INTERNATIONAL JOURNAL OF ADVANCE RESEARCH AND DEVELOPMENT

(Volume 4, Issue 4)

Available online at: www.ijarnd.com

Drug activity prediction of small drug molecules using Random Forest model

Vineet Yadav[1], Ujjawal Goel[2], Tanuj Kumar[3], Utkarsh lakhera[4]

[1,2,3,4] Student, IMS Engineering College, Ghaziabad, Uttar Pradesh

ABSTRACT

The aim of this paper is to develop predictive models that can determine, whether a particular compound is active (1) or not (0). In this, we have different datasets along with the drugs. A dataset contains a number of features. Every feature will be processed by using feature selection algorithms and particular compound also be found with the help of these features. We have taken the dataset from Tox21 databases, clean the data and then apply the different feature selection algorithms in order to predict the desired result by applying different models whether the particular drug will be beneficial or injurious to the health of our patient. All the models will be compared with each other and the suitable one will be selected whose accuracy will be more.

Keywords— *Drug, Feature Selection, Predictive Model, Random Forest Algorithm, Boruta Algorithm, Machine Learning Models, Adaboost*

1. INTRODUCTION

In this, we have different datasets along with the drugs. A dataset contains a number of features. Every feature will be processed by using feature selection algorithms and particular compound also be found with the help of these features and the Final Drug Activity prediction will be done based on these features only. We have to find whether the drug will be having a positive (+ ve) or negative (-ve) effect on the body of the human being.

1.1 Drug

Drugs are typically small organic molecules that achieve their desired activity by binding the target site on a receptor. The first step in the discovery of a new drug is usually to recognize and isolate the receptor to which it might bind, followed by testing many tiny molecules. This makes researchers determine what separates the active (binding) compound from the inactive (non-binding) ones.

1.2 Machine learning models

Machine Learning Models focuses on the development of computer programs that can access data and use it to learn for themselves.ML models for binary classification problems predict a binary outcome (one of two possible classes). We have used these models to classify whether the drugs are active or inactive.

Table 1: Machine learning models

Model	Method	Required Package
Random Forest	Rf	Random forest
Decision Tree	Rpart	Rpart
Adaboost	ada	ada

1.3 Feature selection

Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. In this project, there were 1444 features and out of those 1444 features 139 features are extracted by using the Boruta Feature Selection Algorithm. We can get the feature importance of each feature of your dataset by using the feature importance property of the model. Feature importance gives you a score for each feature of your data; the higher the score more important or relevant is the feature towards your output variable. [6].

1.4 Boruta algorithm

Boruta Algorithm is a feature selection algorithm. Precisely, it does also work with the Random Forest algorithm parallelly. This package [7] derives its name from a demon in Slavic mythology that dwelled in pine forests. We all know that feature selection is a crucial step in predictive modelling. Particularly when one is interested in understanding the mechanisms related to the variable of interest, rather than just building a black box predictive model with good prediction accuracy. [5].

2. CLASSIFICATION MODEL

2.1 AdaBoost

AdaBoost is the best technique to improve the performance of the decision tree on every classification type problems. This technique is given by the authors Freund and Schapire. But now, it is used more in classification rather than regression. Generally, AdaBoost is used to boost the performance of all machine learning algorithms. It is best used for weak learners. After applying this model we can achieve the highest accuracy on a classification problem. [1]

2.2 Decision tree

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. For instance, in the example below, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules, and the fitter the model. [3]

2.3 Random forest

Random forests, also known as random decision forests, are a popular ensemble method that can be used to build predictive models for both classification and regression problems. Ensemble methods use multiple learning models to gain better predictive results in the case of Random forest [7], the model creates an entire forest of random uncorrelated decision trees to arrive at the best possible answer. [7].

3. METHODOLOGY

The Methodology is represented in figure 1. The Dataset contains drug activities, the combination of negative and positive drug activities including 1444 features. A Feature Selection Algorithm [7] can be used to get the subset of important features. This process reduces the space complexity, time complexity as well as increases the accuracy of the model. [2] In the fifth step, the dataset is used to train the classifiers, with their optimum tuning parameters. The used machine learning models are presented in table 1. The proposed model is divided into three phases and all phases are explained below:

- **Phase I:** The Random Forest Algorithm, Adaboost, Decision Tree model can be trained with A% of the dataset and generate predictions from (100-A) % of the dataset.
- **Phase II:** The Boruta Algorithm selects the important features according to the feature of high importance value.
- **Phase III:** Three different models (Decision Tree, Random Forest, AdaBoost) can be applied to the same dataset so that we can find efficient models between these three models.
- **Phase V:** Compared the result of these different models.
- **Phase VI:** Apply the efficient Model to predict the results.

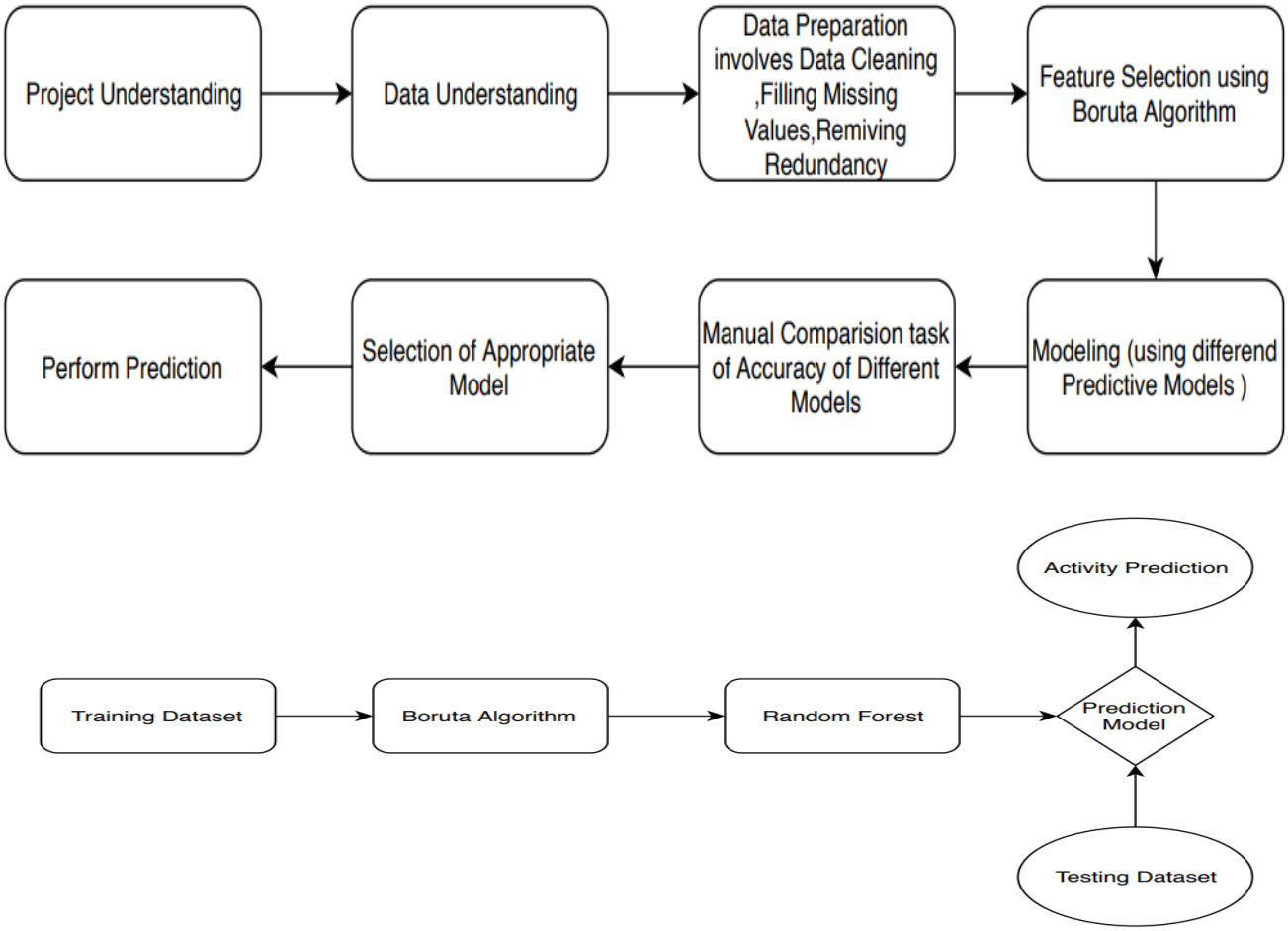

Fig. 1: Methodology of Project

4. RESULT ANALYSIS

Different models are applied on the dataset and the result is compared with each other and founded that the Random Forest model is having the maximum accuracy. The result may vary after considering the different dataset. Another problem is over fitting/ under fitting, to deal with over fitting/under fitting issue, the model should be cross-validated and tested on an independent dataset, and if performance is found to be consistent then models are free from over fitting/under fitting. Over fitting is when the model learns too much and under fitting is when the model learns too less. In cross-validation, models are executed n times and accuracy is recorded if accuracy is highly fluctuating then that model is over fitted/under fitted/biased [Figure 2]. Different Prediction models which are developed manually are compared with each other and found that Random Forest is the best model to be applied for best accuracy and results.

Table 2: Comparison of different models designed by applying algorithms manually

			Confusion Matrix				
SN	ModelName	AUC	TP	FP	TN	FN	Accuracy
1	Decision Tree	1	286	7	272	4	98.07
2	Random Forest	0.997	257	1	304	7	98.59
3	AdaBoast	0.995	375	6	364	14	97.36

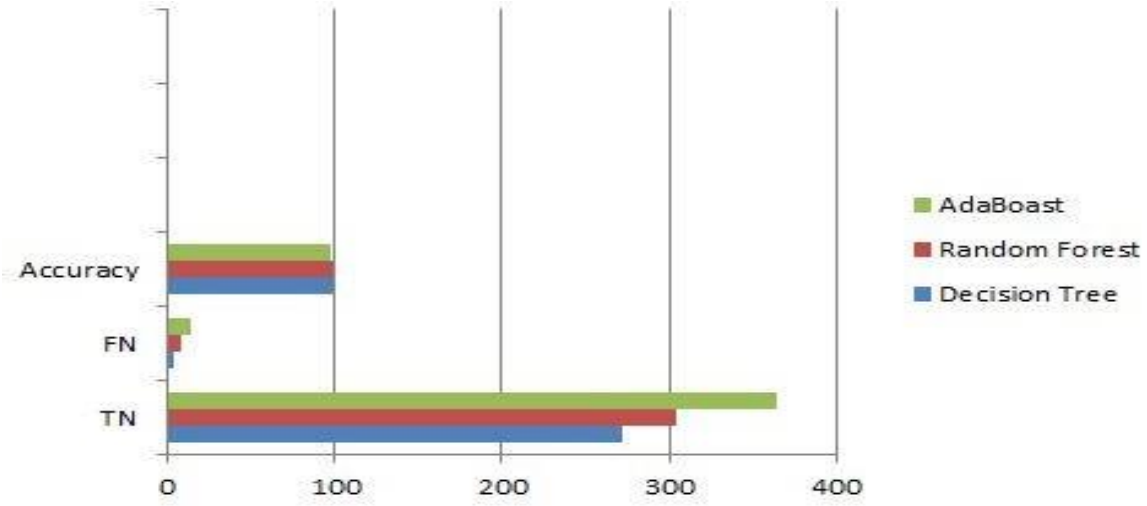

Fig. 2: Comparison of different models designed by applying algorithms manually

5. CONCLUSION

Different methods for Drug Activity Prediction are explored and their accuracies are compared. With these results, we infer that the Random Forest Model is more suitable in handling the classification problem of Drug Activity Prediction, and we recommend the use of these approaches in similar classification problems. This work presents a comparison among the different Data mining classifiers on the database of Drug Activity, by using classification accuracy as well as the prediction of Drugs Activity. It is concluded that accuracy of the Random Forest Algorithm has been compared with other models by applying on the same dataset which contains 139 Features after Feature Selection and about 2000 tuples. The accuracy of the different models is represented below with histogram. It is found that the Accuracy of Random Forest is greater than the other two models.

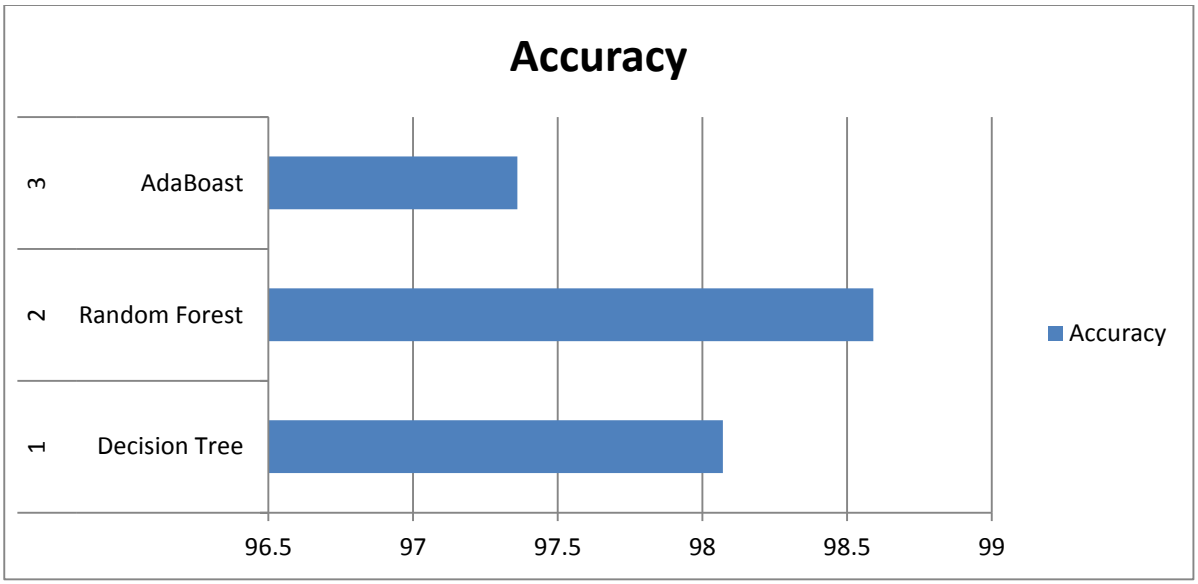

Fig. 3: Accuracy

6. REFERENCES

[1] "Machinelearningmastery.com/boosting-and-adaboos t-for-machine-learning/."
[2] Multilevel ensemble model for prediction by Divya Khanna∗, Prashant Singh Rana.
[3] Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning).
[4] Miron B. Kursa and Witold R. Rudnicki "Feature Selection with the Boruta Package".
[5] Kohavi R, John GH (1997). "Wrappers for Feature Subset Selection." Artificial Intelligence, 97, 273–324.
[6] Leisch F, Dimitriadou E (2010). mlbench: Machine Learning Benchmark Problems. R package version 2.0-0.
[7] Breiman L (2001). "Random Forests." Machine Learning, 45, 5–32.
[8] Liaw A, Wiener M (2002). "Classification and Regression by random Forest." R News, 2(3), 18–22.

