UMBC

Final Project Proposal

DATA 603 Platforms for Big Data Processing

Shreya Patil
ID- HG53212

Problem Statement

Prediction of Drug Binding Affinity of Protein

- Healthcare, Pharmacy and Biological Big Data have entered the digital era. Computational drug discovery is an effective strategy for accelerating and economizing drug discovery and development process.
- Filtering large compound libraries into smaller sets of predicted active compounds using computational models that can be tested experimentally for accurate match is the most common practice.
- In my project I want to predict the likelihood of binding between sample drug compounds and target protein.

Dataset

- https://www.ebi.ac.uk/chembl/
- The ChEMBL Database is a public database that contains curated bioactivity data of more than 2 million compounds.
- I will be filtering and downloading the compound data for single target.
- I will select target protein such that it will have ~2500 compound records.
- Some of the features will be molecular weight, target type, protein classification, standard value.

Problem Solving Approach

- Exploratory Data Analysis on Compound data.
- Divide the data into train and test dataset.
- Using linear Regression model predicting the binding probability of the compound for selected target protein.
- Check the accuracy of the model.