UMBC

Final Project Progress

DATA 603 Platforms for Big Data Processing

Shreya Patil
ID- HG53212

Problem Statement

Prediction of Drug Binding Affinity of Protein

- In my project I want to predict the likelihood of binding between sample drug compounds and target protein.
- If the activity value is less then that is considered good compound for potential drug for the protein activity control

Dataset

- https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL4040/
- For the project I have chosen ‘MAP kinase ERK2’ as my target protein.
- It is one type of Kinase protein which are **intracellular enzymes that regulate cell growth and proliferation as well as the triggering and regulation of immune responses**. Protein kinases are important therapeutic targets in cancer because of their critical role in signaling mechanisms that drive malignant cell characteristics.
- For this target protein we have binding affinity values for 4643 potential drug molecules.

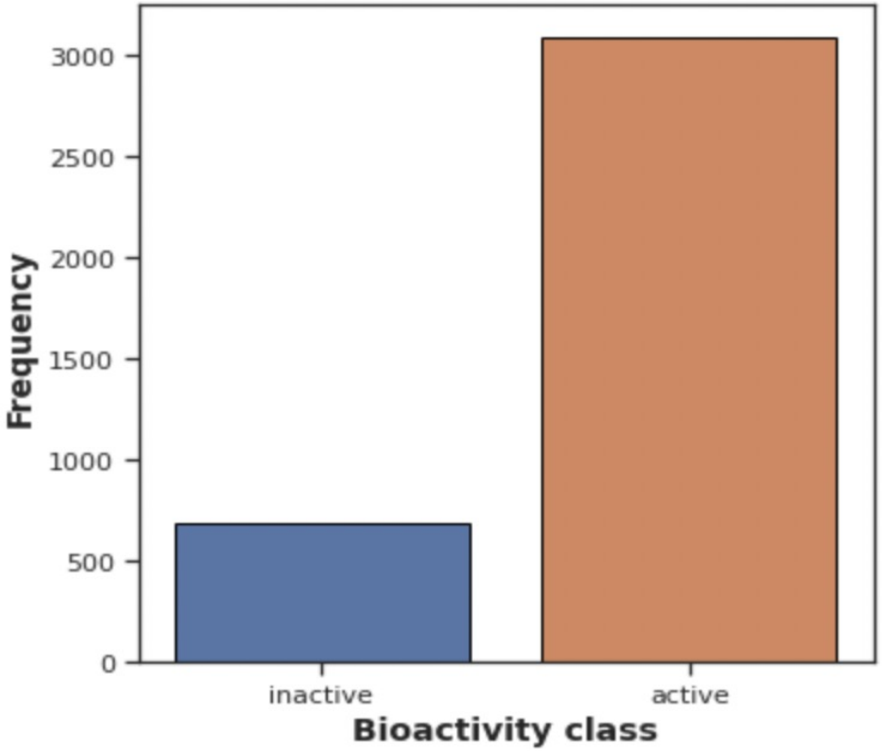
Dataset

- For this dataset I have extracted their 5 features which makes them druglikeness compound . https://en.wikipedia.org/wiki/Lipinski%27s_rule_of_five
  - No more than 5 hydrogen bond donors (the total number of nitrogen–hydrogen and oxygen–hydrogen bonds)
  - No more than 10 hydrogen bond acceptors (all nitrogen or oxygen atoms)
  - A molecular mass less than 500 daltons
  - An octanol-water partition coefficient[10] (log P) that does not exceed 5

EDA results

- Data type of all the columns in float.
- Dropped the values for which has missing values for ‘**standard_value’ column.**
- **Classified ‘standard_value’ column**
    - **>=** 1000 – Inactive class
    - < 1000 – active class
- After cleaning shape of the dataset is

  (3777, 6)



Problem Solving Approach

- Divide the data into train and test dataset.
- Using Logistic Regression model predicting the binding probability of the compound for selected target protein.
- Check the accuracy of the model.