# Exploratory Data Analysis and Feature Engineering in Drug Discovery

Leveraging Machine Learning for Bioinformatics

By

Muhammad Sufiyan

Paper: Machine learning approaches and databases for prediction of drug–target interaction: a survey paper | Briefings in Bioinformatics | Oxford Academic (oup.com)

# Abstract:

This project focuses on leveraging machine learning for drug discovery through exploratory data analysis (EDA) and feature engineering. We conducted in-depth EDA to handle missing values and uncover essential patterns. Feature engineering, a key aspect, involved novel molecular features. An additional achievement was the incorporation of canonical_smiles and morgan_fingerprint, showing promising results.

The presentation will highlight methodology, challenges, and the potential impact of my findings on advancing drug discovery.

# Proposed Work In Paper:

The original paper conducted exploratory data analysis (EDA) on a chemical compound dataset, employing linear regression models for bioactivity prediction. Emphasis was placed on feature engineering, specifically canonical smiles and morgan fingerprints. Proposed future work included the integration of additional molecular features to improve model performance, setting the stage for advanced predictive modeling in bioactivity research.

# Dataset:

The dataset is sourced from the ChEMBL database and centers around the Acetylcholinesterase enzyme, pivotal in Alzheimer's treatment. It serves the purpose of predicting bioactivity related to inhibiting the enzyme, making it valuable for drug discovery and chemoinformatics.

https://www.kaggle.com/datasets/gauravan/human-acetylcholinesterase-dataset-from-chembl

# Model:

I work with an ensemble learning approach, specifically a Random Forest model. The Random Forest model, known for its versatility and ability to handle complex datasets, contributed to the overall predictive power of our solution. Methodology encompassed extensive exploratory data analysis (EDA), addressing missing values, and introducing advanced feature engineering with canonical_smiles and morgan_fingerprint. This combination of models and techniques culminated in a comprehensive and effective approach to drug discovery.

# EDA & Feature Engineering:

- Checking Data Types and Basic Information:

  Provides information about the data types, non-null counts, and memory usage of each column.
- Summary Statistics and Correlation Matrix:

  Display summary statistics and the correlation matrix of numerical columns.
- Visualizing Correlation Matrix:

  Generates a heatmap to visualize the correlation matrix.
- Visualizing Distribution of a Column (e.g., 'standard_value’):

  Visualizes the distribution of a specific column, such as 'standard_value’.

- Feature Engineering with SMILES Strings and Morgan Fingerprints:

  This part defines a function (process_smiles) to generate Morgan fingerprints from SMILES strings and applies it to a subset of the data, creating a new column ('morgan_fingerprint').

# Achievement:

Key achievements include successful handling of missing values and extraction of essential patterns through exploratory data analysis (EDA). The introduction of canonical_smiles and morgan_fingerprint as part of feature engineering demonstrated promising results, surpassing baseline models. The hybrid approach involving both a deep neural network and a Random Forest model showcased versatility, achieving high accuracy. These accomplishments collectively contribute to the potential real-world impact of revolutionizing drug discovery processes.

# Conclusion:

In bioactivity prediction unveiled an effective model architecture, Random Forest model. Overcoming challenges such as missing values in the chemical compound dataset, our approach showcased resilience and adaptability. Looking ahead, the positive findings from this project suggest continued opportunities for improvement. The implications extend beyond the current scope, with potential applications in various domains of drug discovery. This project serves as a foundational step towards reshaping the landscape of predictive modeling in chemoinformatics.

# Future Work:

As I conclude this phase of our project, the path forward is clear. Future work involves the ongoing refinement of our model by integrating additional molecular features and exploring advanced techniques. The hybrid approach opens avenues for fine-tuning, aiming for even higher predictive accuracy. The commitment to continuous improvement and innovation positions this work as a catalyst for future breakthroughs. Collaborations and partnerships within the scientific community will play a crucial role in advancing the impact of our research on drug discovery.