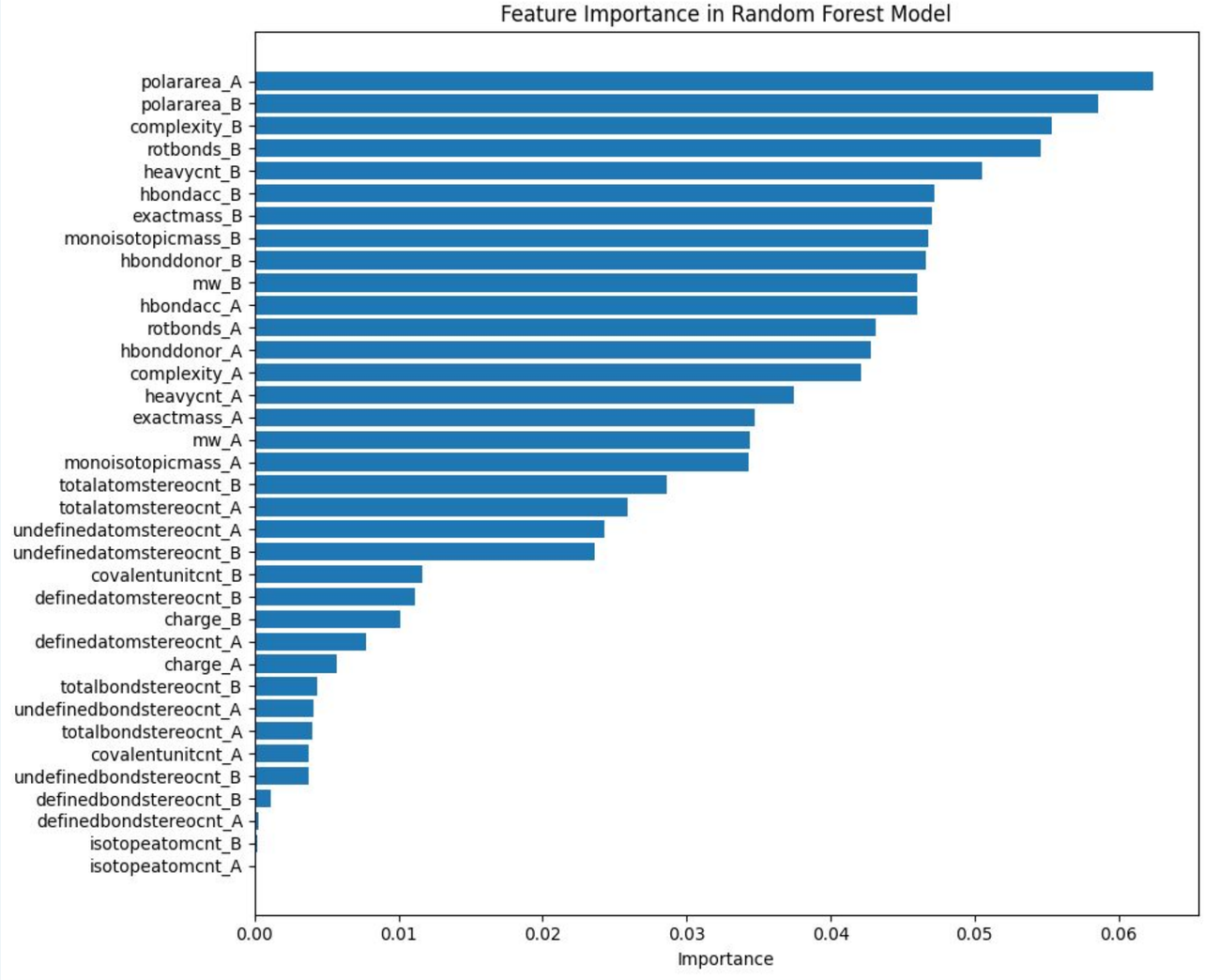
# Random Forest



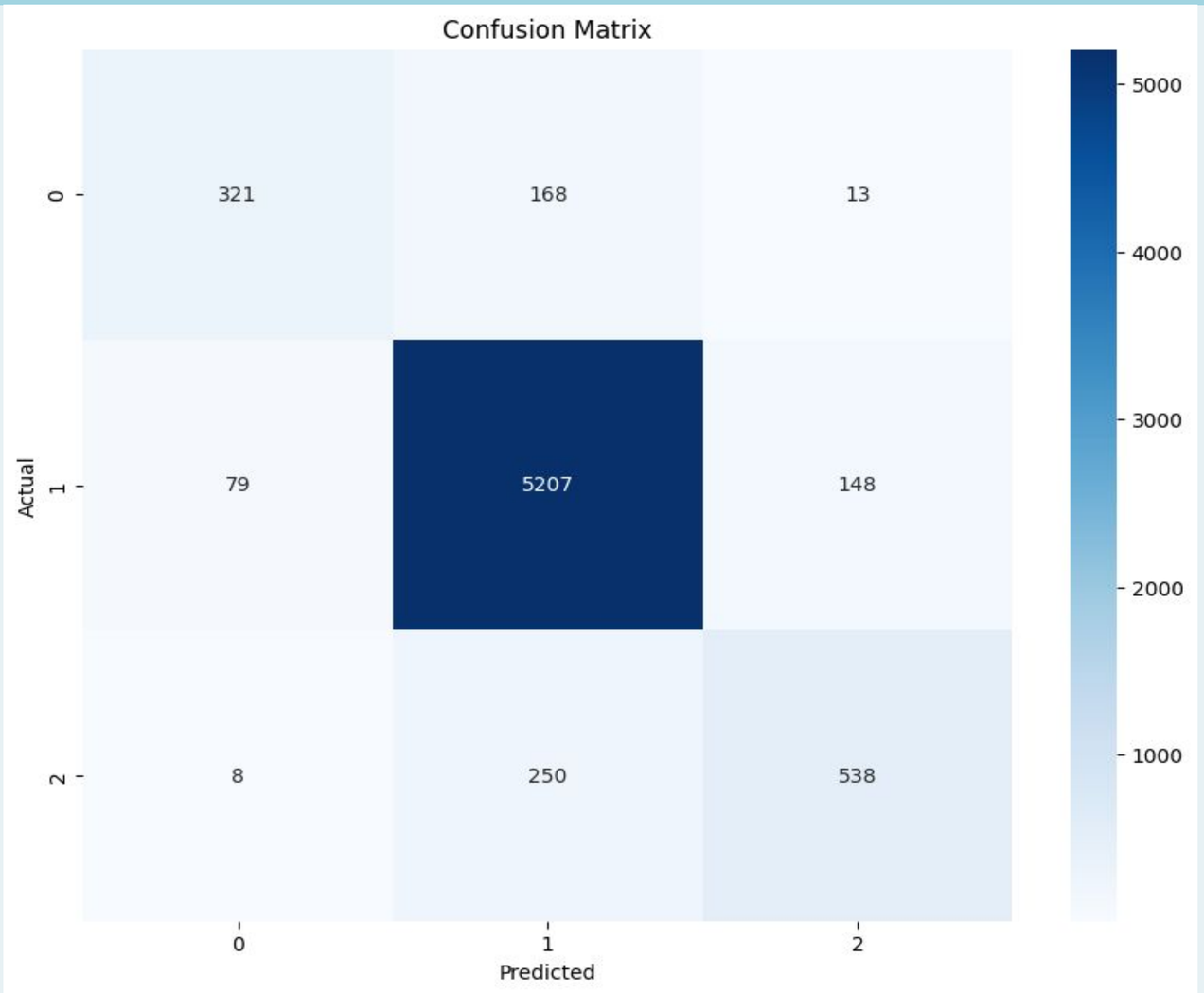



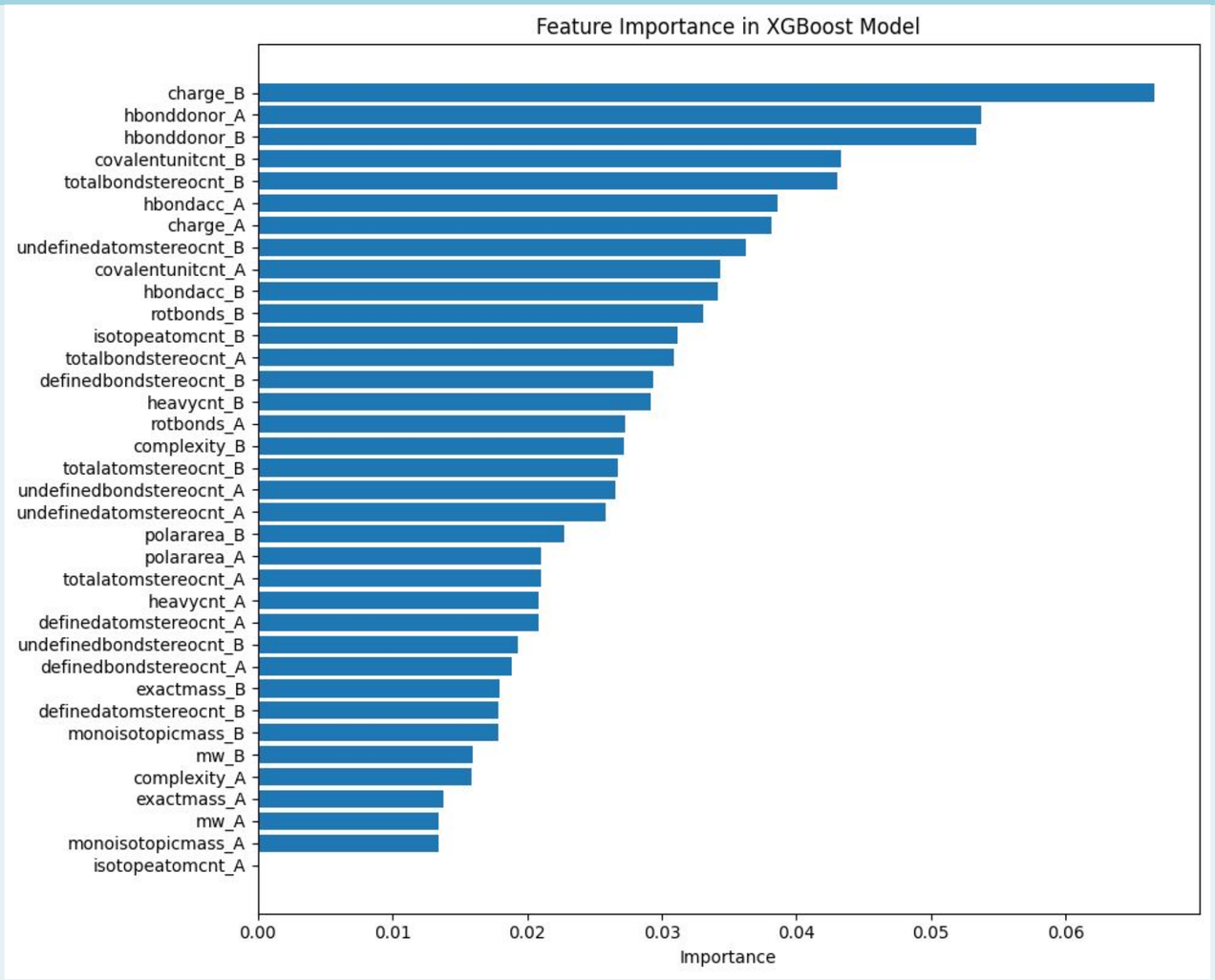
# Random Forest W/O Duplicates



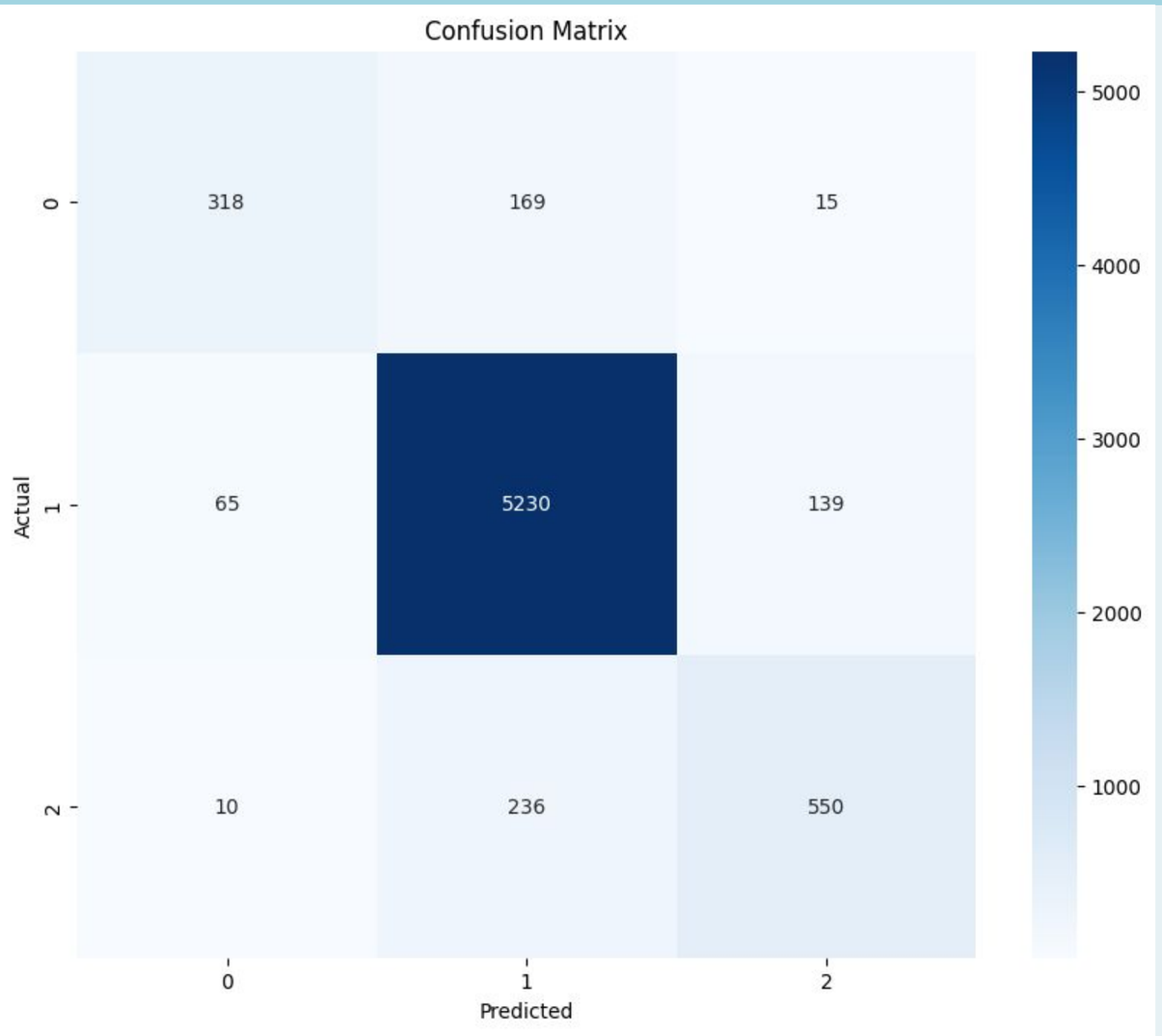



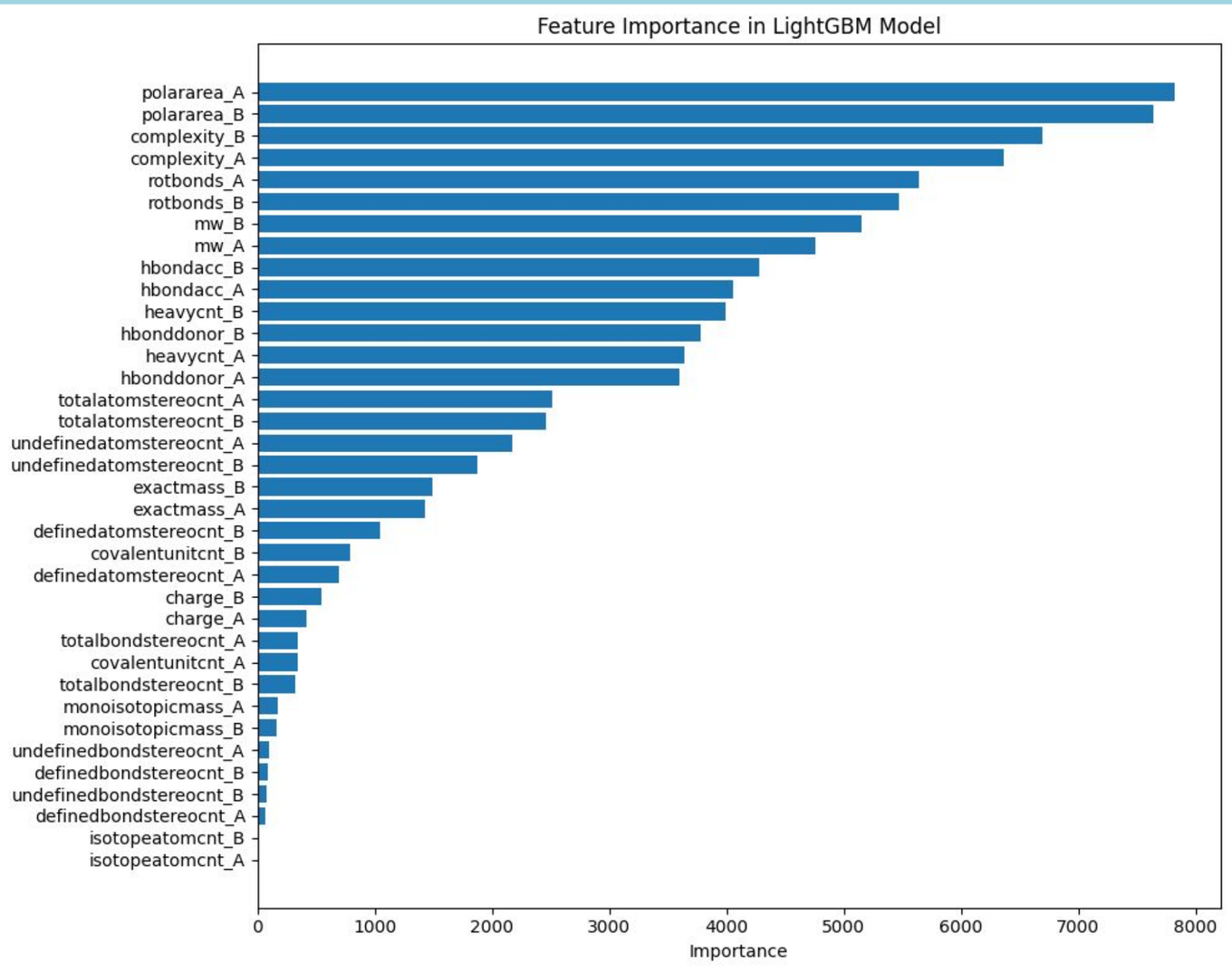
# XGBoost W/O Duplicates



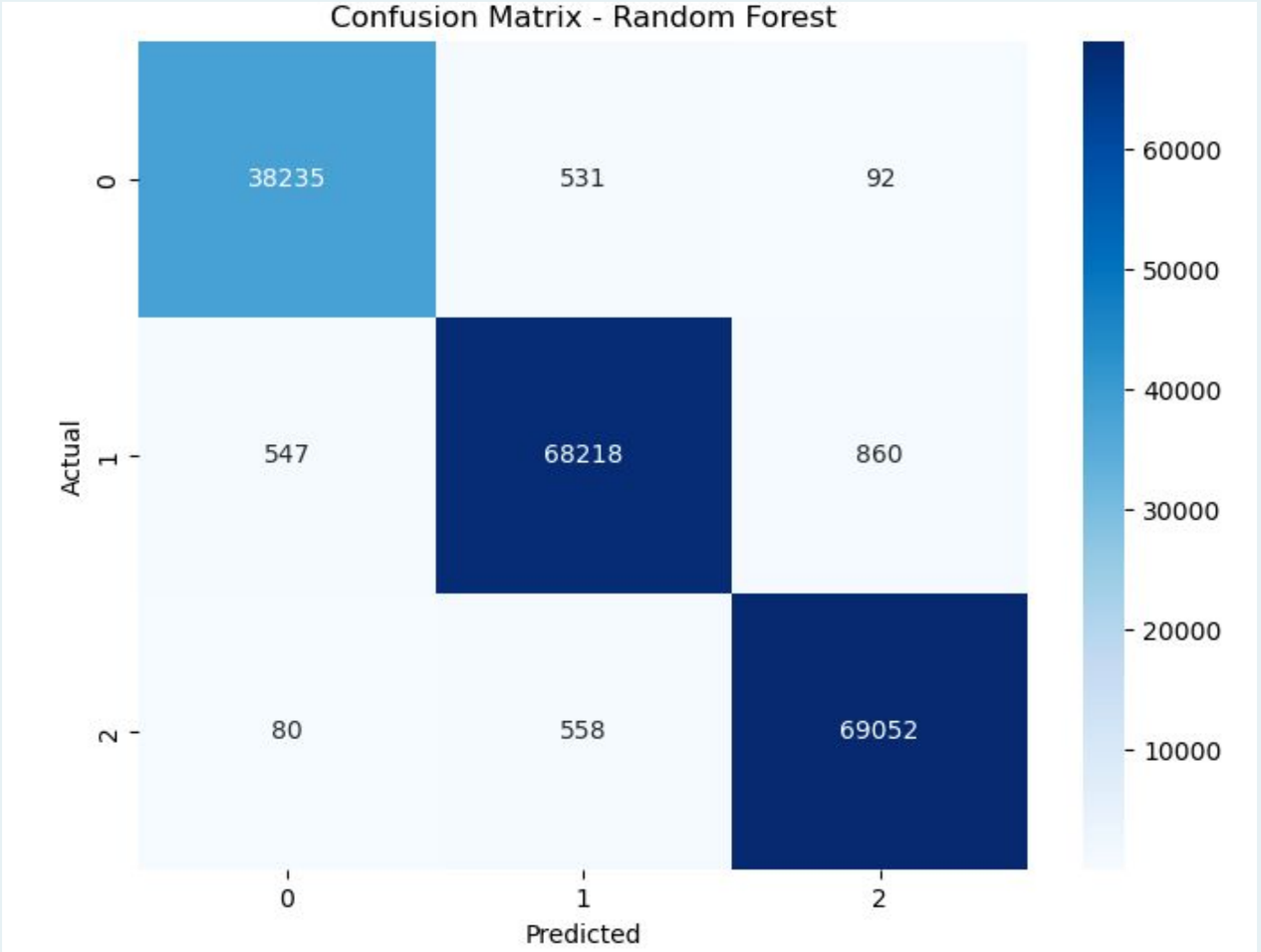



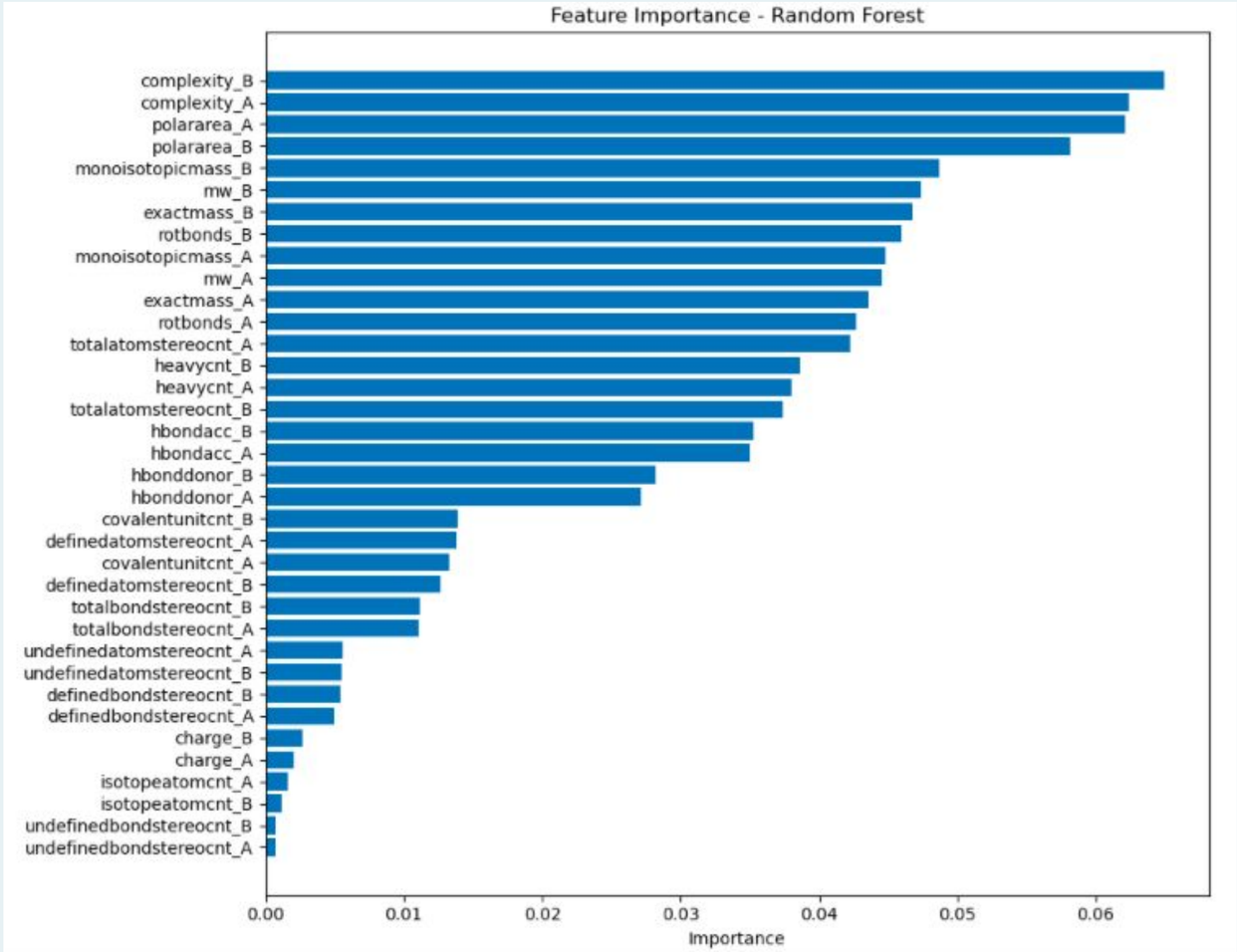
# LightGBM W/O Duplicates



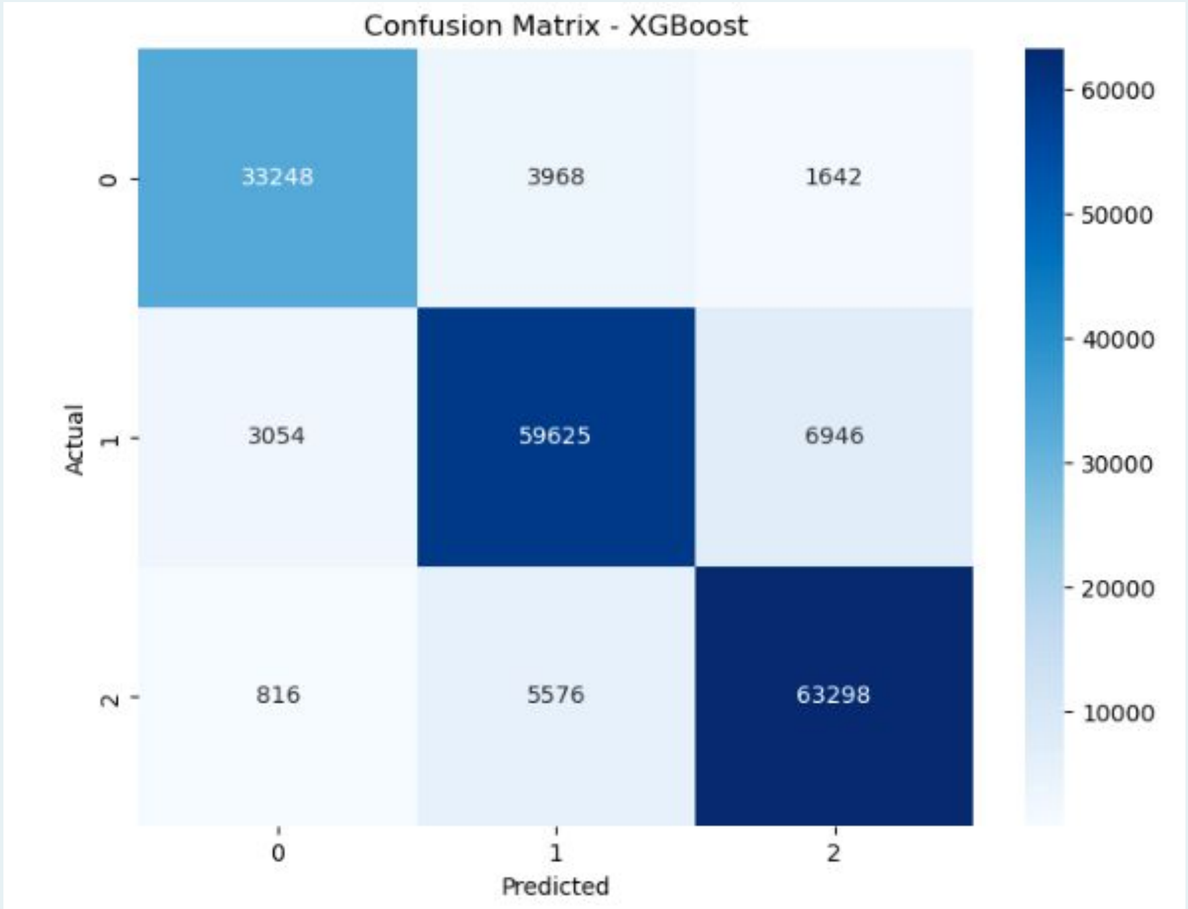



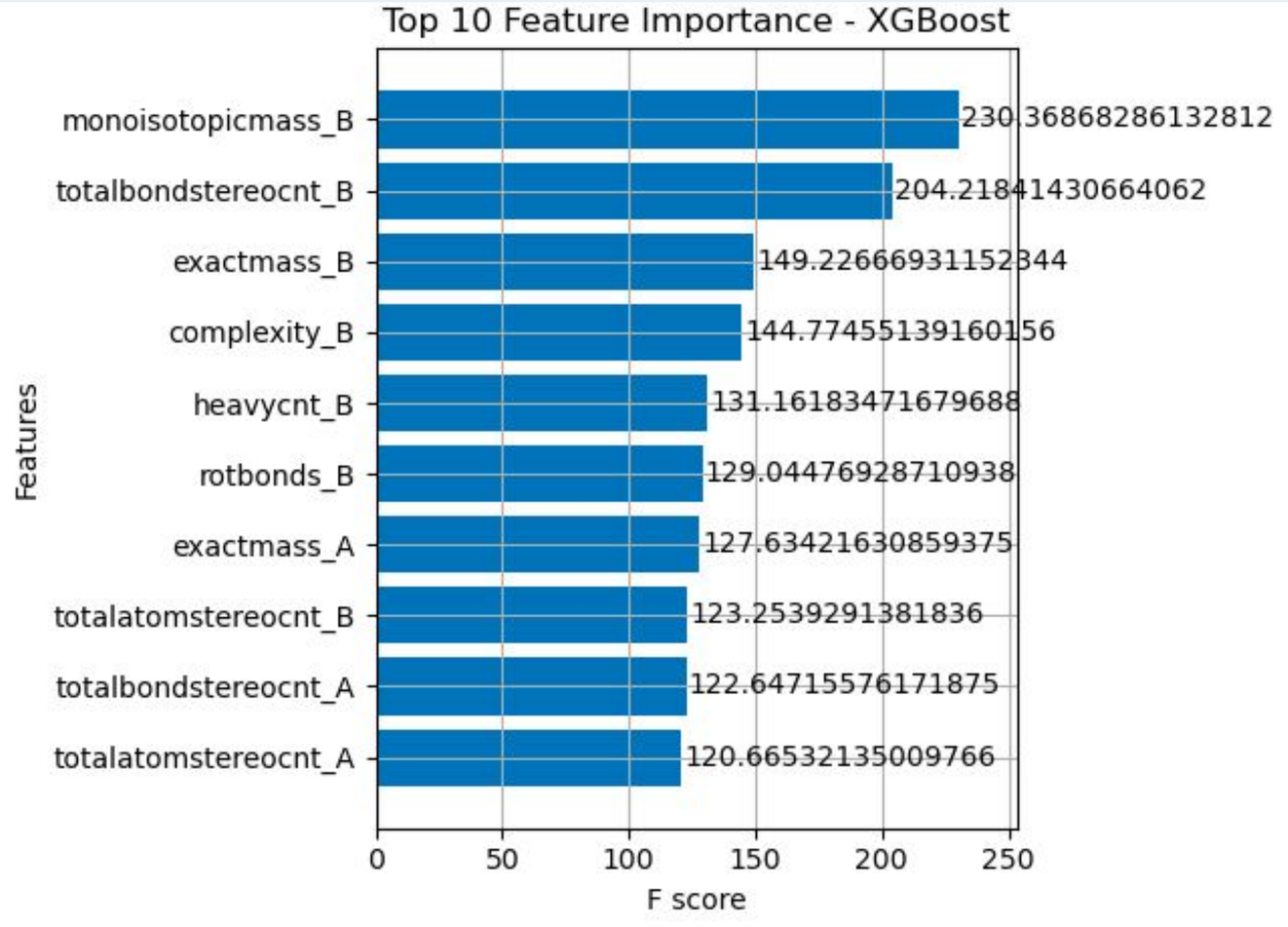
# Random Forest - Class Manipulation



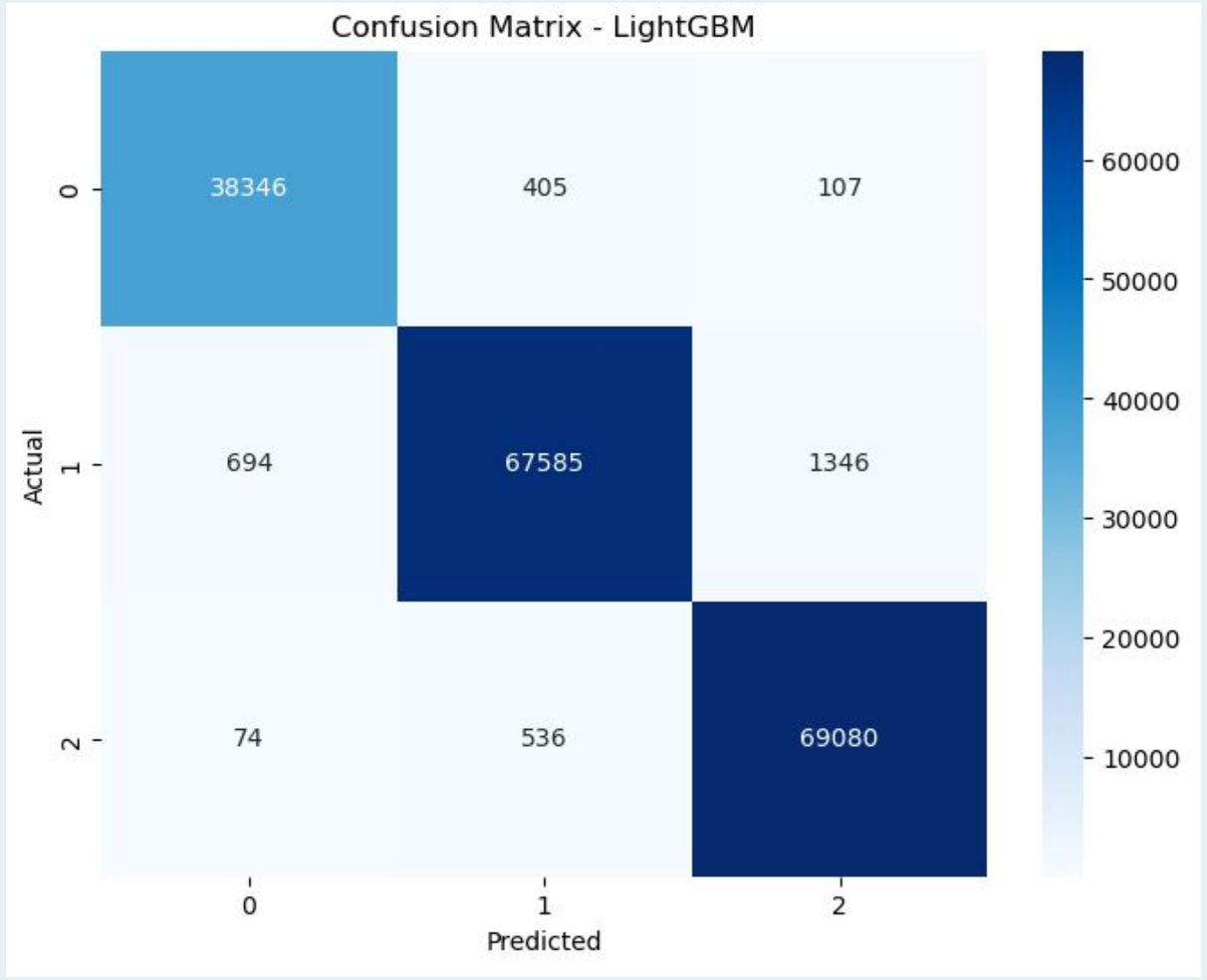



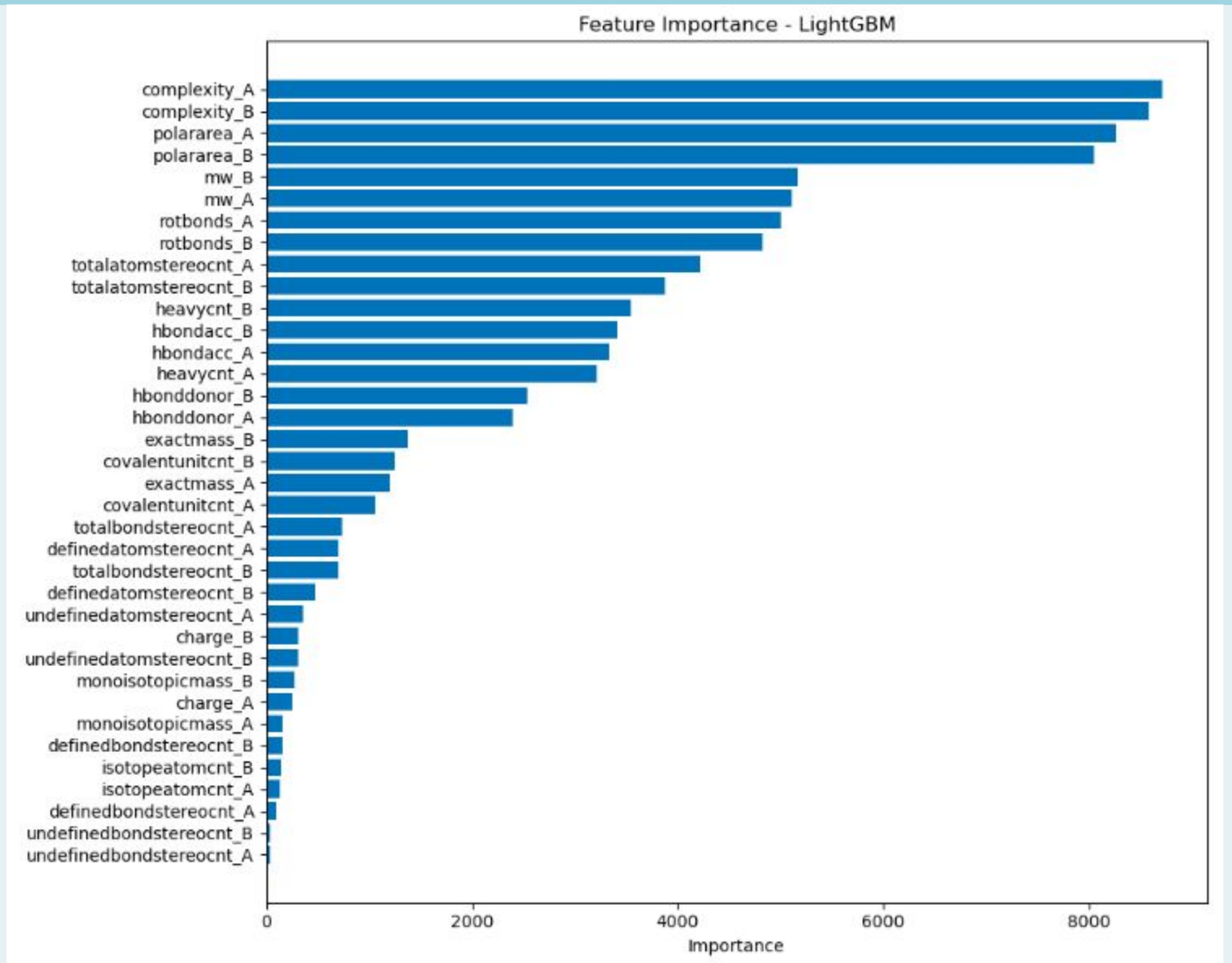
# XGBoost - Class Manipulation



# LightGBM - Class Manipulation

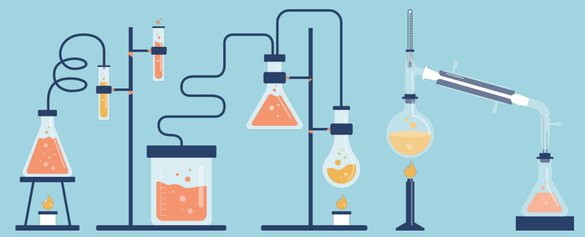


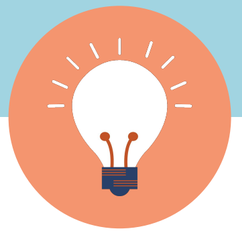
# Method Accuracy Comparison

| Model | W/ Duplicates | W/O Duplicates | Class Manipulation |
|---|---|---|---|
| Random Forest | 0.991 | 0.886 | 0.9850 |
| XGBoost | 0.899 | 0.901 | 0.8765 |
| Light GBM | 0.866 | 0.9058 | 0.9823 |

# Learning outcomes



- [x] Handling redundancy in datasets
- [x] Addressing Class Imbalances
- [x] Analyzed how dataset preprocessing choices affect model accuracy



Gained an appreciation for the complexities of working with biological and chemical datasets, such as duplicate entries and imbalanced classes.

# Thank you!