Predicting Pharmacokinetics Properties

Devbrat Anuragi
17078

I have Used Properties like Physicochemical Properties, Lipophilicity, Water Solubility, Druglikeness and Medicinal Chemistry to predict Pharmacokinetics Properties of a molecule Using Random Forest.

Predicting Pharmacokinetics Properties

Workflow

Select the Target.
I have selected **hemoglobin subunit beta** as the target. This is the target gene for the Sickle Cell Anemia.

Browse Activities for target.
A csv file of Activities of the molecules related to hemoglobin subunit beta was downloaded from the ChEMBL database

Web crawl properties
I found a website http://www.swissadme.ch/index.php which convert smiles to its Physicochemical Properties, Lipophilicity, Water Solubility, Druglikeness and Medicinal Chemistry and Pharmacokinetics. So using the ChEMBL ID I get the Smiles of each molecule and then used smiles to get the above mentioned properties.

Train the Random Forest
With the properties I have trained the Random Forest Classifier and Random Forest Regressor to classify and predict Pharmacokinetics properties

Random Forest Visalization
So for each classifier I have created 10 decisioon tree. You can find all the decision tree in the separate folder attached with this assignment.

What data I have scraped from the webpage?

Physicochemical Properties	
Formula	C13H18O2
Molecular weight	206.28 g/mol
Num. heavy atoms	15
Num. arom. heavy atoms	6
Fraction Csp3	0.46
Num. rotatable bonds	4
Num. H-bond acceptors	2
Num. H-bond donors	1
Molar Refractivity	62.18
TPSA	37.30 Å²

Lipophilicity	
Log $P_{o/w}$ (iLOGP)	2.17
Log $P_{o/w}$ (XLOGP3)	3.50
Log $P_{o/w}$ (WLOGP)	3.07
Log $P_{o/w}$ (MLOGP)	3.13
Log $P_{o/w}$ (SILICOS-IT)	3.15
Consensus Log $P_{o/w}$	3.00

Water Solubility	
Log *S* (ESOL)	-3.36
Solubility	9.09e-02 mg/ml ; 4.41e-04 mol/l
Class	Soluble
Log *S* (Ali)	-3.97
Solubility	2.23e-02 mg/ml ; 1.08e-04 mol/l
Class	Soluble
Log *S* (SILICOS-IT)	-3.44
Solubility	7.49e-02 mg/ml ; 3.63e-04 mol/l
Class	Soluble

Pharmacokinetics	
GI absorption	High
BBB permeant	Yes
P-gp substrate	No
CYP1A2 inhibitor	No
CYP2C19 inhibitor	No
CYP2C9 inhibitor	No
CYP2D6 inhibitor	No
CYP3A4 inhibitor	No
Log K_p (skin permeation)	-5.07 cm/s

Druglikeness	
Lipinski	Yes; 0 violation
Ghose	Yes
Veber	Yes
Egan	Yes
Muegge	Yes
Bioavailability Score	0.85

Medicinal Chemistry	
PAINS	0 alert
Brenk	0 alert
Leadlikeness	No; 1 violation: MW<250
Synthetic accessibility	1.92

All the image shown above are taken from the website and all these table consist of the properties related to the molecles. So I have trained Random Forest Classifiers for predicting each of the Pharmoinetics properties.

Train Set set and test set

Dependent Variable =

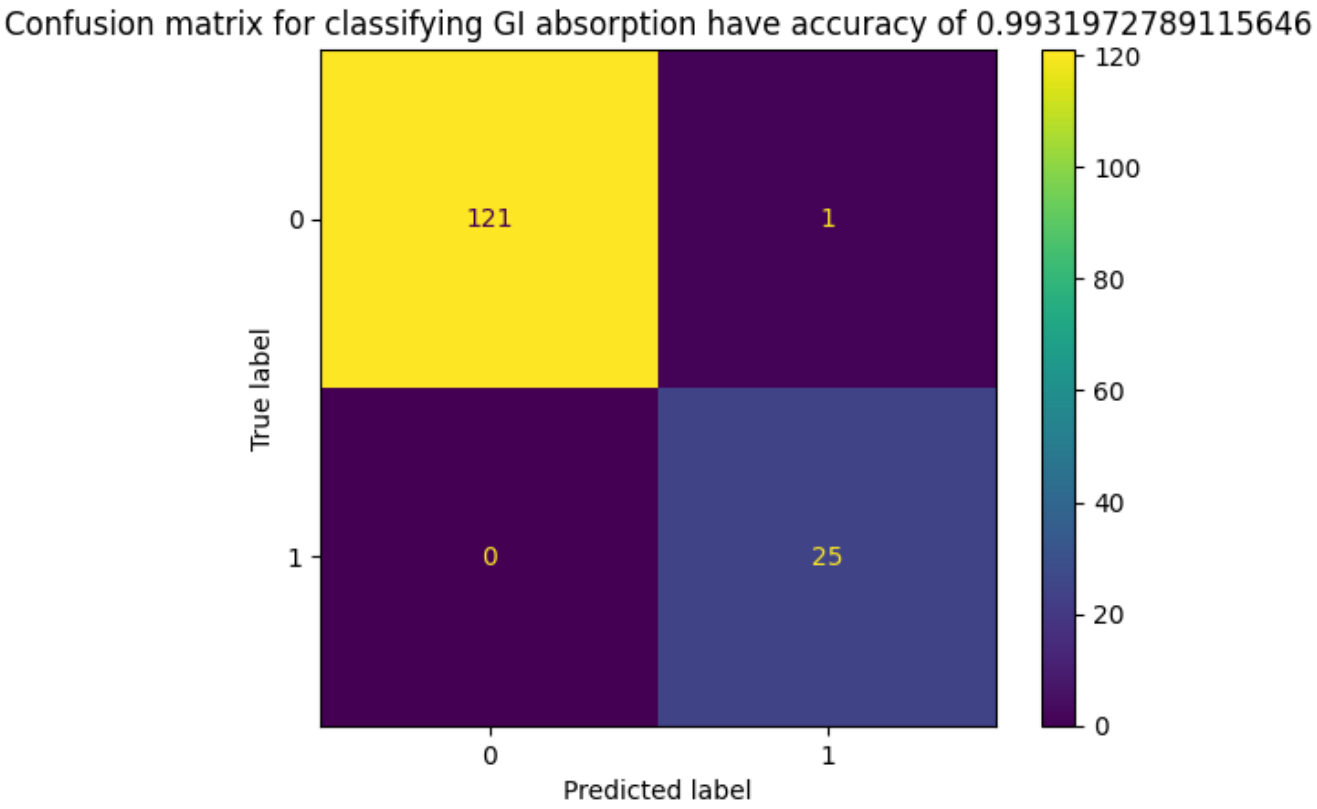
- Molecular weight",
- "Num. heavy atoms",
- "Num. arom. heavy atoms",
- "Fraction Csp3",
- "Num. rotatable bonds",
- "Num. H-bond acceptors",
- "Num. Hbond donors",
- "Molar Refractivity",
- "TPSA", "Log Po/w (iLOGP) ",
- "Log Po/w (XLOGP3) ",
- "Log Po/w (WLOGP) ",
- "Log Po/w (MLOGP) ",
- "Log Po/w (SILICOSIT) ",
- "Consensus Log Po/w",
- "Class_cat",
- "Class .1_cat",
- "Class .2_cat",
- "Log S (ESOL) ",
- "Log S (Ali) ",
- "Log S (SILICOSIT) ",
- "Solubility .1",
- "Solubility .2",
- "Solubility ",
- "Lipinski_cat",
- "Ghose_cat",
- "Veber_cat",
- "Egan_cat",
- "Muegge_cat",
- "Bioavailability Score "

The Variable which are ending with *_cat* like Class_cat are categorical variable. All the remaining variables are the numerical values.

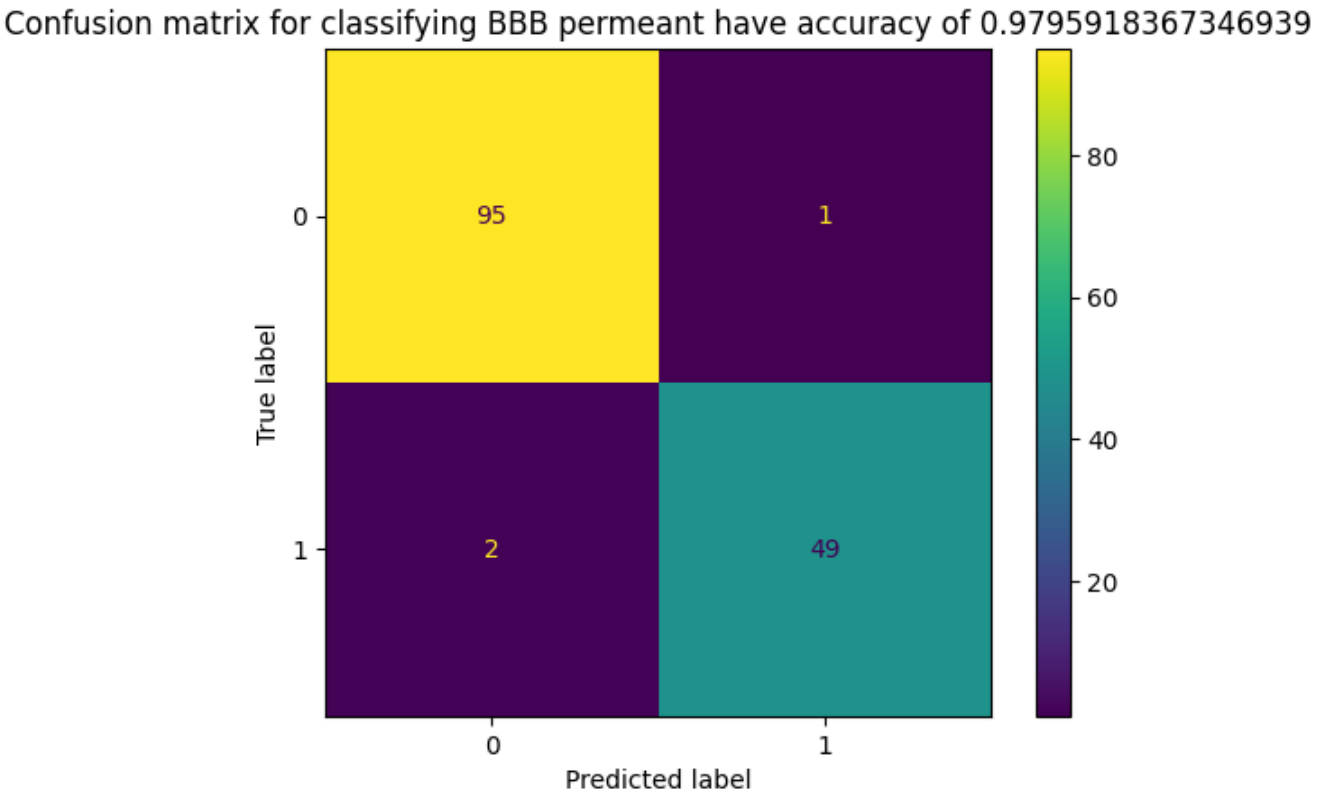
Independent Variable:

- GI absorption
- BBB permeant
- P-gp substrate
- CYP1A2 inhibitor
- CYP2C19 inhibitor
- CYP2C9 inhibitor
- CYP2D6 inhibitor
- CYP3A4 inhibitor
- Log K_p (skin permeation)

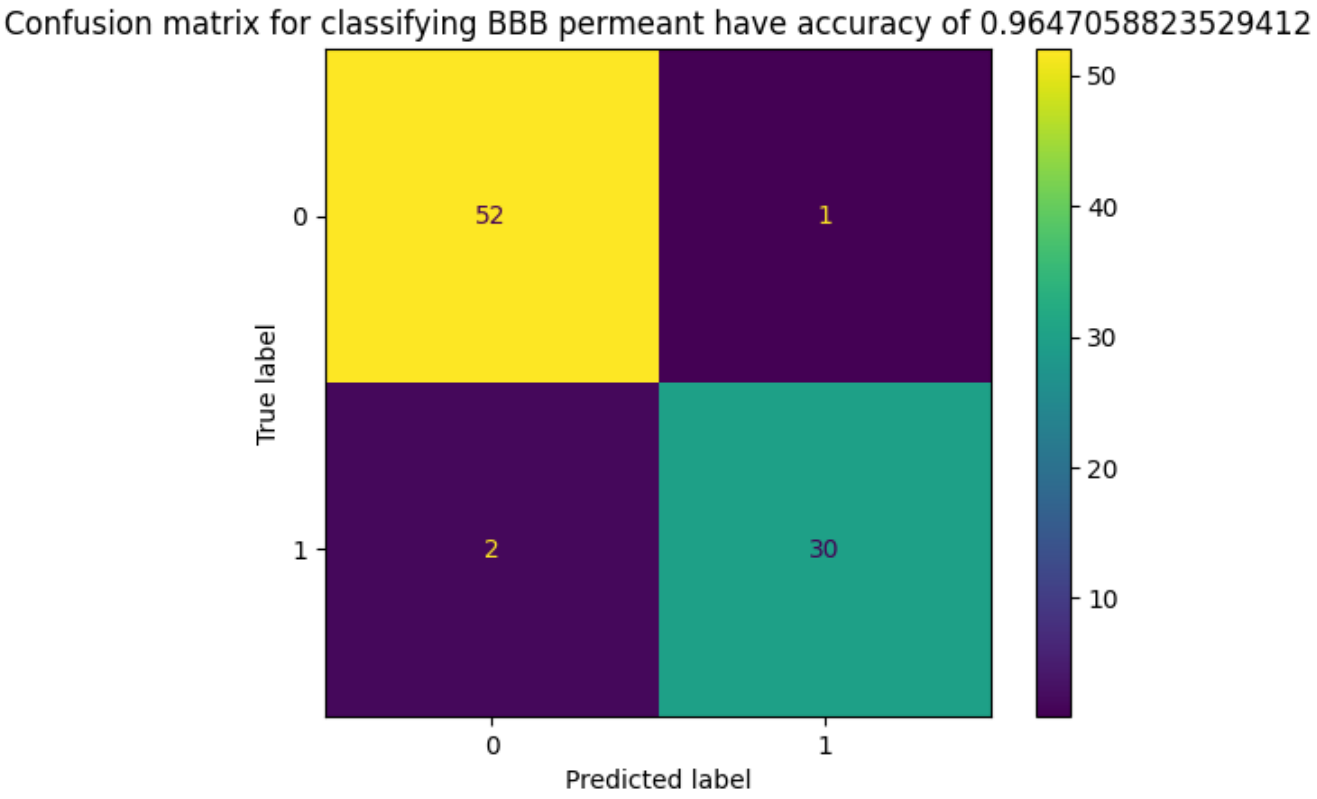
Random Forrest Classifier for Classifying GI Absorption:



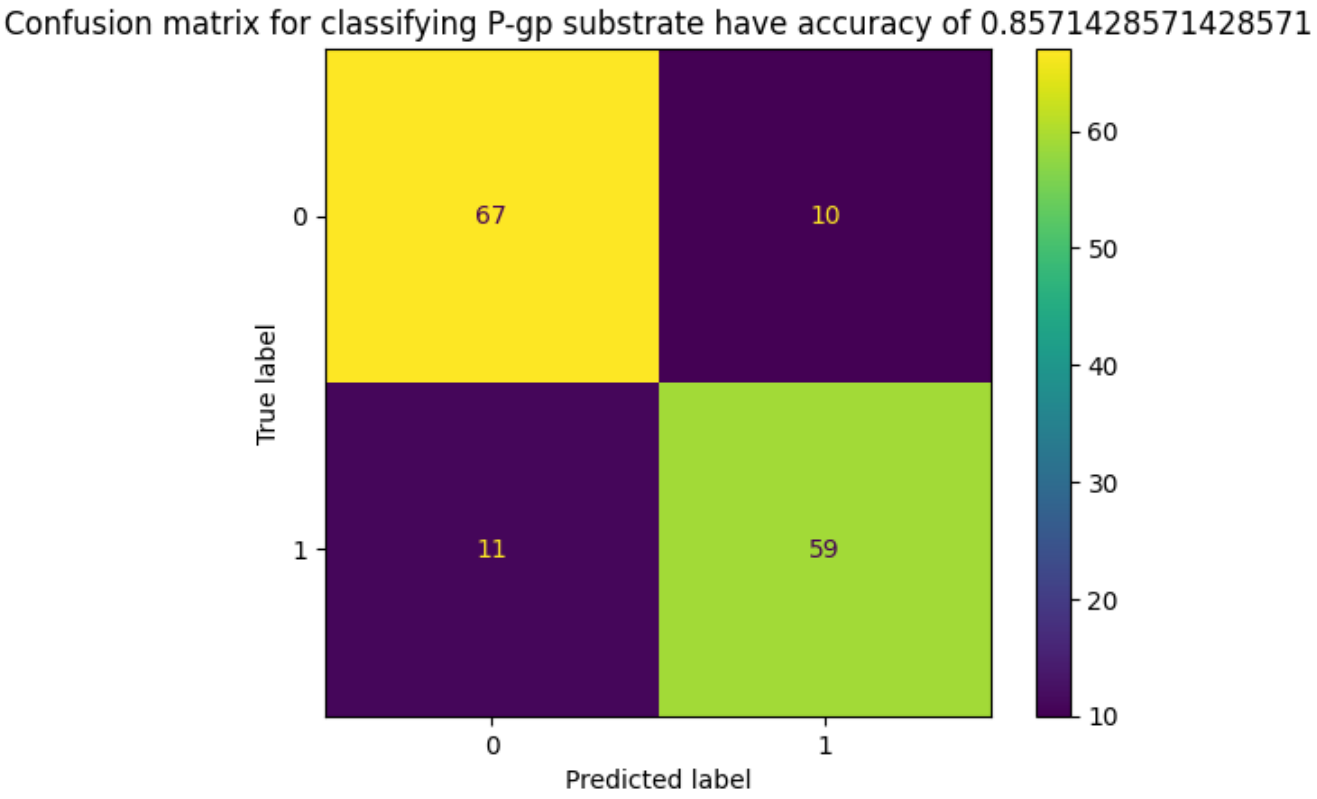
Random Forrest Classifier for Classifying BBB permeant



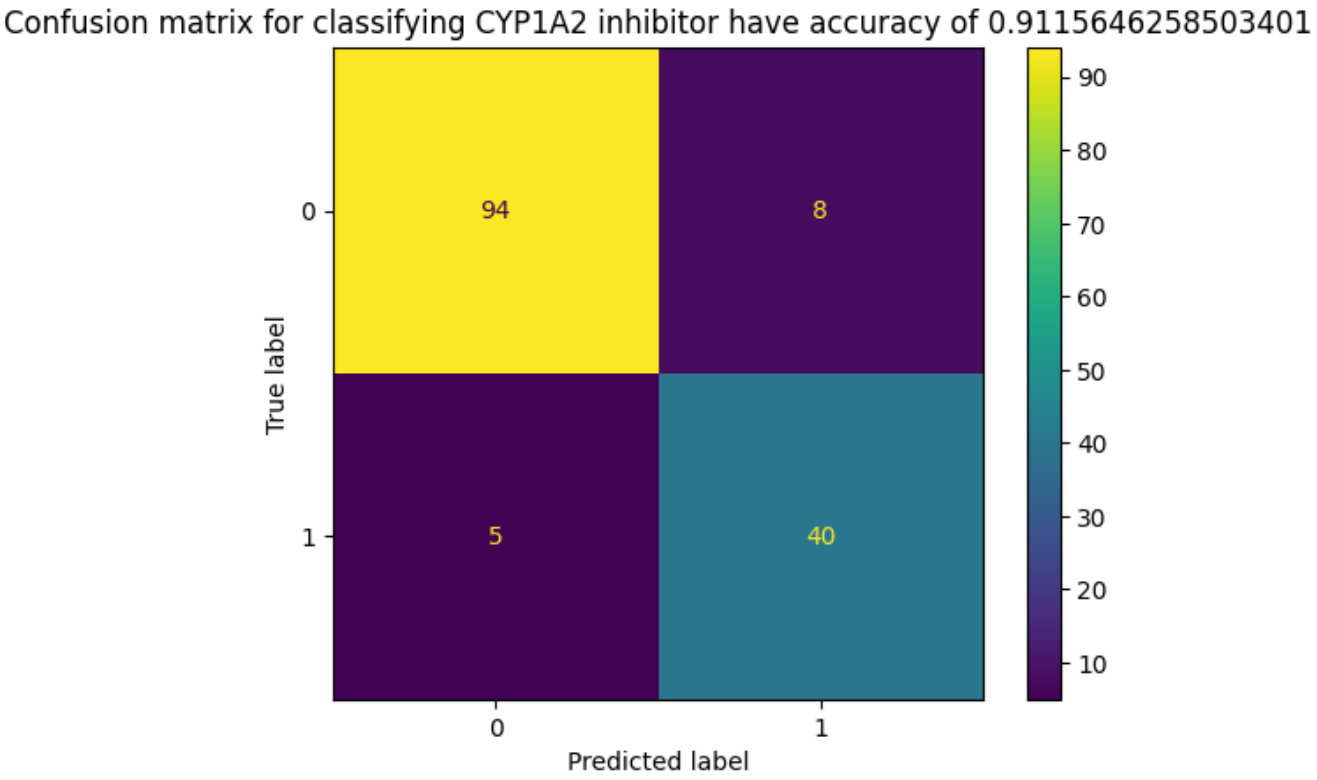
Random Forrest Classifier for Classifying BBB permeant



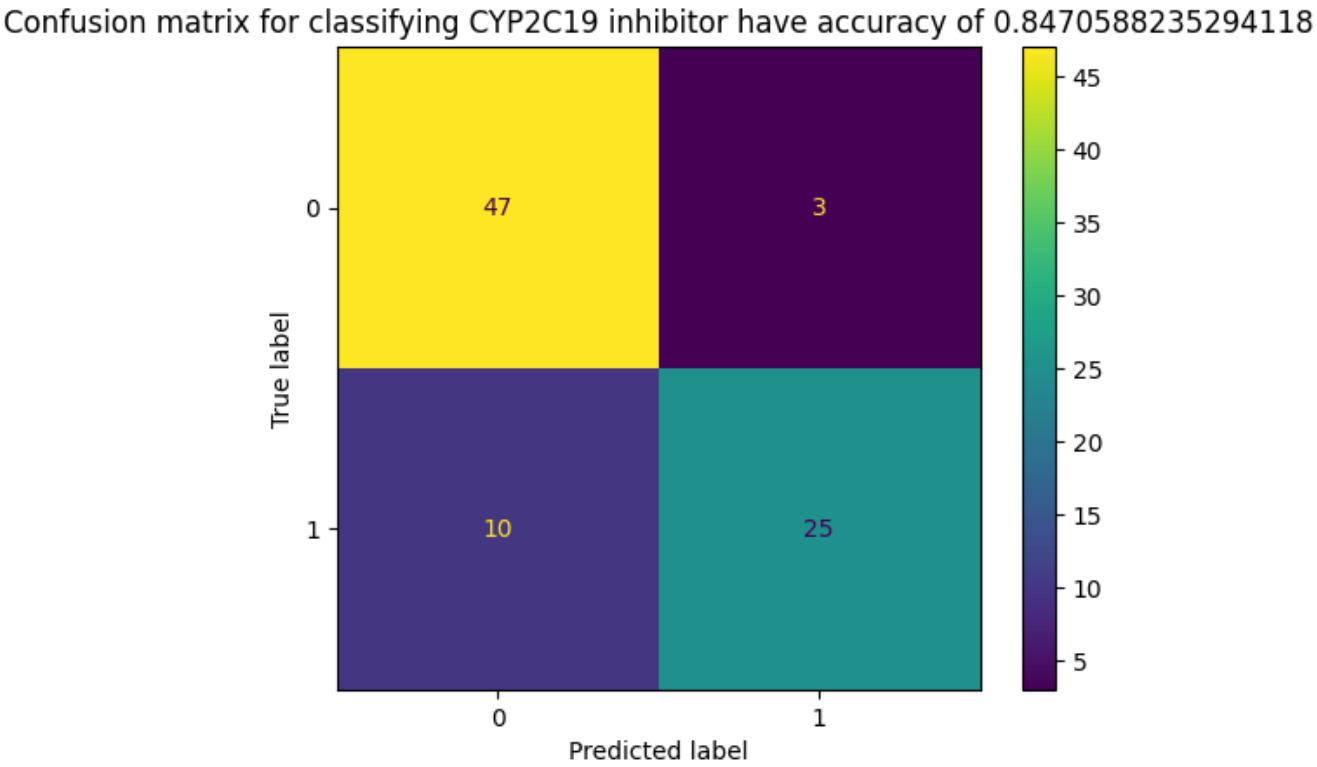
Random Forrest Classifier for Classifying P-gp substrate



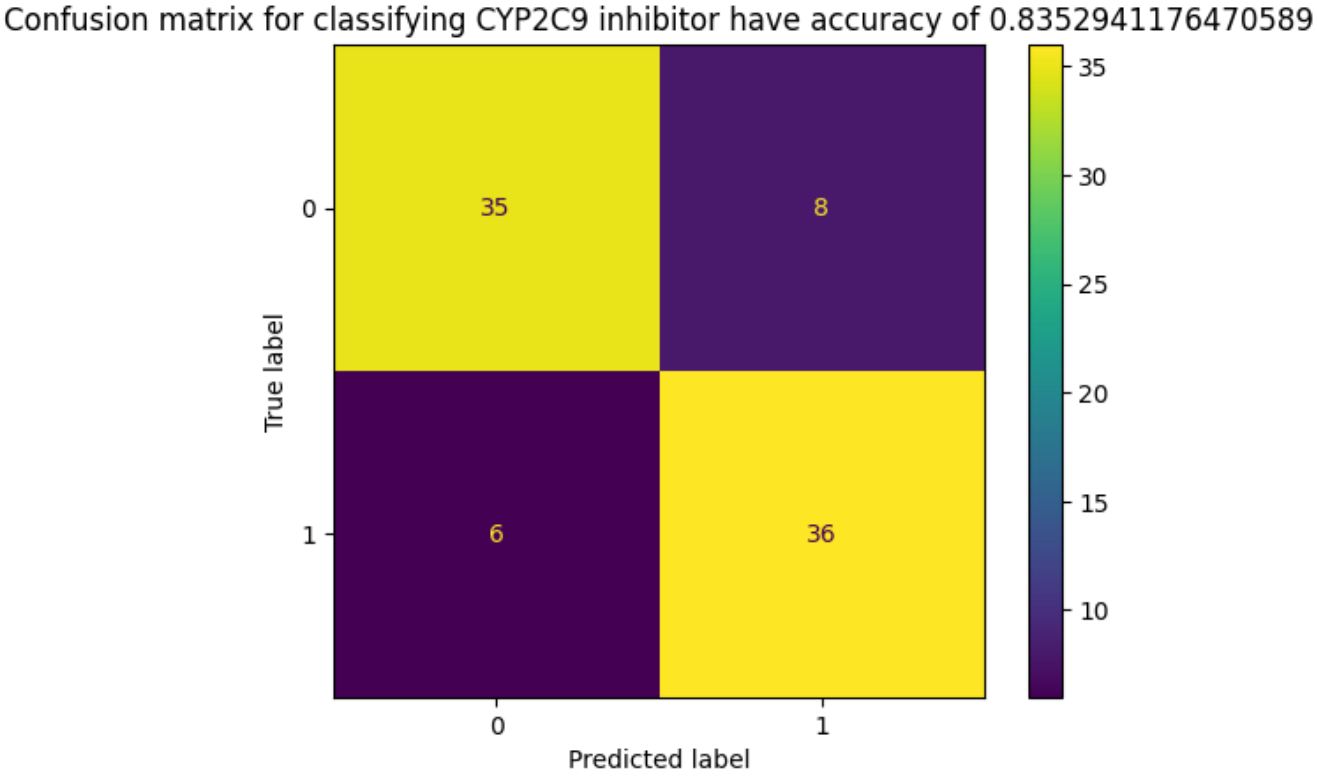
Random Forrest Classifier for Classifying CYP1A2 inhibitor



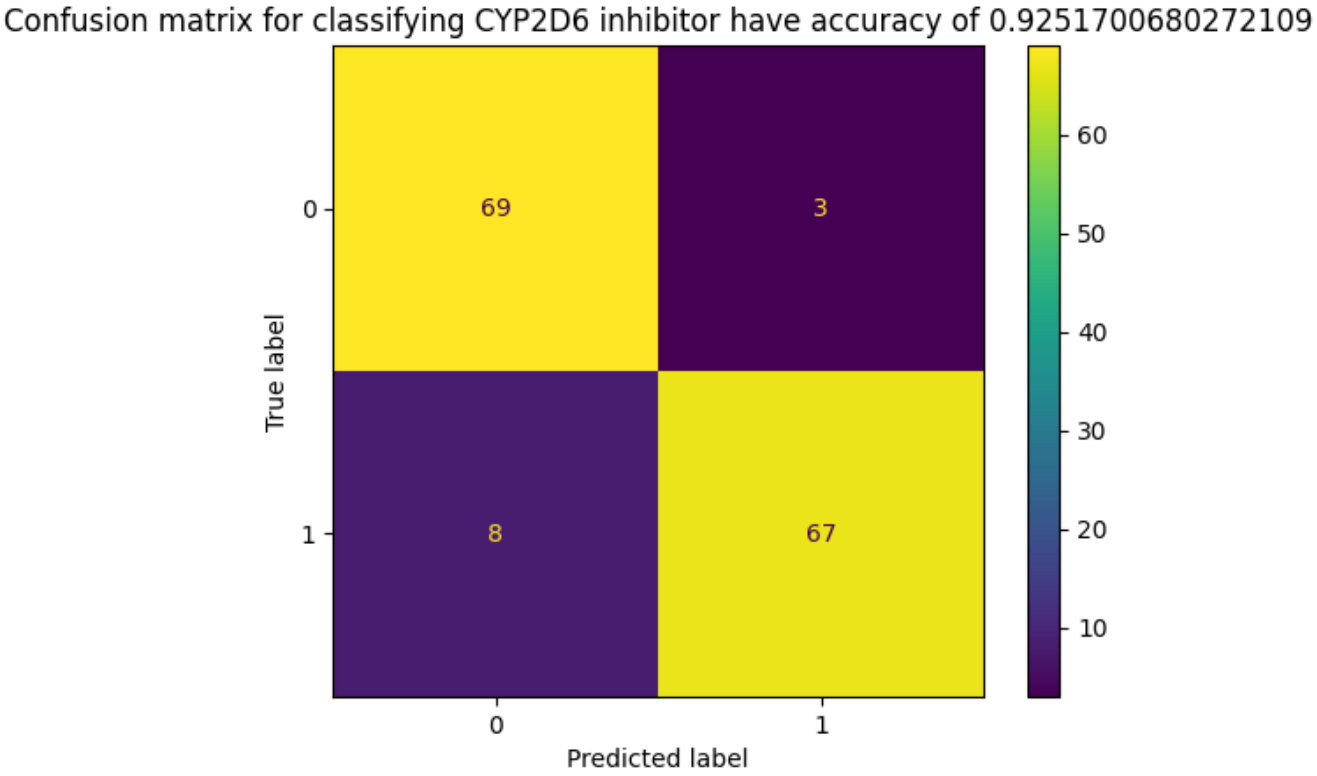
Random Forrest Classifier for Classifying CYP2C19 inhibitor



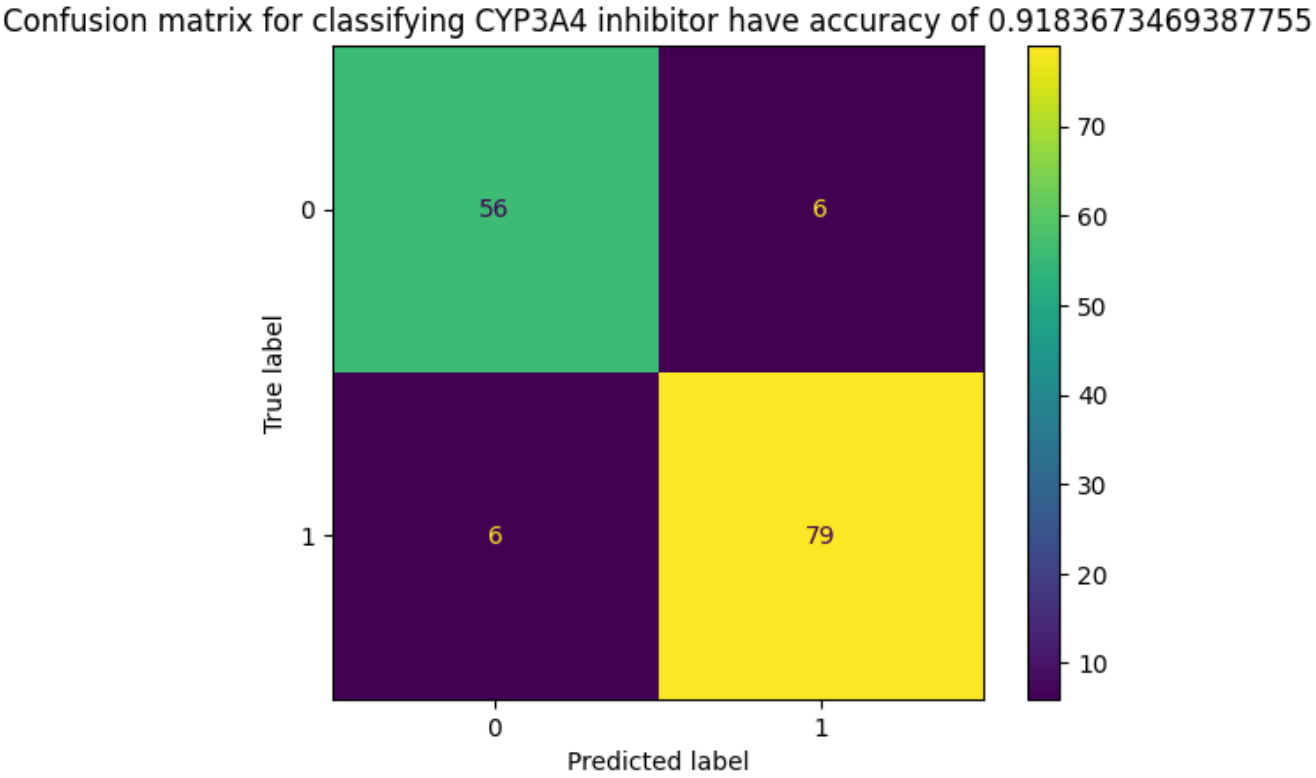
Random Forrest Classifier for Classifying CYP2C9 inhibitor



Random Forrest Classifier for Classifying CYP2D6 inhibitor



Random Forrest Classifier for Classifying CYP3A4 inhibitor



Random Forrest Regressor for Classifying CYP3A4 inhibitor

Model accuracy score with 10 decision-trees : 0.9859746165068916