Bioinformatics Project - Computational Drug Discovery

📌 Overview This project is part of my learning journey in bioinformatics and computational drug discovery. It focuses on building machine learning models to predict the bioactivity of compounds targeting the acetylcholinesterase enzyme, which is relevant for Alzheimer's disease treatment. The project was developed based on the tutorial by Data Professor (<https://www.youtube.com/watch?v=zD2focOkQ48>), with additional enhancements including feature selection optimization and identification of the top 10 most promising compounds.

🧬 Project Components

1. Data Preparation

* Collected bioactivity data from ChEMBL database
* Processed IC50 values to pIC50 (negative log of IC50) for better model performance
* Calculated molecular fingerprints using PaDEL descriptors

1. Feature Engineering

* Applied variance threshold to remove low-variance features
* Used SelectKBest for feature selection (top 100 most relevant features)
* Split data into training (80%) and test (20%) sets

1. Model Development

* Implemented Random Forest regression as primary model
* Compared performance of 42 different machine learning algorithms using LazyPredict
* Visualized model performance metrics (R-squared, RMSE)

1. Enhancements

* Identified and displayed the top 10 most promising compounds with:
* ChEMBL IDs
* Canonical SMILES
* Experimental pIC50 values
* Predicted pIC50 values

🛠️ Technologies Used

* Python
* Pandas
* Scikit-learn
* RDKit (via PaDEL descriptors)
* LazyPredict
* Matplotlib/Seaborn

📊 Key Results Random Forest achieved R² = 0.54 on test set

* HistGradientBoosting showed best performance among all algorithms tested (R² = 0.45)
* Identified promising compounds for further investigation

🎯 Learning Outcomes Through this project, I gained practical experience in:

* Computational drug discovery pipelines
* Molecular descriptor calculation
* Machine learning for bioactivity prediction
* Model evaluation and comparison
* Feature selection techniques

The project serves as a foundation for more advanced work in cheminformatics and drug discovery.