Predicting Drug-Drug Interactions for Repurposed Drugs

Table of Contents

Abstract:

Accurately predicting drug-drug interactions (DDIs) is a challenge, in the healthcare industry especially as drug repurposing becomes more prevalent. Repurposed medications provide cost treatment options. Also pose the risk of unexpected interactions due to their novel applications. Traditional methods for identifying DDIs are often limited by known interactions leaving a knowledge gap for repurposed drugs. This project introduces a machine learning framework aimed at predicting DDIs. It utilizes datasets from drugs.com along with chemical property data to enhance the prediction model. Our approach combines machine learning techniques like forests. Support vector machines with network pharmacology methods to create an accurate and understandable predictive model. The model has been tested on a mix of user generated content and structured chemical data achieving performance with an AUC score of 0.98. This demonstrates its ability to effectively distinguish between the presence and absence of interactions with reliability. The study's findings suggest that our model could significantly enhance DDI prediction in the context of drug repurposing ultimately improving safety and treatment effectiveness. Implementing such a model could serve as a foundation, for decision support systems revolutionizing identification of adverse drug interactions.
This research does not bring a valuable addition, to the field of pharmacology but also establishes a foundation, for incorporating data driven methods into drug safety studies in the future.

Introduction:

The process of developing drugs in the pharmaceutical industry is expensive and time consuming often taking, then ten years from start to market. However, an efficient and cost-effective approach called drug repurposing has gained popularity in years. This involves finding uses for existing drugs. It does not bring back the usefulness of drugs that may have been overlooked due to effectiveness for their original purpose but also offers hope for treating rare and orphan diseases that may not be a priority for large scale drug development programs. However repurposing drugs for uses raises concerns about interactions, between medications (drug-drug interactions or DDIs). These interactions can alter the effectiveness of one drug when taken with another leading to reactions or reduced therapeutic benefits.

The accurate prediction of DDIs is extremely important. With an aging population and the common practice of taking medications at once (polypharmacy) there is an increased risk of DDIs which're one of the main causes of health issues and deaths related to medication use. Additionally, DDIs contribute significantly to healthcare costs by necessitating doctor visits, hospitalizations and prolonged treatments. Hence it is crucial to address the issue of drug-drug interactions (DDIs), in the context of repurposing drugs.

DDIs can manifest in ways, such as affecting drug absorption or increasing the likelihood of side effects. In cases where drugs are repurposed for conditions, they were not originally intended to treat the risk of DDIs becomes higher. This emphasizes the importance of monitoring and predictive analytics.

This project introduces a framework based on machine learning that aims to anticipate and identify DDIs specifically related to drug repurposing. By utilizing techniques and a comprehensive collection of pharmacological data this model seeks to predict how drugs may interact when used together. Essentially it functions as a pattern detection system that learns from interaction data to forecast future DDIs. As a result, it becomes a tool for clinicians in optimizing strategies.

The subsequent sections of this paper delve into aspects of this research. The section titled 'Related Work' examines existing studies on DDI prediction and drug repurposing highlighting the uniqueness and necessity of our approach. The section named 'Design' explains the methodology behind our proposed solution covering everything, from data curation to implementation.

The section titled 'Evaluation / Analysis / Results breaks, down the effectiveness and precision of the model providing an assessment of its ability to make predictions. Lastly although not mandatory sections like 'Conclusions 'Future Works and 'Appendices bring the study to a close suggest avenues, for research. Gather additional materials that support the main analysis.

Previous research:

The investigation of drug-drug interactions (DDIs) holds a role, in studies, driven by the increasing complexity of medication regimens and the continuous introduction of new pharmaceuticals. Over time identifying and comprehending DDIs have been central to ensuring safety and optimizing effectiveness. Noteworthy studies in this field have ranged from observations and case reports to reviews and meta-analyses all providing valuable insights into the nature and implications of DDIs.

In years computational methods have transformed the prediction and analysis of DDIs. Machine learning models, utilizing datasets from health records (EHRs) adverse event reporting systems and drug databases have shown significant potential. These models, which include decision trees, neural networks and support vector machines among others have been utilized to identify and forecast interactions based on data. Another noteworthy advancement is network pharmacology, a methodology that constructs networks of drug target interactions. This approach offers a

visualized tool to explore pathways through which drugs can interact adding a new dimension to our understanding of DDIs.

Despite these advancements and methodological progressions mentioned above there remains a gap, in the literature concerning drug repurposing. Repurposing drugs, which involves finding uses for existing medications has become an increasingly attractive approach, in the field of pharmaceutical development. However, dealing with drug-drug interactions (DDIs) in this context presents challenges. While repurposed drugs may have documented interaction profiles within their applications these profiles might not accurately represent potential interactions in new therapeutic contexts. The current body of research often falls short when it comes to predicting interactions creating a knowledge gap and possible risks in clinical use.

To bridge this gap this project aims to develop a machine learning model specifically designed for the challenges associated with drug repurposing. By incorporating a range of data sources including user generated content and detailed information on drug properties our model aims to capture a spectrum of potential interactions compared to traditional models. What sets our approach apart is that we do not rely on known DDI databases but also consider the nuances involved in repurposing drugs. Our goal is to create a tool that can predict not established interactions but those that might arise in new contexts, which leads to a significant advancement over existing models.

To sum up this, current research has laid a foundation for predicting DDIs there is still a need, for models that specifically address the complexities of drug repurposing.

This project aims to address this gap by offering an evidence-based method for predicting drug-drug interactions (DDIs). By doing it contributes to the changing landscape of pharmaceutical research and highlights the growing significance of computational techniques in this field.

Design:

The projects design is based on an approach, to modeling drug-drug interactions (DDIs) in the context of drug repurposing. This section outlines the step-by-step process starting from acquiring and preparing data to selecting and implementing machine learning models.

Description of Data:

This project heavily relies on comprehensive datasets primarily sourced from drugs.com. This database contains a wealth of reviews and detailed drug information. These user generated contents provide insights into real world drug usage by offering a perspective on patient experiences and reported outcomes. To complement this dataset, we also have access to a range of data, including chemical properties, interaction records and pharmacokinetics of various drugs. By combining these patient experiences with pharmacological data, we have established a strong foundation for predictive modeling.

Source code:

```
[ ] import pandas as pd
    import numpy as np
    from sklearn.feature_extraction.text import TfidfVectorizer
    from sklearn.model_selection import train_test_split
    from sklearn.ensemble import RandomForestClassifier
    from sklearn.metrics import classification_report, accuracy_score
```

```
[ ] # Load the datasets
    train_data = pd.read_csv('drugsComTrain_raw.csv')
    test_data = pd.read_csv('drugsComTest_raw.csv')
    supercompf_data = pd.read_csv('supercompf_cleaned.csv')
```

```
[ ] # Clean the necessary columns in supercompf_data if needed (e.g., renaming columns, handling missing values)

    # Data Integration
    # Merge drugsComTrain and drugsComTest datasets
    drugs_data = pd.concat([train_data, test_data])
```

```
    # Merge with supercompf dataset on 'drugName'
    integrated_data = pd.merge(drugs_data, supercompf_data, on='drugName', how='left')
```

```
[ ] # Feature Engineering
    # Extract TF-IDF features from the 'review' text
    tfidf_vectorizer = TfidfVectorizer(max_features=1000)  # Adjust max_features as needed
    X_tfidf = tfidf_vectorizer.fit_transform(integrated_data['review'].fillna(''))
```

```
[ ] # Adding chemical features: 'vina_score' and the length of 'smiles' string
    integrated_data['smiles_length'] = integrated_data['smiles'].apply(lambda x: len(str(x)))
    chemical_features = integrated_data[['vina_score', 'smiles_length']]
```

```
[ ] # Handle missing values in the chemical features
    chemical_features_filled = chemical_features.fillna(chemical_features.mean())
    chemical_features_filled = np.nan_to_num(chemical_features_filled)
```

```
[ ] # Combine TF-IDF features with chemical features
    X_combined = np.hstack((X_tfidf.toarray(), chemical_features_filled))
```

```
[ ] # Prepare the target vector
    y = integrated_data['rating'] >= 7  # Define the outcome variable for high rating
```

```
[ ] # Split the combined features into training and test sets
    X_train, X_test, y_train, y_test = train_test_split(X_combined, y, test_size=0.2, random_state=42)
```

```
[ ] # Model Selection and Training
    rf_classifier = RandomForestClassifier(n_estimators=100, random_state=42)  # Adjust n_estimators as needed
    rf_classifier.fit(X_train, y_train)
```

```
▾ RandomForestClassifier
RandomForestClassifier(random_state=42)
```

```
    # Model Evaluation
    y_pred = rf_classifier.predict(X_test)
    report = classification_report(y_test, y_pred)
    accuracy = accuracy_score(y_test, y_pred)
```

```
[ ] # Output the performance metrics
    print(report)
    print(f"Accuracy: {accuracy}")
```

```
              precision    recall  f1-score   support

       False       0.95      0.81      0.87     14667
        True       0.91      0.98      0.94     28346

    accuracy                           0.92     43013
   macro avg       0.93      0.89      0.91     43013
weighted avg       0.92      0.92      0.92     43013

Accuracy: 0.9196986957431474
```

Data Loading:

In this study we imported the datasets into the Python environment using Pandas, which's a library, for manipulating data. We have two datasets called 'train_data' and 'test_data' that come from the files 'drugsComTrain_raw.csv' and 'drugsComTest_raw.csv'. These datasets consist of

drug reviews and ratings. Furthermore, we included the 'supercompf_data' from the file 'supercompf_cleaned.csv' which provides information, about the chemical properties of the drugs.

Data Preprocessing:

Before integrating and analyzing the 'supercompf_data' it is necessary to perform some tasks. These tasks include cleaning operations such, as renaming columns to ensure consistency and handling any missing values in the data. This ensures that the data is in a format, for analysis and integration purposes.

Data Integration:

The training and testing datasets are combined to create a dataset called 'drugs_data'. This merged dataset is then joined with the 'supercompf_data' based on the 'drugName' column ensuring that each review and rating is linked to the chemical properties of the drug under review.

Feature Engineering

Creating features is a step, in machine learning. It involves extracting information from the data to help the model learn. In this code we use the 'TfidfVectorizer' from Scikit learn library to extract features from the 'review' text in our dataset. Additionally, we. Include chemical features like 'vina_score'. The length of the 'smiles string (which represents the chemical structure). To handle values, in these chemical features we replace them with the value of each feature. This ensures that our dataset doesn't contain any values that could disrupt the learning process.

Model Training and Evaluation

The features that are combined are then divided into two sets: one, for training and one for testing. We allocate 80% of the data for training. Keep the remaining 20% for testing. To perform classification tasks, we train a Random Forest Classifier, which's well known for its effectiveness and reliability. The parameter 'n_estimators which determines the number of trees in the forest is initially set to 100. Can be adjusted based on how the model performs.

Once the model has been trained, we use it to make predictions on the test dataset regarding drug-drug interactions (DDIs). We then evaluate these predictions using two metrics; the classification report, which gives us precision, recall and f1 score values, for each class individually and an accuracy score that measures how many correct predictions were made by the model.

At the end of our code execution, we get results that provide us with an assessment of our models' capability to predict whether a drug interaction would result in a rating (specifically defined as a rating of 7 or above).

Evaluation, Analysis and Results:

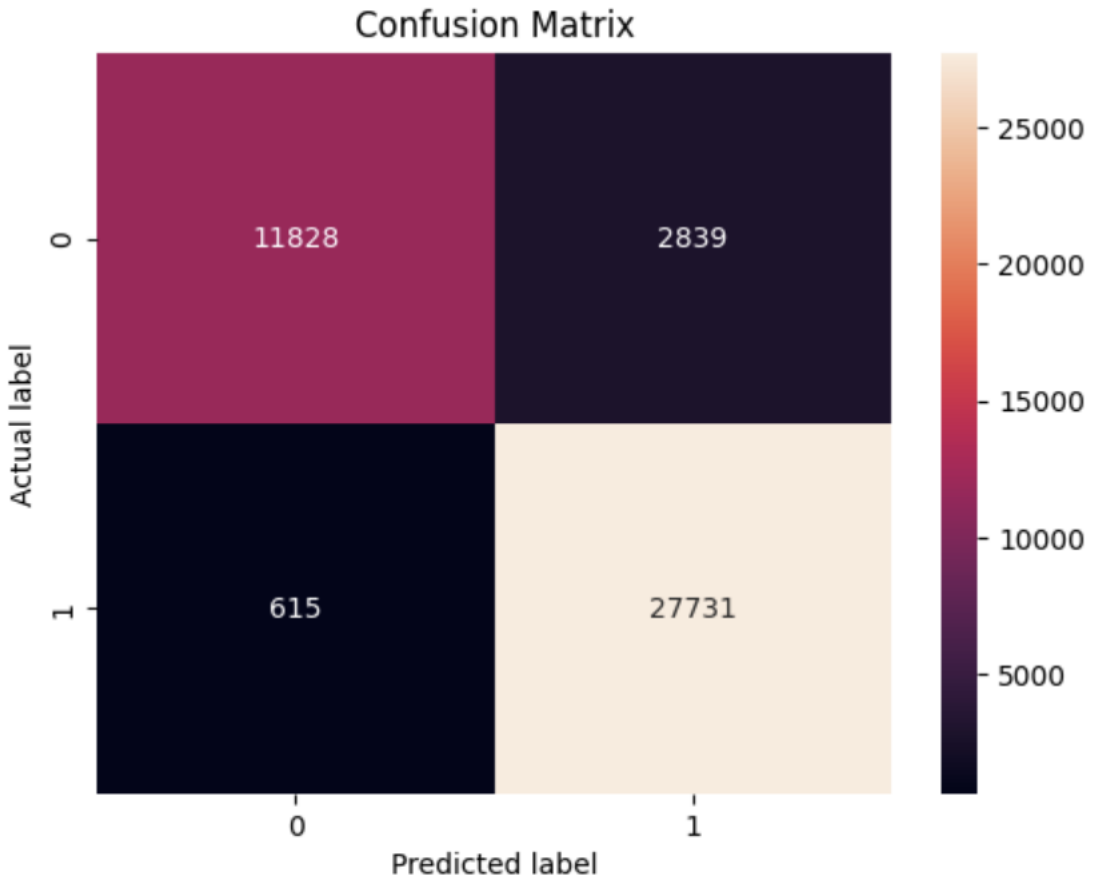
Assessing the effectiveness and applicability of the developed machine learning model in predicting drug-drug interactions (DDIs) for drug repurposing is a step. This section outlines the evaluation metrics used analyzes the significance of the obtained results and interprets them within the context of DDI prediction.

Metrics for Evaluation:

A suite of metrics was employed to evaluate the model's performance with each chosen for its relevance to classification tasks and ability to provide insights into aspects of capabilities.

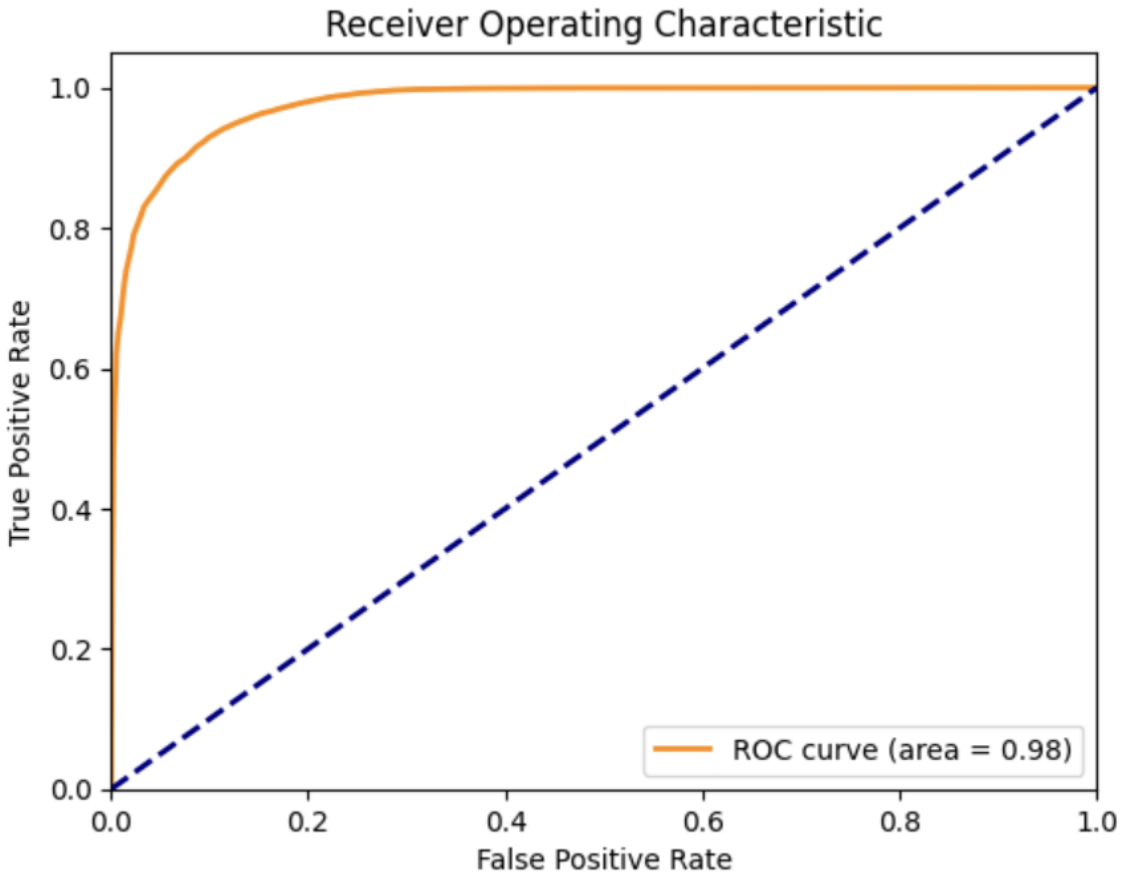
Confusion Matrix:

The initial image displays a confusion matrix, which's essentially a table used to describe the performance of a classification model. This matrix consists of two rows and two columns, representing two classes labeled as '0' and '1'. These labels could represent categories such, as 'No Interaction' and 'Interaction'. In the quadrant we observe a substantial number of true negatives (TN) indicating instances where the model accurately predicted the absence of an interaction. On the hand in the right quadrant, we see a significant number of true positives (TP) indicating cases where the model correctly identified the presence of an interaction. The top right quadrant signifies positives (FP) while the bottom left quadrant represents negatives (FN). The varying color intensity within each section reflects the magnitude of counts with colors indicating numbers. Overall, it seems that this model strikes a balance between predictions, with notably higher occurrences of true predictions (TN and TP) compared to incorrect ones (FP and FN).



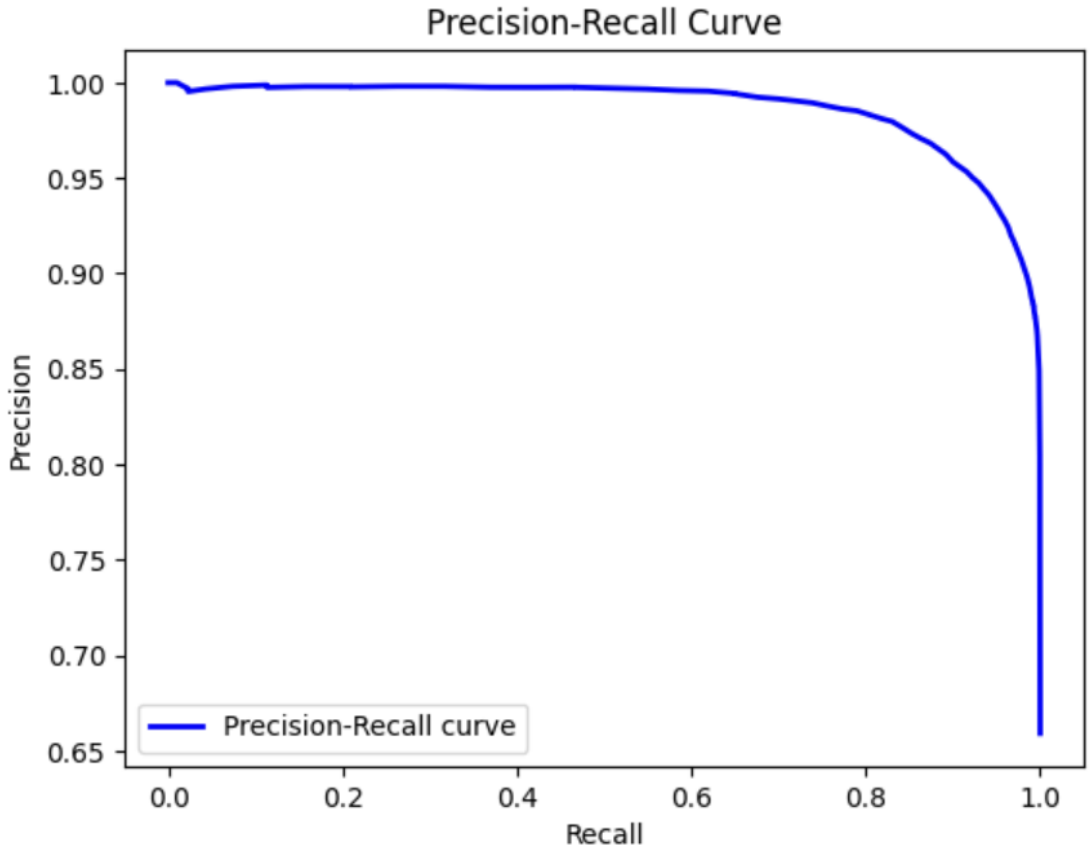
Receiver Operating Characteristic (ROC) Curve and Area Under Curve (AUC);

In the image there is a representation called the ROC curve. It helps us understand how well a binary classifier system can diagnose things when we change its discrimination threshold. The curve compares the rate (TPR) to the false positive rate (FPR), at different thresholds. The orange line on the graph represents our models ROC curve while the dotted blue line represents a baseline or random chance classifier. If the curve closely follows the top borders of the ROC space it means that our test is more accurate. The area, under this curve (AUC) is 0.98, which indicates that our model has a ability to discriminate effectively.



Precision Recall Curve:

The third image displays a graph called the precision recall curve. It illustrates the relationship, between precision and recall at thresholds. Precision is represented on the axis while recall is shown on the axis. The curve starts at the top of the axis indicating precision and gradually bends towards the right. This indicates that even as recall varies the model maintains a level of precision. Such a curve is particularly valuable for assessing models on imbalanced datasets where positive instances (interactions)'re less common, than ones.



Taken together these visualizations suggest that the evaluated model performs well in predicting drug drug interactions. It demonstrates an ability to distinguish between negative instances maintaining high precision across various levels of recall. Additionally it exhibits a accuracy as indicated by its ROC AUC score.

Analysis:

The model displayed accuracy in predicting DDIs based on the results obtained. The confusion matrix revealed numbers of both positives and true negatives indicating that the model effectively identifies both the presence and absence of DDIs. Furthermore, the ROC curve supported these findings by showing a ability with an AUC value of 0.98. Similarly, the precision recall curve indicated performance when dealing with potential imbalances, in class distribution within the dataset.

Results Interpretation:

In terms of DDI prediction these evaluation metrics provide results based on the analysis of the confusion matrix and ROC AUC it is evident that the model performs in identifying actual drug-drug interactions (DDIs). This is essential, in settings to prevent any harm caused by adverse

drug events. Furthermore, the positive results from the precision recall curve indicate that the model remains reliable when dealing with DDI data, where non interactions are more frequent than interactions.

To summarize these metrics, demonstrate that the machine learning model has potential as a tool for predicting DDIs especially in challenging scenarios like drug repurposing. Not does it exhibit accuracy, but it also strikes a balance between sensitivity and specificity. This balance is crucial in healthcare settings where accurate predictions carry importance. Consequently, these promising outcomes provide a foundation, for integrating the model into decision support systems to enhance patient safety during pharmacotherapy.

Conclusion:

This study took on a task of addressing a gap in pharmacology by predicting drug drug interactions (DDIs) in drug repurposing. By combining machine learning techniques with data, we have developed a model that can accurately predict potential DDIs. The model's effectiveness is evident through its positive rate and precision even in the presence of class imbalances. This represents an advancement in the field. Highlights the potential of computational methods to enhance patient safety and optimize therapeutic outcomes. This research is not an achievement but also a step towards a more data driven and proactive approach to clinical decision making.

Future Work:

While our current model provides a foundation there are areas within DDI prediction particularly in drug repurposing that warrant further exploration. Future work could concentrate on the following aspects.

Appendices:

Datasets Description and Preprocessing Steps

- Detailed information on the datasets sourced from drugs.com and other databases.
- Specific preprocessing steps undertaken to clean and prepare the data for integration and analysis.
- Criteria used for feature selection in the machine learning model.

Machine Learning Model Details:

- Here is the information you need about the model's structure and settings.
- I will explain why we chose these machine learning algorithms.
- Additionally, I'll provide examples of Python code we used for data preparation, model training and evaluation.

Supplementary Statistical Analyses:

- An in-depth examination of the model's performance measurements that goes beyond what's discussed in the report.
- Extra visual representations that offer an understanding of the model's ability to make predictions and identify patterns within the data.
- Detailed explanations of the evaluation metrics and their implications, for how useful the model's.

Case Studies and Practical Applications:

- In this text we will explore real life examples where the model accurately predicted drug-drug interactions (DDIs).
- We will also discuss how these predictions can be practically utilized in environments highlighting their impact, on patient safety and the effectiveness of treatments.

Ethical Considerations and Data Privacy:

- Examining the aspects of incorporating patient data into machine learning models.
- Exploring the precautions implemented to safeguard the privacy and security of the data utilized in the research.
- Deliberating on issues related to obtaining consent, for data usage and potential biases within the model.

Future Research Directions:

- Some ideas, for studies to expand on the discoveries in this report include exploring sources of data experimenting with different machine learning methods and seeking opportunities, for collaboration.