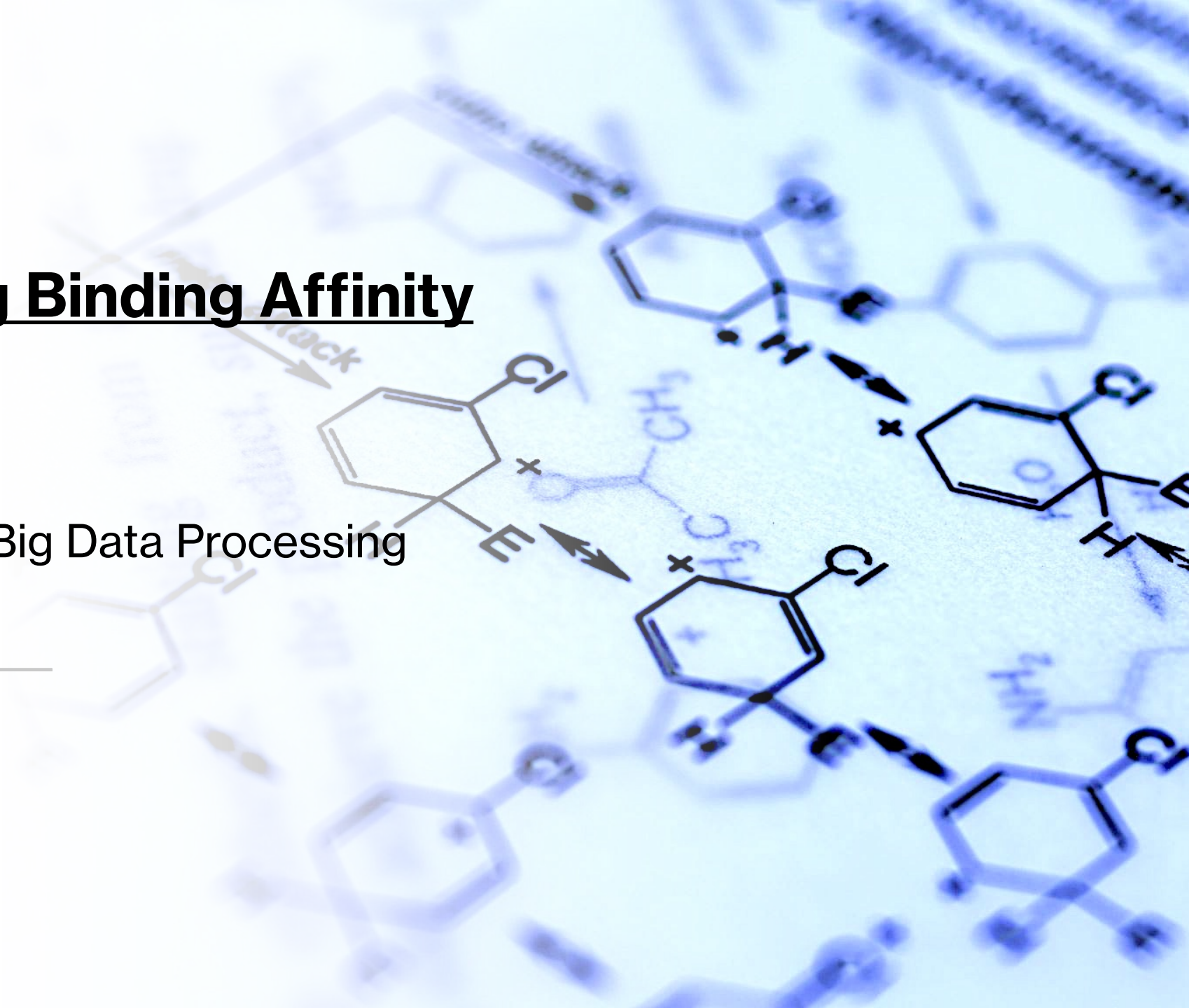
UMBC

Prediction of Drug Binding Affinity of Protein

DATA 603 Platforms for Big Data Processing

Shreya Patil
ID- HG53212



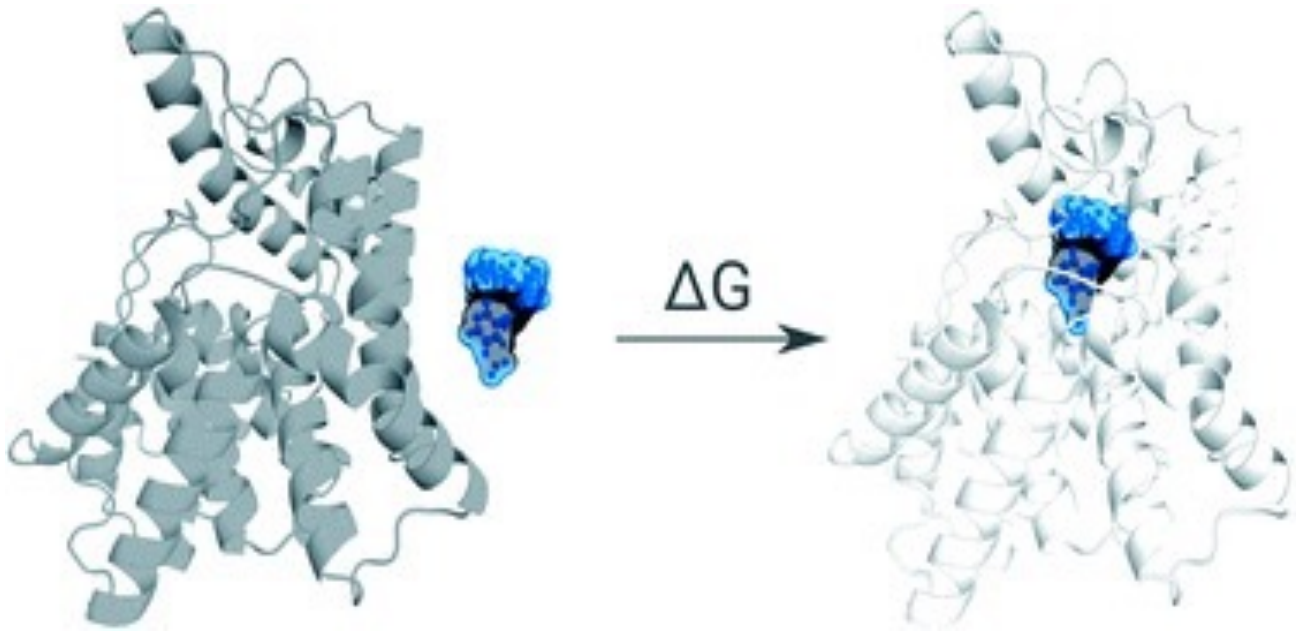
Introduction

- Healthcare, Pharmacy and Bio-engineering big data driven fields.
- Computational drug discovery.
- Filtering large compound libraries into smaller sets of predicted active compounds using computational methods.
- In my project I want to predict the likelihood of binding between sample drug compounds and target protein.

Problem Statement

To Predict Drug Binding Affinity of Protein

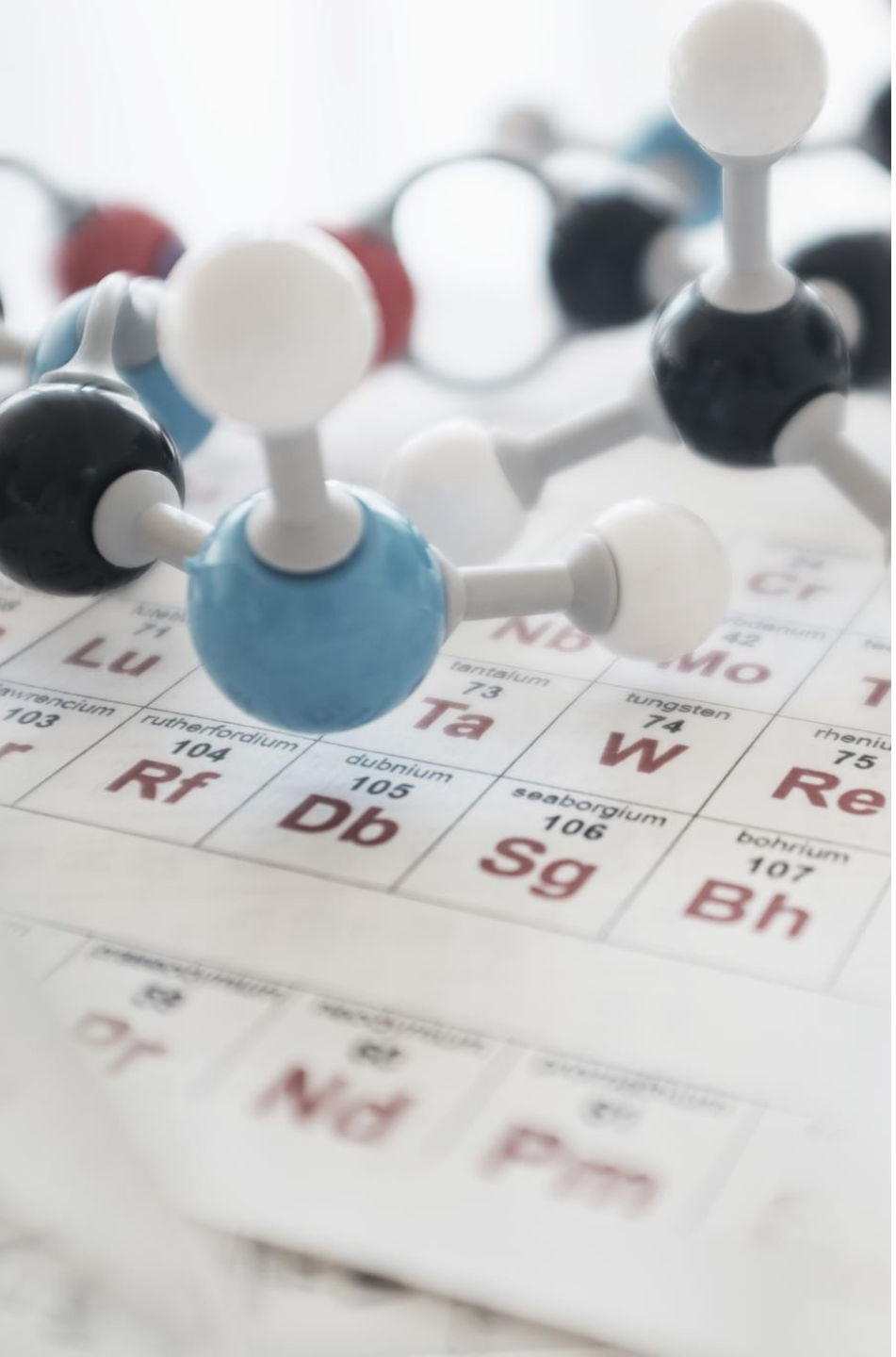
Less standard activity value = Drug likely Compound



ΔG = free energy of binding

Dataset

- ChEMBL Database
- Target Protein - ‘MAP kinase ERK2’
- Kinase protein are **intracellular enzymes that regulate cell growth as well as the triggering and regulate immune responses of human body**.
- Protein kinases are important therapeutic targets in cancer because of their critical role in signaling mechanisms that drive malignant cell characteristics.
- 4643 Potential drug molecules.



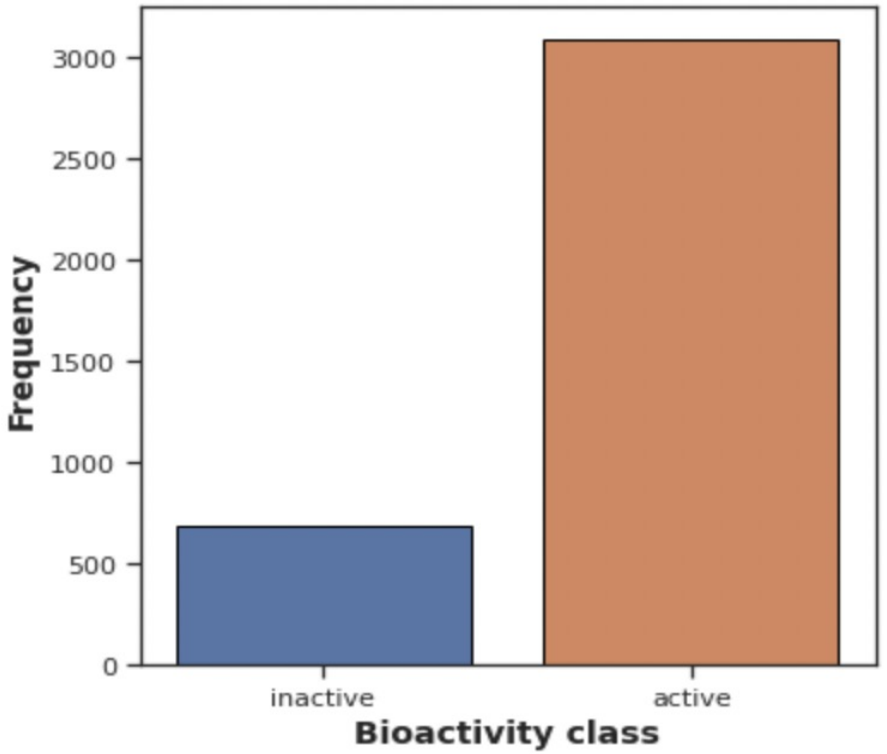
Dataset

- **Lipinski's rule of five**, also known as **Pfizer's 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

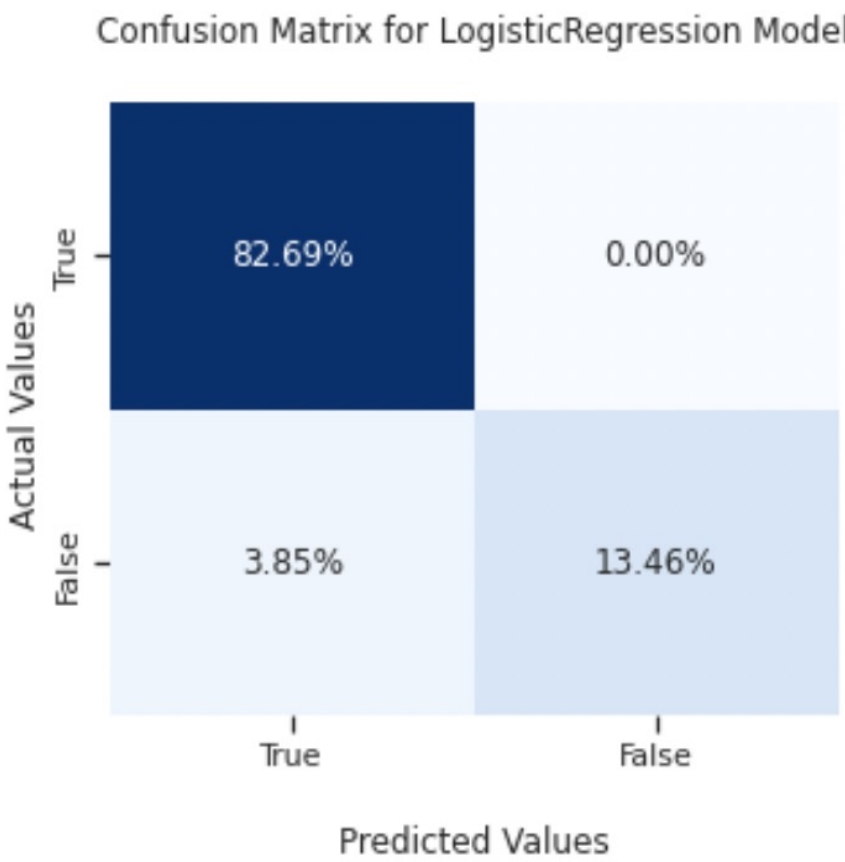
EDA Results

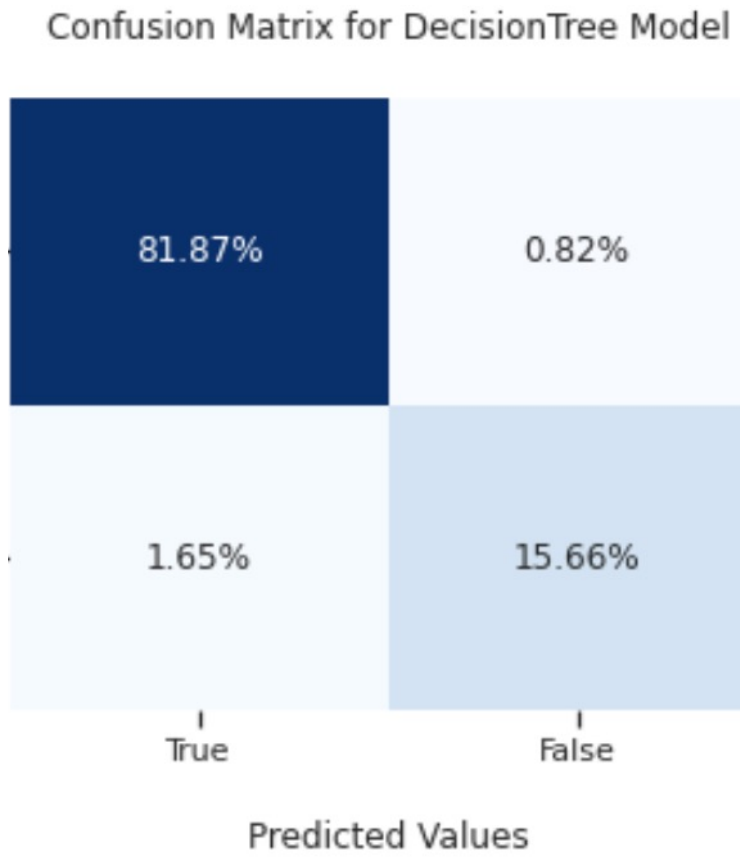
- Data type of all the columns in float.
- Dropped the values for which has missing values for '**standard_value'** column.
- **Classified 'standard_value'** column
  - **>=** 1000 – Inactive class
  - < 1000 – active class
- After cleaning shape of the dataset is

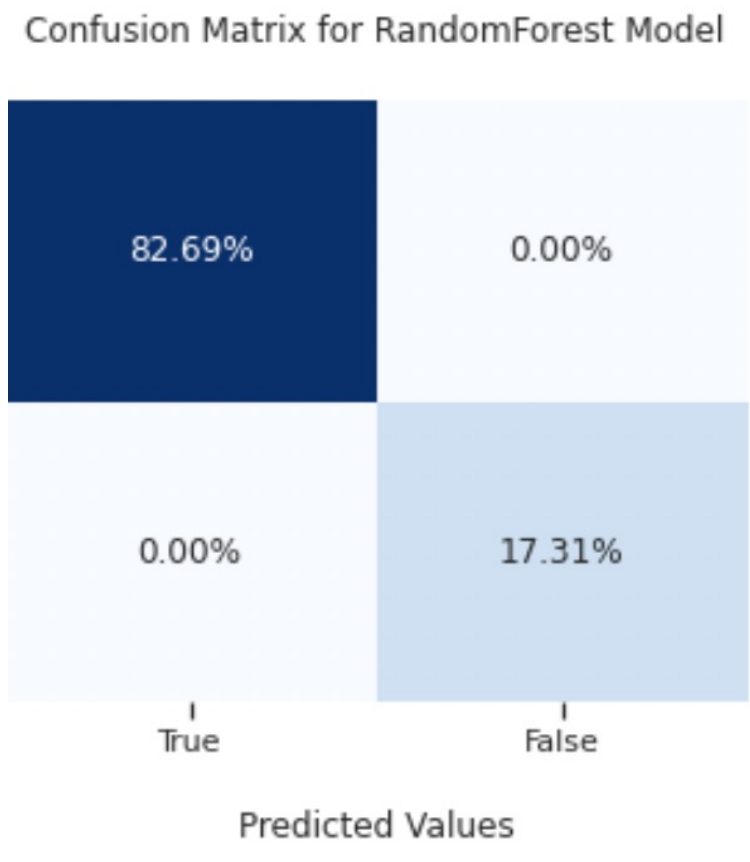
  (3777, 5)



Modeling and Result







Modeling and Result

	Models	Performance_(areaUnderROC)	Accuracy
0	Logistic Regression	0.998431	0.961538
1	Decision Tree	0.998102	0.975275
2	Random Forest	1.000000	1.000000

Future Scope

1

Using additional features other than rule of five.

2

Using Multiclass Classifier

3

Targeting more Proteins to find drug compounds.

Thank You!