**Final project assignment**

**Part A: Dose finding project**

**1 Introduction**

**1.1 Background**

Malaria, an fatal infectious disease caused by Plasmodium parasites, leading to much sever illness and death especially in AfricaThe most common treatment for Malaria is the combination of a long-acting and a fast-acting antimalarial drug in order to avoid resistance. Artimisinin derivatives, as the long-acting antimalarial drugs, cases of resistance development to artemisinin derivatives has been reported in several countries(Nosten, et al 2002; Dondorp, et al 2009; Noedl, et al 2008; Phyo, et al 2012).

**1.2 A phase IIb study**

In this study, we have totally 8 study arms and 120 individuals in each of them. All individual were given combination therapy with 500 mg piperakine (long-acting antimalarial drug) and, either artemisinin or AMP1050 (100 mg, 150 mg, 250 mg, 500 mg, 750 mg, 1000 mg, 2000 mg). The highest drug concentration was recorded as well as the drug effect which was calculated as the parasites level in time 0 and 6 hour and the difference was represented as logarithm.

**2 Methodology**

**2.1 Examine the data**

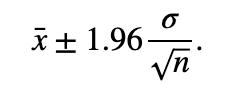
‘data\_110.csv’ was imported and loaded to R. The effect-variable was calculated by logarithm of the difference of parasites level in time 0 and 6 hour. 2 graphs were plotted in order to visualize the difference in the drug effect and side effect variable to each study arms respectively.

**2.2 Standard statistical analysis**

Two pairwise t-test was performed to compare the drug effect/side effect between treatment arms and study arm as we are interested in the difference between the tested drugs and positive control. 2 hypothesis were set: **1 H0**: difference of mean effect/side effect between groups is zero; **2 Ha**: mean effect/side effect in the population of an arm is greater than the population mean of arm. After comparison, the highest dose of AMP1050 that results in a better effect and has a side-effect profile that is not inferior than artemisinin study arm was chosen.

**2.3 Model based analysis**

A plot describing the correlation between drug effect and maximum drug concentration for AMP1050 was showed. A linear model, and Emax model and an Sigmoid Emax model were fit to and described the decrease in parasite levels in the blood. The best model was chosen based on the AIC value and the confidence interval(ci) for each x value was calculated in order to account for uncertainty. The average effect of 500 mg artimisin (+ piperakin) was predicted by the mean of ci and visualized on the effect curve. Ci was calculated by the following equation, which x represent the mean of drug effect, σ represent Standard deviation and n represent number of samples.

****

To calculate the minimum Cmax on the Cmax effect curve, the value should be higher than the lwr of ci in case the tested drug at least as good an effect as positive control. **Max. and Min. Cmax was calculated by the following equation :**

0.2916667 = 0.0013415x -1.0285104

6.487789 = 0.0453x - 3.224

where -1.0285104 and -3.224 , 0.0013415 and 0.0453 represent the y-intercept and slope of the fit and linear model(Table 3 and 6). 0.2916667 represent the mean of the probability of getting side effect and 6.487789 represent the lower boundary of drug effect(artimisin). Predict() function was performed to account for the uncertainty.

**The highest dose was calculated by the equation :**

984.1052 = x1 + 256.2 \* 1.96

214.3883 = x2 - 256.2 \* 1.96

where x1 and x2 represent the mean of Cmax, 984.1052 and 249.697 represent the upper boundary and lower boundary of Cmax, 256.2 is the residual standard error and 1.96 is the constant for 95% confidence interval(Table 7). x1 and x2 were applied to the predict() function to calculated the dose and account for uncertainty.

**3 Result**

**3.1 Examine the data**

Arms 2 – 4 will not be the good choice as replacements for artimisin because of its low drug effect compared to arm 1(**Figure 1**). On the other hand, the side effects in arm 4 is similar to arm 1 which can be chosen as the least preferred replacement(**Figure 2)**. Arm 8 has highest drug effect as well as side effect which will not be chosen as the replacement because of its toxicity(**Figure 2)**. Therefore, arms 5 – 7 should be the most promising replacement for artimisin because they are more effective than artimisin but of course have higher side effect(**Figure 1 and 2)**.

**3.2 Examine the data**

The pairwise t-test results (**Table 1**) show that arms 5-8 rejected H0, therefore, we can consider arm 5 -7 have significantly greater effect than artimisin. Arm 5 was significant at alpha = 0.001 while arms 6-8 was significant at alpha = <0.001. On the other hand, The pairwise t-test results (**Table 2**) show that arm 8 rejected H0, therefore, we can consider arm 8 have significantly greater side effect than artimisin. Arm 8 was significant at alpha = <0.001. To conclude, arm 7 (1000 mg) of AMP1050 should be chose as it results in a better effect than the artemisinin positive control, and has a side-effect profile that is not inferior.

**3.3 Model Based Analysis**

**3.3.1 Visualization of correlation between drug effect and Cmax(AMP1050)**

Figure 3 shows that the maximum drug concentration (Cmax) covaries with the drug effect. The scatter plot shows a clear covariation between the variables but not linear as the linear regression line(blue line) does not go through the points at all (figure 4). Red line and green line represent how Emax model and sigmoid model fit the data and it seems that sigmoid model show better performance(figure 4). Figure 5 shows effect curve including parameter uncertainty on sigmoid model, the blue shadow gives the visualisation of uncertainty by calculating the confidence intervals on each x-value(figure 5).

**3.3.2 Fitting linear model, Emax model and sigmoid model to data**

The linear regression equation shows that parasite level decreases by is 0.045log parasite as the Cmax increases by 1.00 ng/mL (the regression coefficient (slope), is 0.045)(Table 3). The coefficient of determination (R^2) shows that 90.0% of the variation in drug effect between different individuals can be explained by the individuals having different Cmax. Since the regression coefficient is positive and the correlation coefficient (+-0.94)are positive, so it means that there is a strong, positive, linear relationship between the variables Cmax and drug effect)(Table 3). The Emax model and sigmoid model(Table 4 and 5) shows that maximum effect of the drug will be a 237.8 and 120.8 log parasite reduction from 0 to 6 hours. Half of that effect is reached when Cmax reaches 3798.4 ng/ml and 1360.7 ng/ml respectively. The most appropriate model of the 3 models would be sigmoid model because of the lowest AIC. (AIC of linear model, Emax model and sigmoid model are 5420.571, 5317.266 and 5301.503 respectively).

**3.3.3 Cmax range and highest dose selection**

**3.3.3.1 Minimum Cmax**

Confidence interval (ci) of 500 mg positive control drug(artimisin) was calculated(red lines in figure 6), the upper boundary was around 8.143 and the lower boundary was 6.488, the average drug effect was 7.3156. Minimum Cmax was 214.3883 ng/ml. For uncertainty, the lower boundary of drug effect (artimisin) 6.488 drop within the 95% ci(fit = 6.490, lwr = 6.356, upr = 6.945).

**3.3.3.2 Maximum Cmax**

The maximum Cmax was 984.1052 ng/ml. For uncertainty, the mean of probability of getting side effect(artimisin) was 0.2916667, which drop within the 95% ci(fit = 0.2916186, lwr = 0.03206983, upr = 0.5511673).

**3.3.3.3 Highest dose selection and validation**

The dose range from 792.1327mg(fit = 792.1327,lwr = 774.4358, upr = 809.8296) to 1123.4mg(fit = 1123.324, lwr = 1101.229, upr = 1145.419).

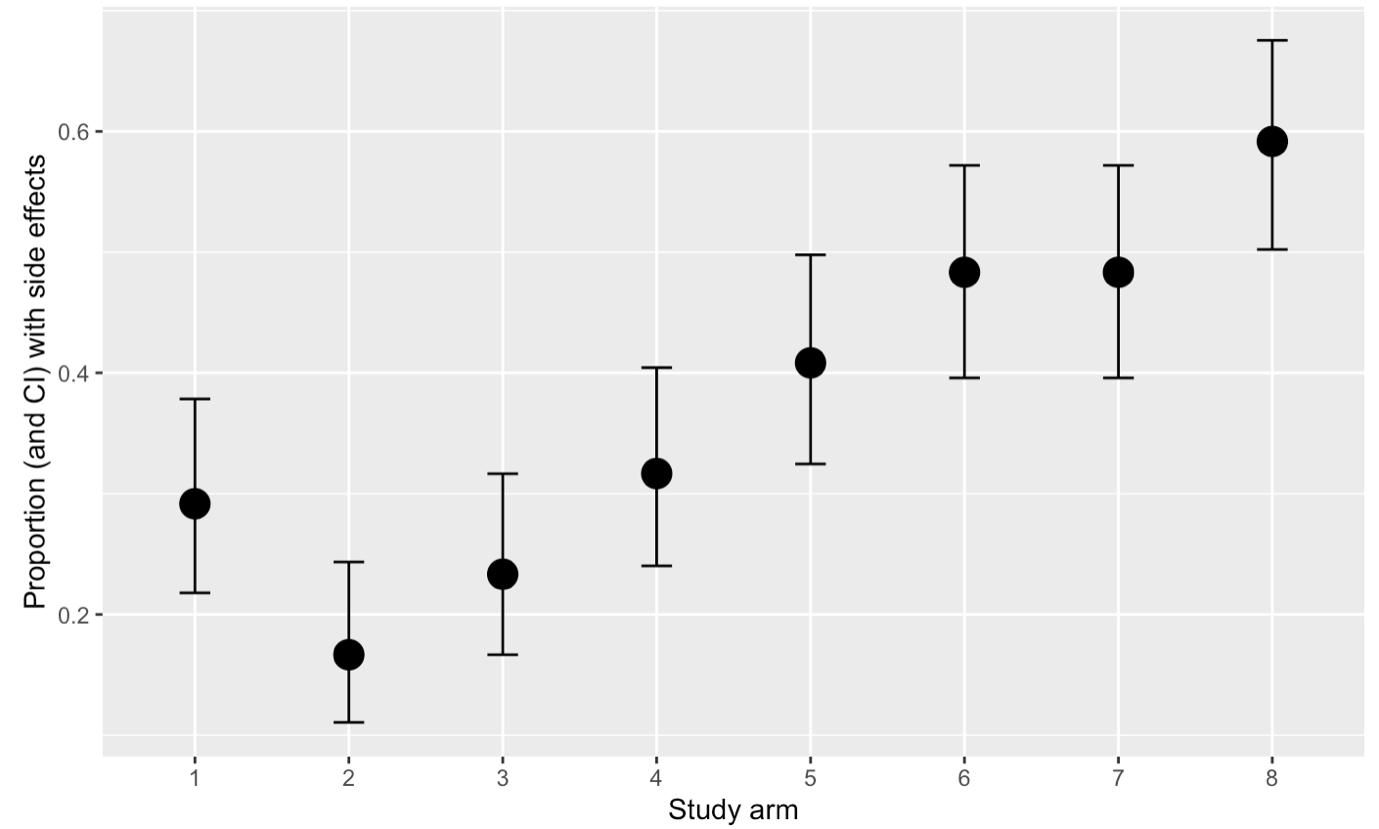
Figure 7 shows the validation of dose selection, the red boundary represent the Cmax range and prove that the highest dose should be around 1000 mg(most drug with dose = 1000 mg drop within the red boundary).

**4 Conclusion**

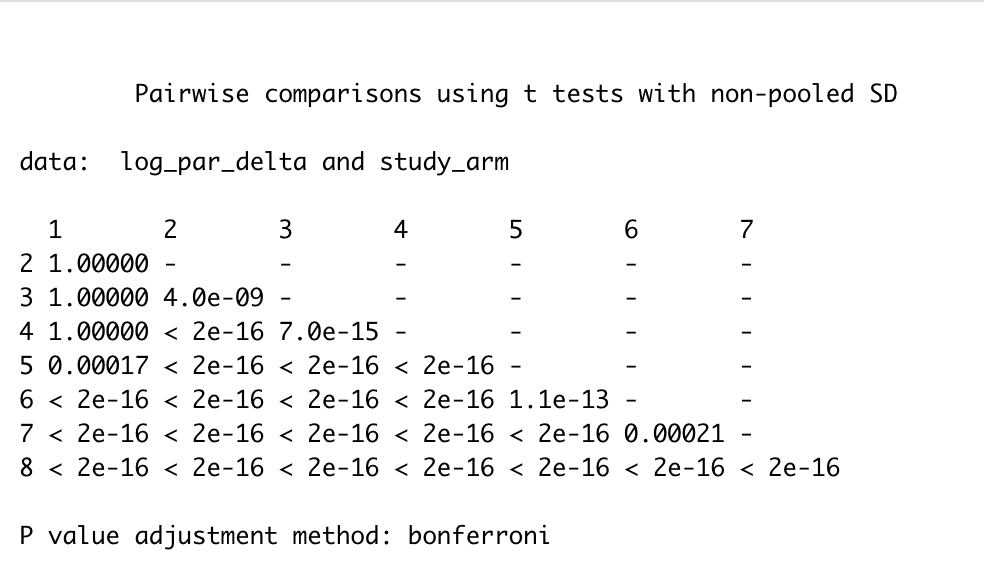
The highest dose from statistical and model part were similar(1000 mg and 1123 mg). A further validation of the dose selection should be included in coming study.

****

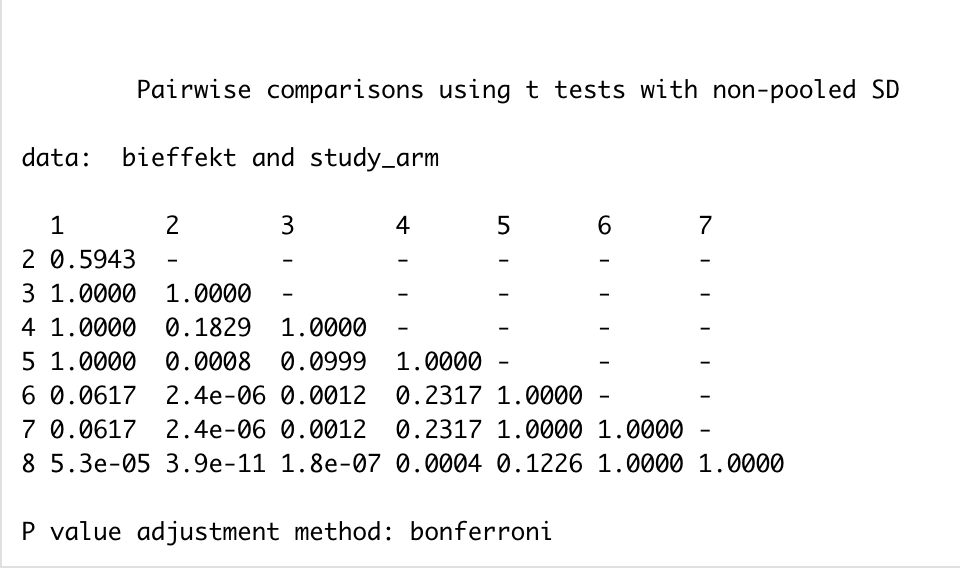
**Figure 1:** Effect variable with respect to study arms

****

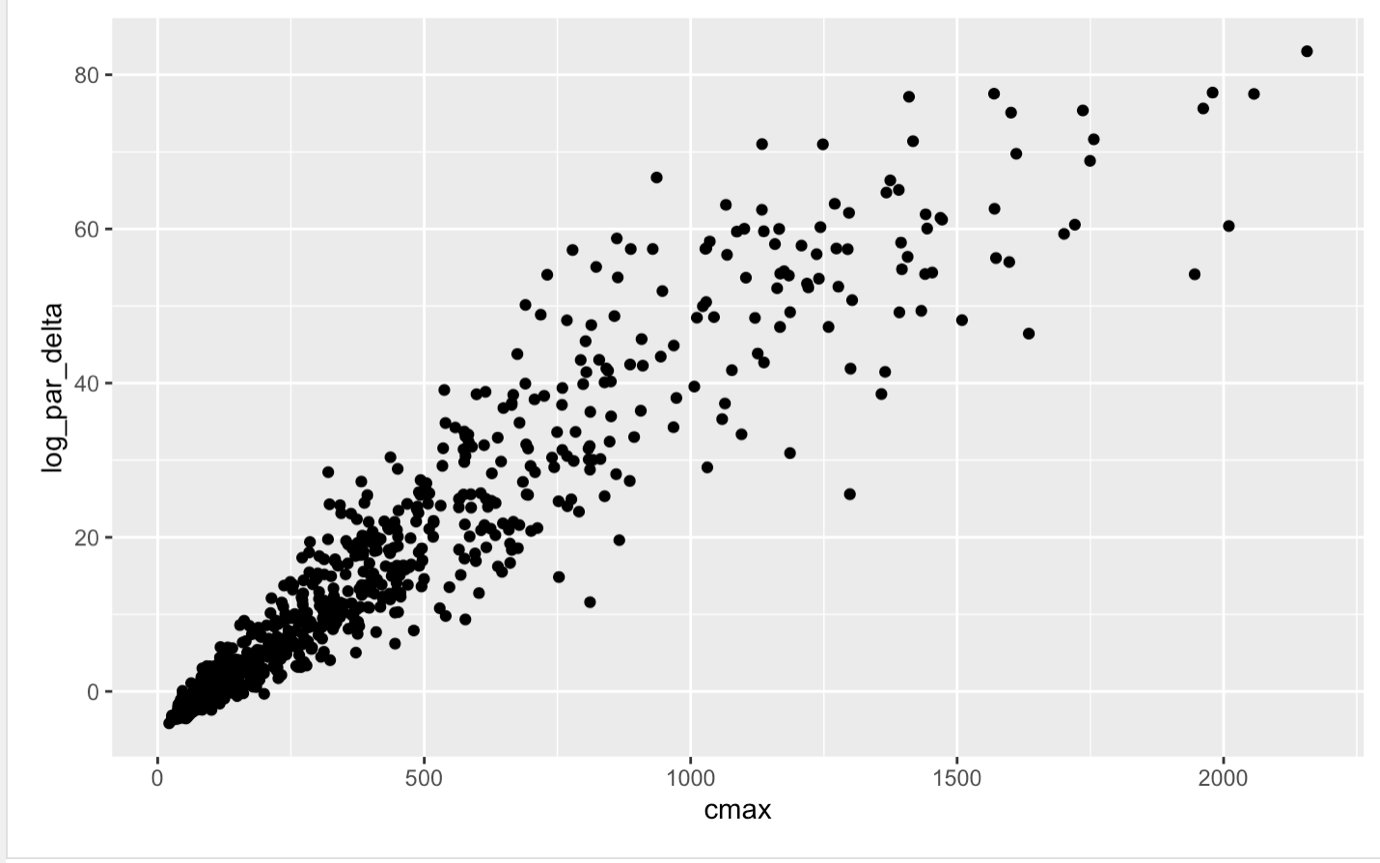
**Figure 2:** Side effect variable with respect to study arms

****

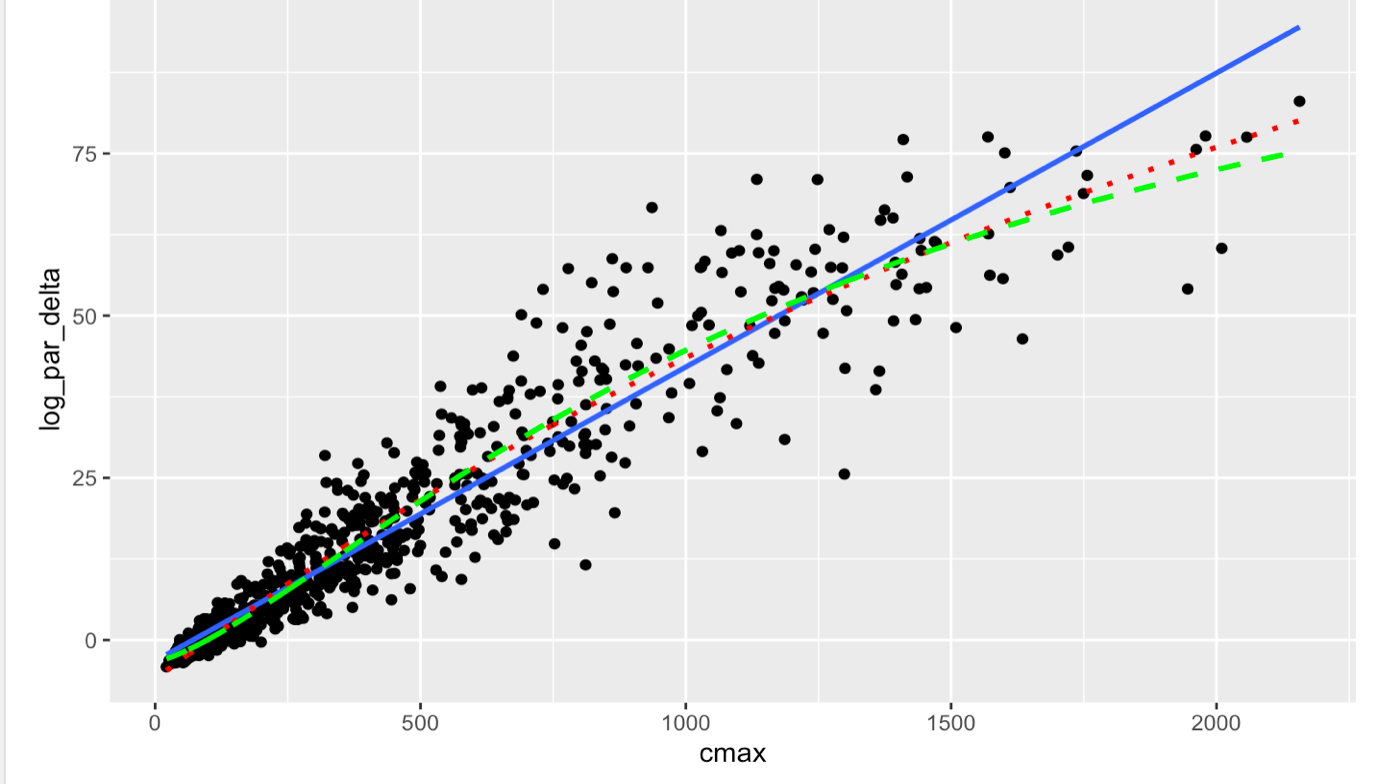
**Table 1** Pairwise comparison using t tests(drug effect)



**Table 2** Pairwise comparison using t tests(side effect)

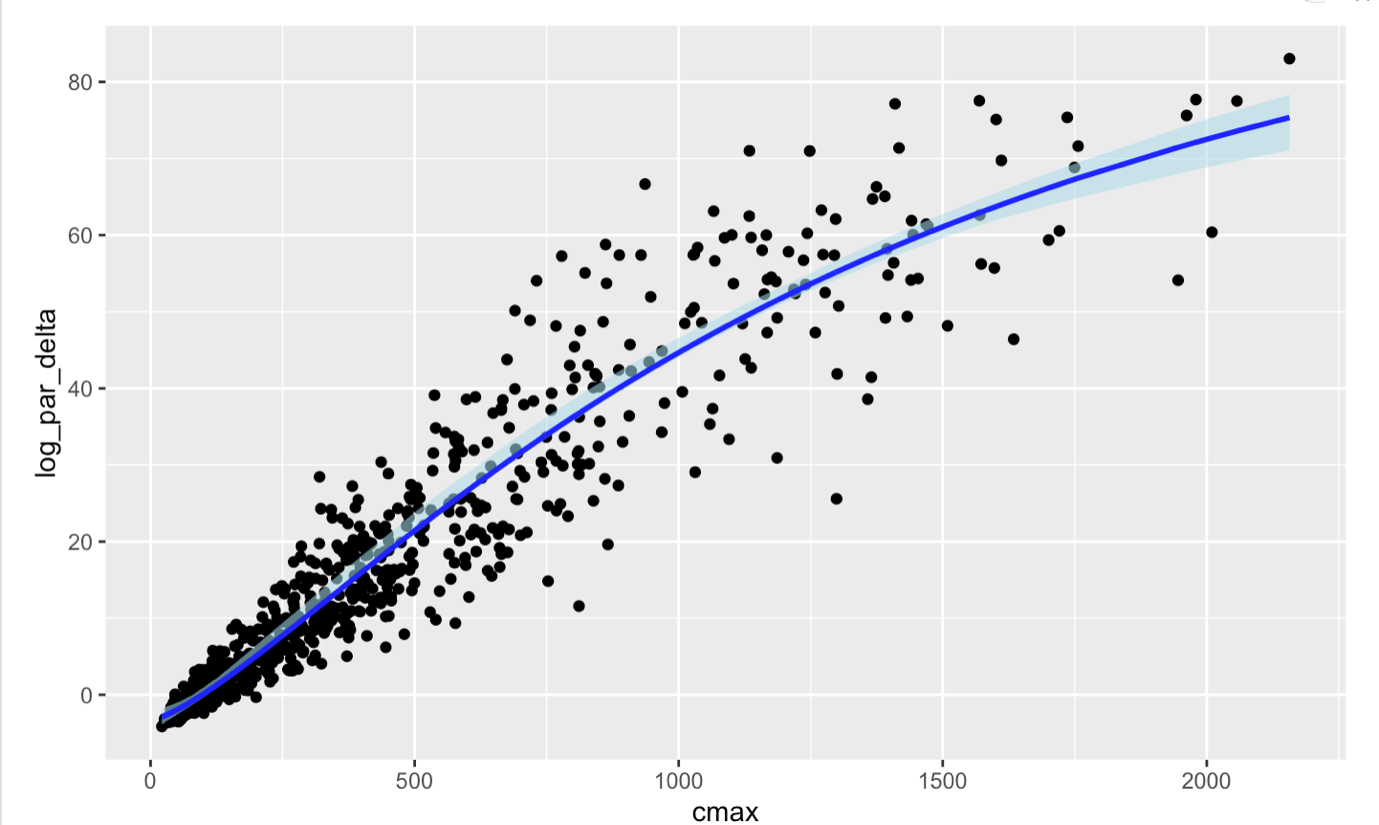
****

**Figure 3** Correlation between drug effect and Cmax

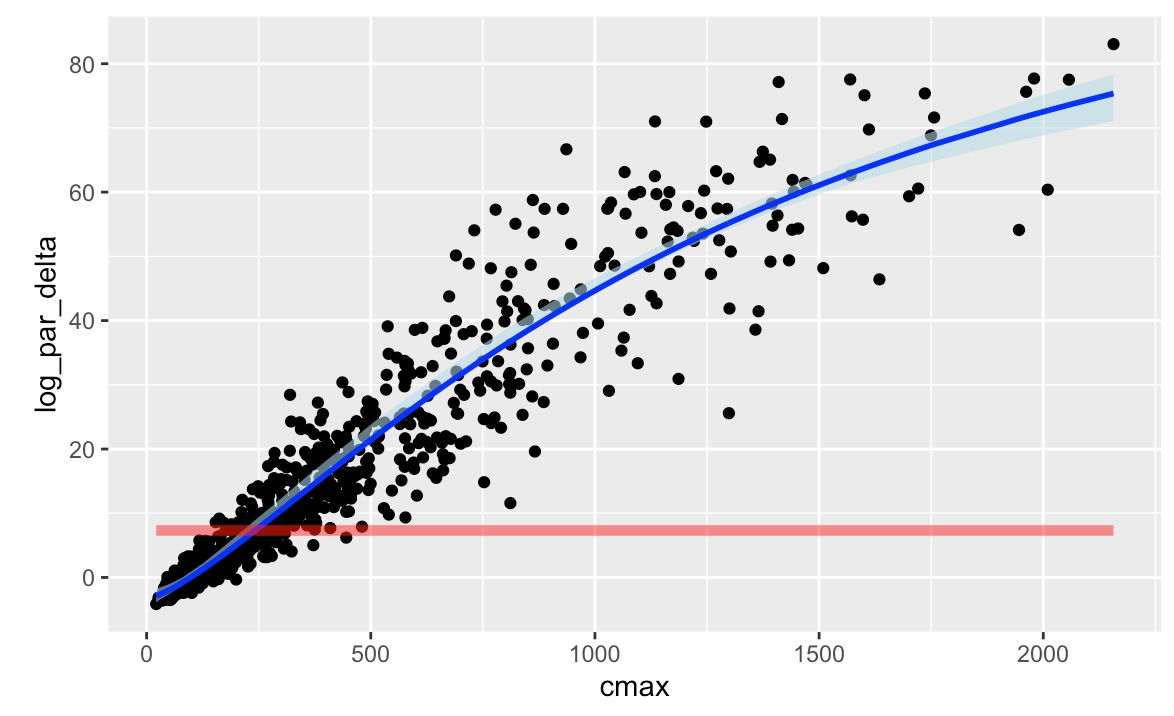


**Figure 4** Correlation between drug effect and Cmax with a linear model,

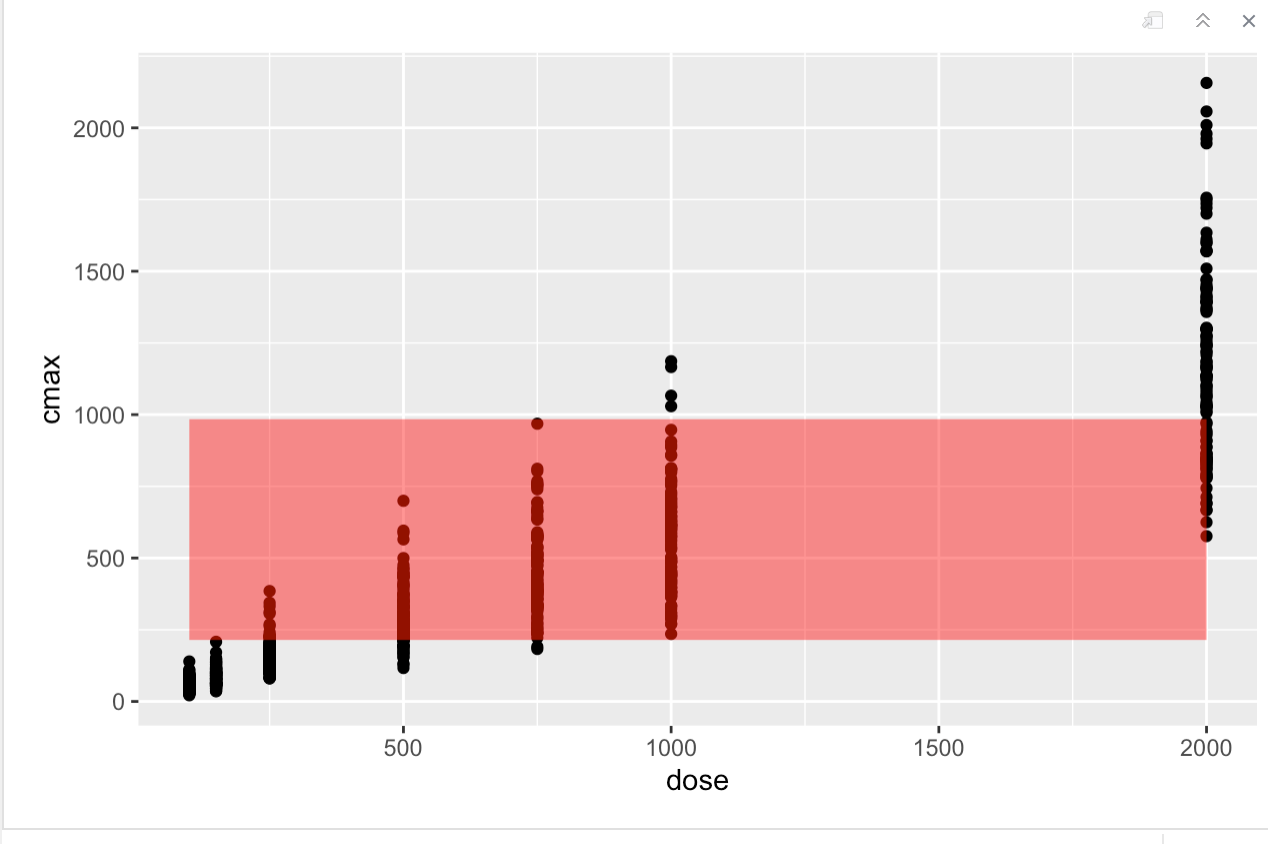
Emax model and sigmoid model



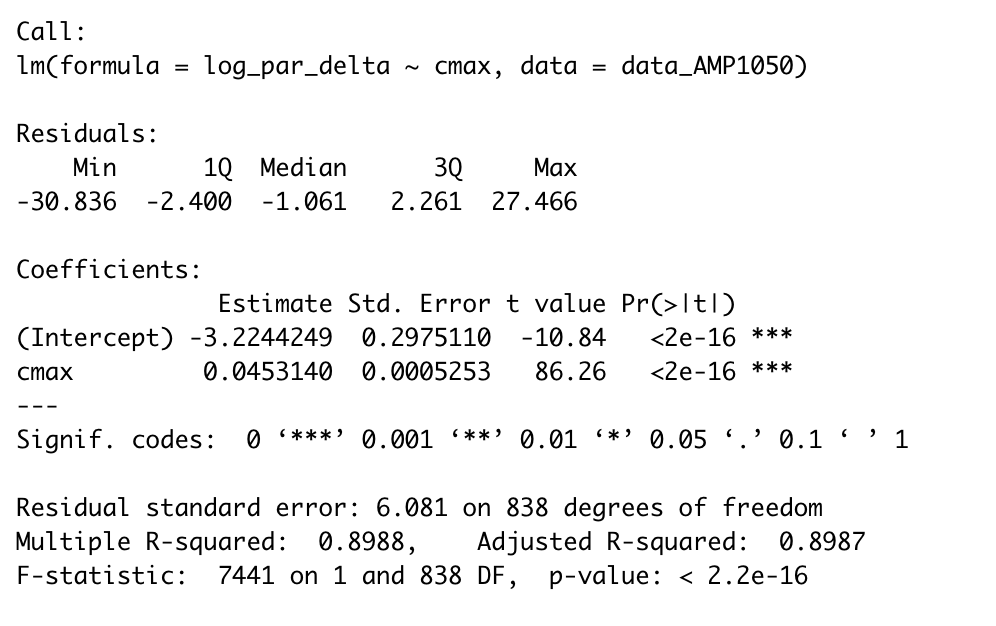
**Figure 5** Effect curve including parameter uncertainty on sigmoid model



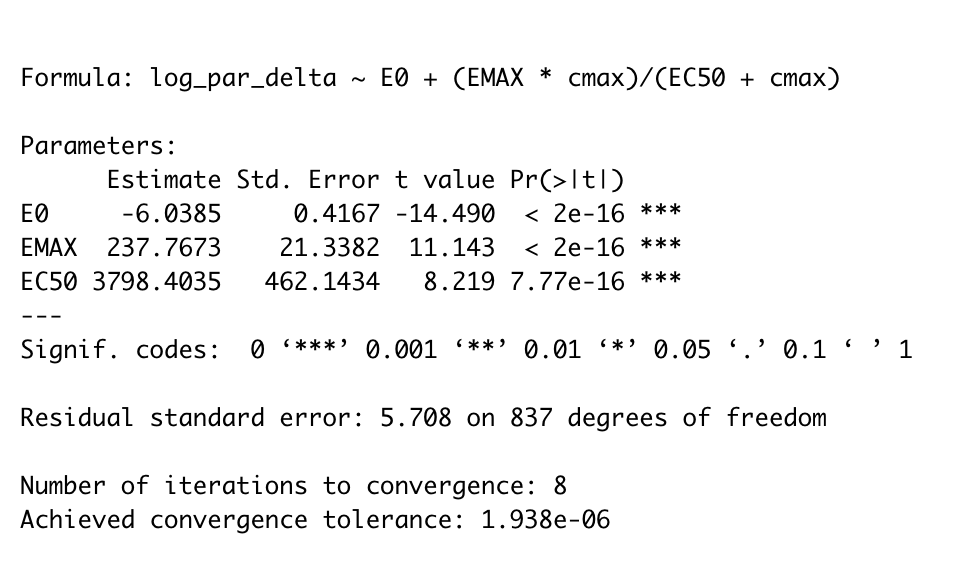
**Figure 6** Prediction of 500 mg artimisin average effect with confidence interval on effect curve



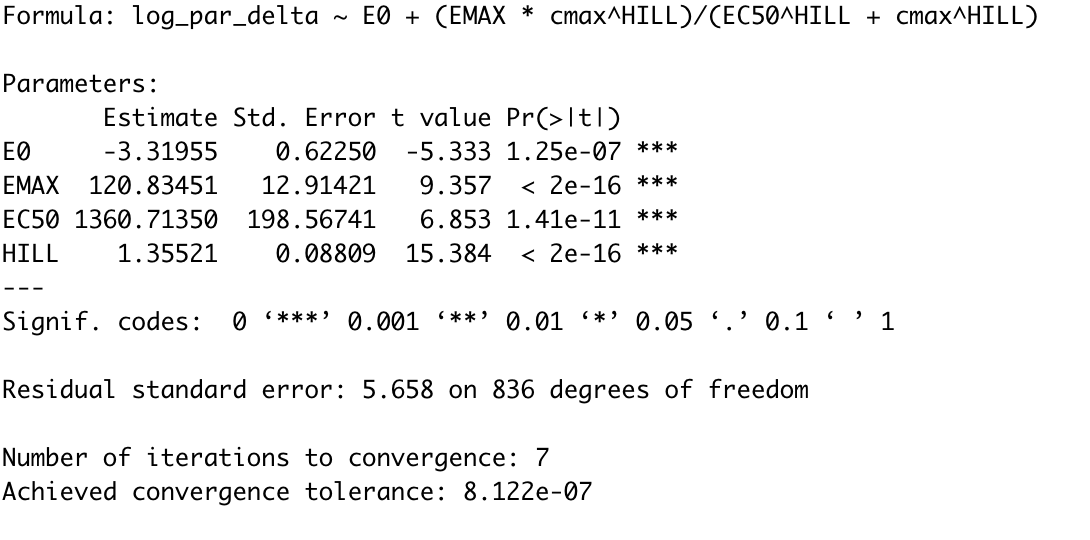
**Figure 7** Validation of dose selection



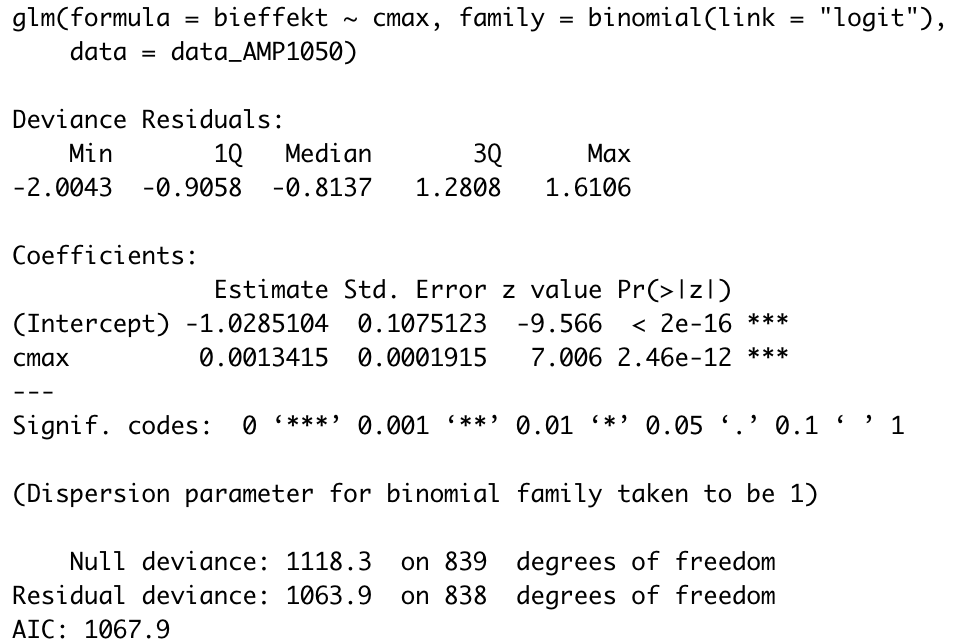
**Table 3** Linear-regression model of drug effect



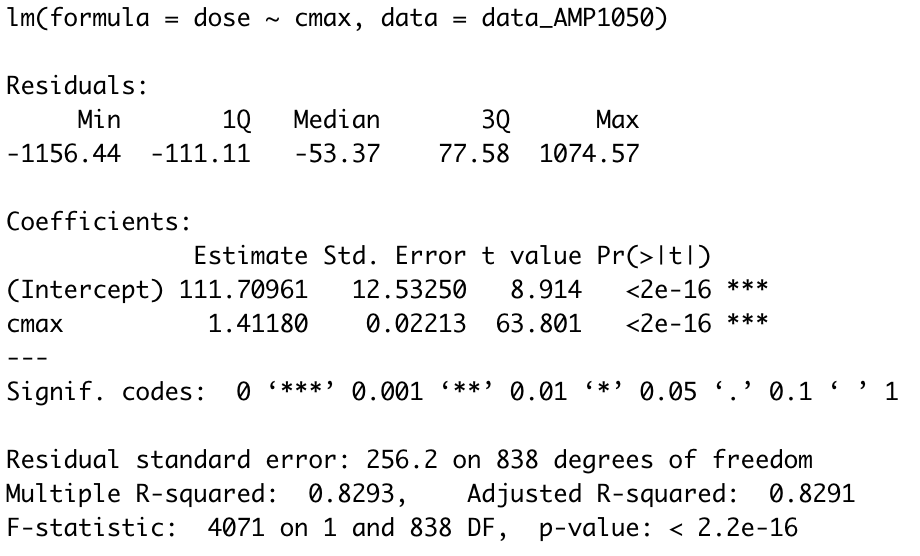
**Table 4** Emax model of drug effect



**Table 5** Sigmoid Emax model of drug effect



**Table 6** Logistic-regression model for the probability of side effects in the study arms with AMP1050



**Table 7** logistic-regression model for dose prediction

**Part B: Machine learning project**

**1 Introduction**

The goal of the project is to classify the bioactivity of drug based on Lipinski molecular descriptors. The target protein is a Tyrosine ABL kinase which caused chronic myelogenous leukaemia (CML) if mutation occurred. A binary classification problem, where features (X) are Lipinski molecular descriptors while target vector (y) for classification is bioactivity of the drug(active/inactive), is shown in the project. The bioactivity data is expressed as IC50 values(Active: <1000nM; inactive: > 10000nM; other: intermediate). All data were obtained from ChEMBL Database.

**2 Methodology**

**2.1 Exploratory data analysis**

Summary of dataset was read. No missing value was found.

**2.1.1 Pair-plot of all variables and correlations investigation**

One class **‘Bioactivity\_class'** and 4 feature **'MW', 'LogP','NumHDonors','NumHAcceptors’** were selected for pair-plot and the pearson correlation was calculated.

**2.1.2 Class of bioactivity distribution in dataset**

Bioactivity distribution (active/inactive) was performed and showed in bar chart.

**2.2 Machine learning model**

**2.2.1 Data preparation**

Features(**'MW', 'LogP','NumHDonors','NumHAcceptors’**) and target (‘**Bioactivity**’) were selected as training and test sets. Common ‘stratify=y’ was set in order to ensure the random split between the percentage of ‘active’ and ‘inactive’.

**2.2.2 Models selection and training**

**Logistic Regression, Decision Tree, Random Forest and Gradient Boosting** were selected. Model accuracy was performed. 4 import features including **'LogP','NumHDonors','NumHAcceptors’** and **‘MW’** were used todetermine 1 variable ‘**bioactivity’.** Confusion matrix was made. **Logistic Regression** wasselected**,** the parameter C was used for sensitivity analysis on change of accuracy. K-fold cross validation was also performed.

**2.3 Model comparison**

The 4 selected models were compared based on accuracy on training and test set as well as the feature prediction. In addition, the accuracy in active and inactive prediction was performed.

**3 Result**

**3.1 Pair-plot of all variables and correlations investigation**

Pair-plot and correlations investigation(Figure 1 and 2) shows no or little correlation between variables. Figure 2 shows that **MW** and **NumHAcceptors** has significant correlation(0.52).

**3.2 Class of bioactivity distribution in dataset**

The distribution of active compounds was higher than inactive compounds(Figure 3 ). It is not a balanced dataset for bioactivity.

**3.3 Logistic Regression**

**C : 1 Training set accuracy: 0.812**

**C : 1 Test set accuracy: 0.820**

C : 1 Training set F1-score: 0.896

C : 1 Test set F1-score: 0.901

C : 0.1 Training set accuracy: 0.813

C : 0.1 Test set accuracy: 0.819

C : 0.1 Training set F1-score: 0.896

C : 0.1 Test set F1-score: 0.900

C : 100 Training set accuracy: 0.812

C : 100 Test set accuracy: 0.820

C : 100 Training set F1-score: 0.896

C : 100 Test set F1-score: 0.901

The parameter C determine the value of accuracy for both the training and test set. The lower the C, the stronger the regulation. The higher the C, the chance of overfitting increase, the trust on training set increases. Therefore, when C is lower, the accuracy of training set will be higher, while when C is higher, the accuracy of test set will be higher.

A 10-fold cross validation was performed on logistic regression as the optimal parameter C was determined. Logistic regression has good performance based on 10 fold cross validation.

Average accuracy after 10 fold cross validation :0.79 +/- 0.09

Average F1-score after 10 fold cross validation :0.72 +/- 0.05

Figure 4 shows that the rank of important feature are **LogP > MW > NumHDonor = NumHAcceptor.** The feature selection was based on the coefficient values (C1, C100 and C0.01). The positive coefficient values indicate important feature. Figure 5 shows the confusion matrix of logistic regression and the classification accuracy were **4%** and **100%** in inactive and active compounds.

**3.4 Decision Tree**

Figure 6 shows that the training accuracy is way higher than testing accuracy, the F1 score shows that the training set performs better than testing set. From figure 7, is was showed that the rank of important feature are **LogP > MW > NumHDonor > NumHAcceptor**. Confusion matrix of Decision tree showed that the classification accuracy were **47%** and **94%** in inactive and active compounds.

**3.5 Random Forest**

**Accuracy on test: 0.854**

F1-score on test set: 0.913

From figure 10, is was showed that the rank of important feature are **LogP > MW > NumHDonor > NumHAcceptor.** Confusion matrix of Decision tree showed that the classification accuracy were **57%** and **96%** in inactive and active compounds.

**3.6 Gradient Boosting**

**Accuracy on train: 0.906**

**Accuracy on test: 0.874**

F1-score on test set: 0.927

From figure 11, is was showed that the rank of important feature are **LogP > MW > NumHDonor > NumHAcceptor**. Confusion matrix of Decision tree showed that the classification accuracy were **41%** and **98%** in inactive and active compounds.

**3.7 Model comparison**

**3.7.1 Accuracy on testing and training set**

**Gradient Boosting** hashighest accuracy on both testing (0.906) and training set(0.874).

**3.7.2 Feature importance**

All 4 models give the same result for feature ranking**: LogP > MW > NumHDonor > NumHAcceptor**, which indicated that **LogP** is the most important feature.

**3.7.3 Classification accuracy on active drug prediction**

Figure 13 shows that the accuracy on active drug prediction of the 4 models, which Logistic Regression has the best performance (100%), Gradient Boosting(98%) has higher accuracy than Decision Tree(96%) and Random Forest (94%).

**3.7.5 Classification accuracy on inactive drug prediction**

Figure 14 shows that the accuracy on inactive drug prediction of the 4 models, which Logistic Regression has the worst performance (4%), Random Forest (57%) has higher accuracy than Decision Tree(47%) and Gradient Boosting (41%).

The high accuracy for active drug predation may due to samples unbalance, the number of active compounds is almost 4 times than inactive compounds.

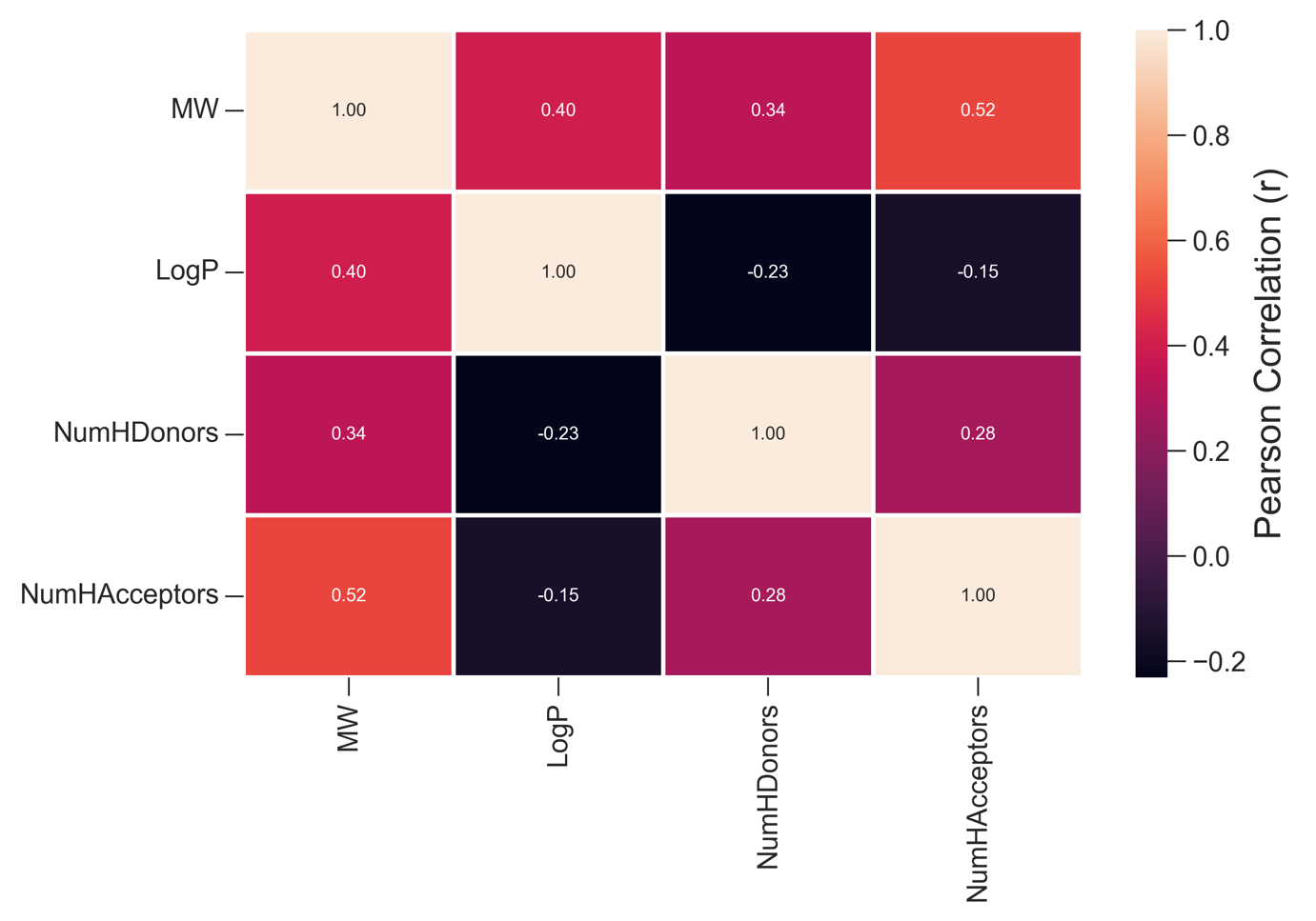
**4 Conclusion**

**Gradient boosting** seems to providing the best performance.

Feature selection suggests **LogP** and **MW** are the most important factor for the successful prediction of bioactivity of the drug.



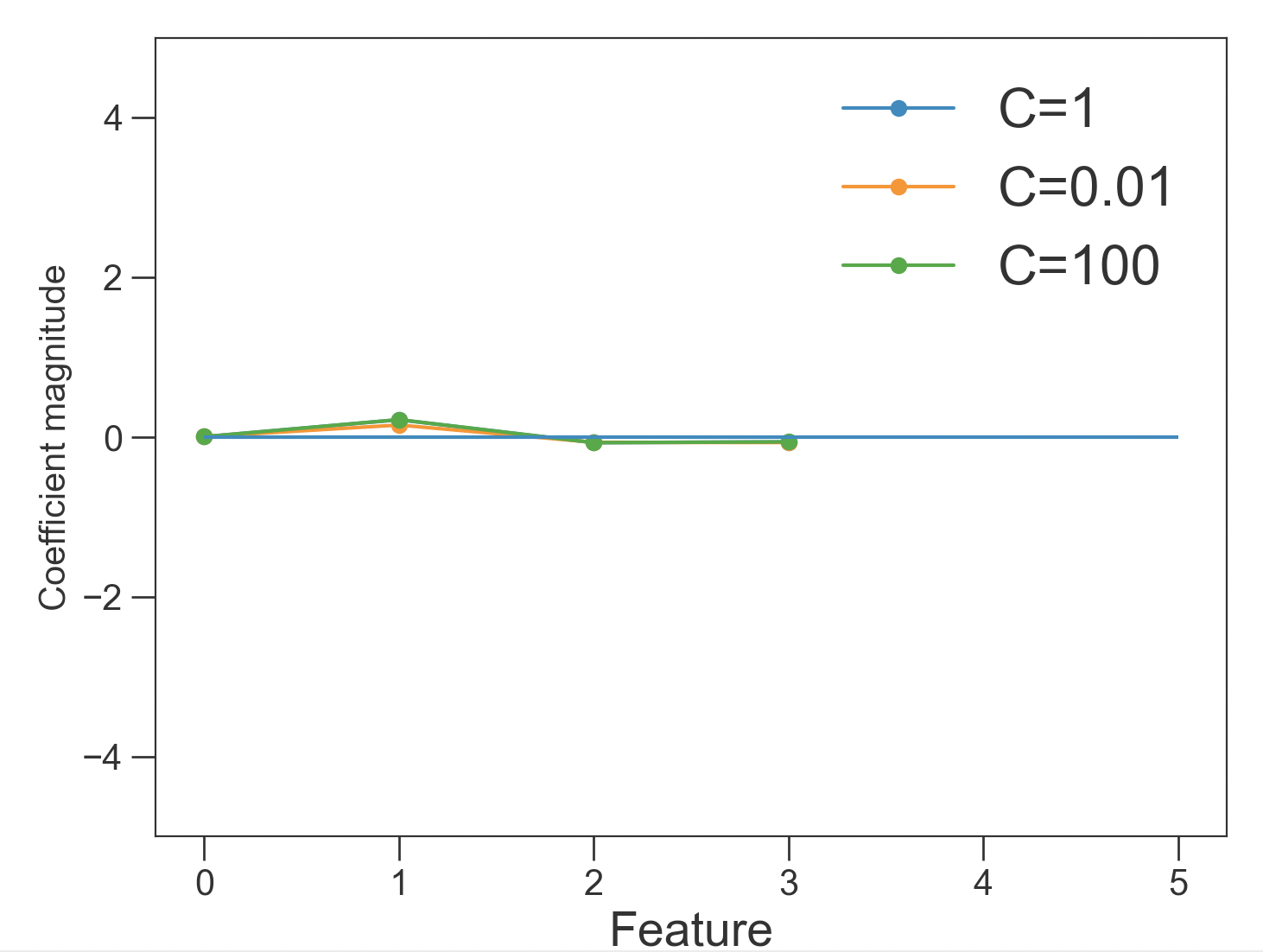
**Figure 1** Pair-plot of all variables



**Figure 2** Pearson correlation of variables

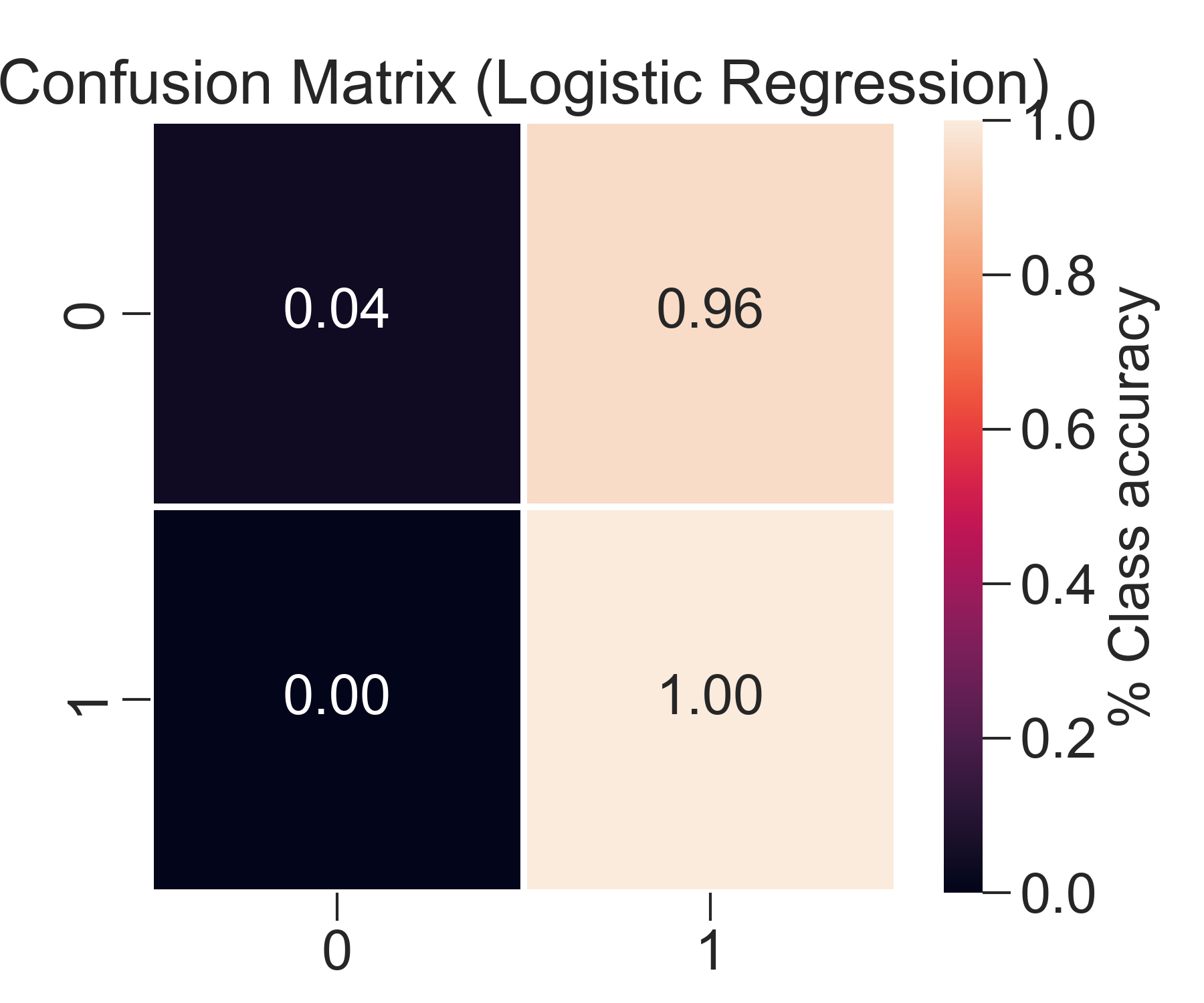


**Figure 3** Class distribution of active and inactive class



**Figure 4** Feature determination of logistic regression

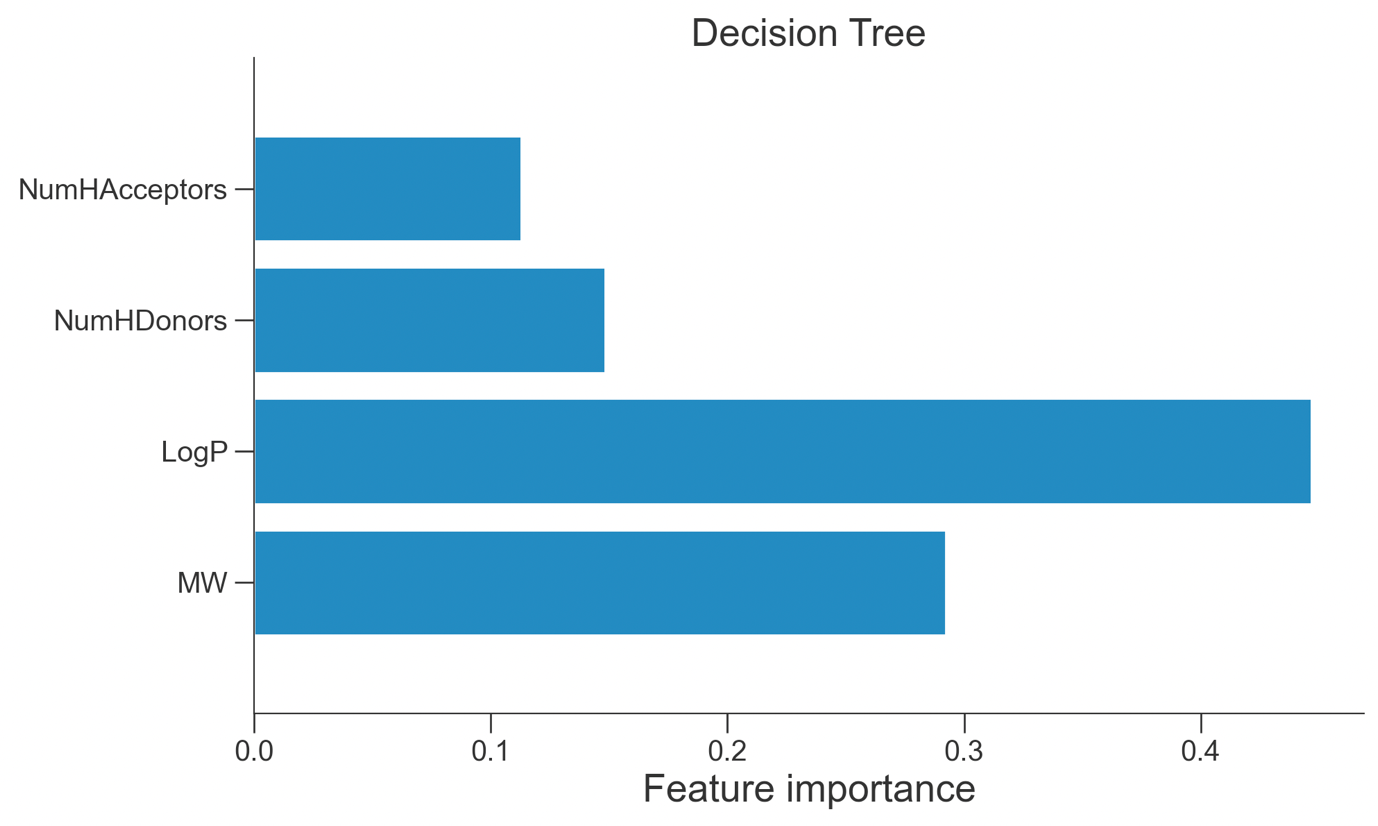
(0:MW, 1:LogP, 2: NumHDonor, 3:NumHAcceptor)



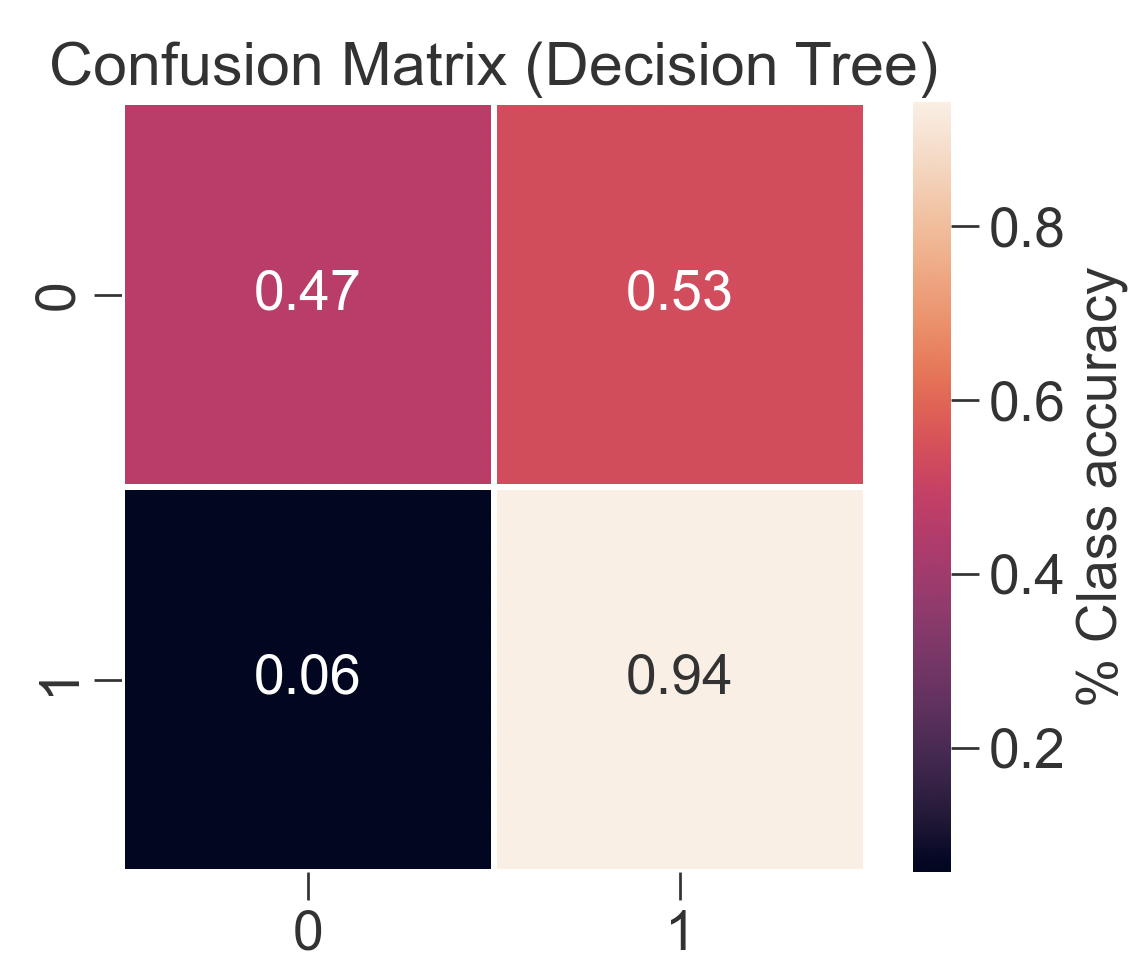
**Figure 5** Confusion matrix of logistic regression



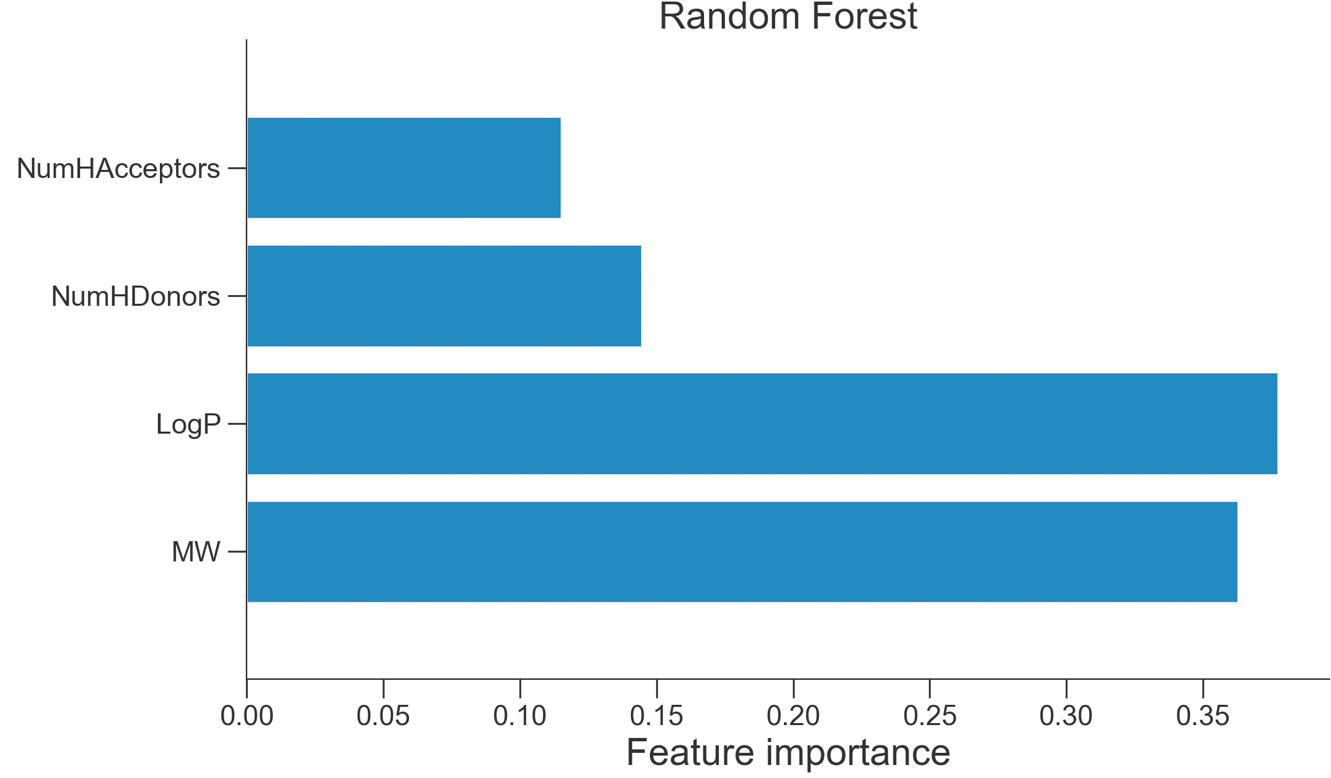
**Figure 6** Accuracy of training and testing set of decision tree



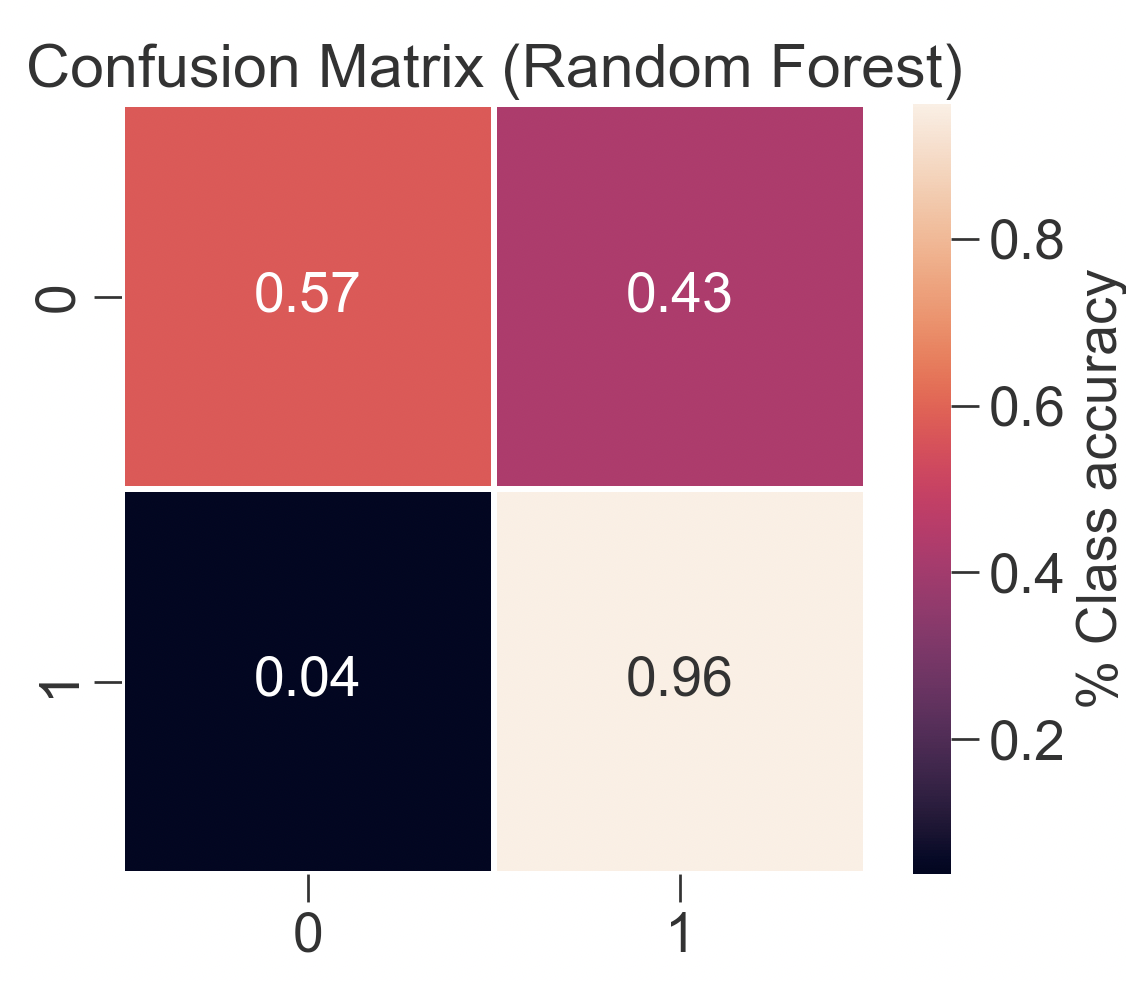
**Figure 7** Feature determination of Decision tree



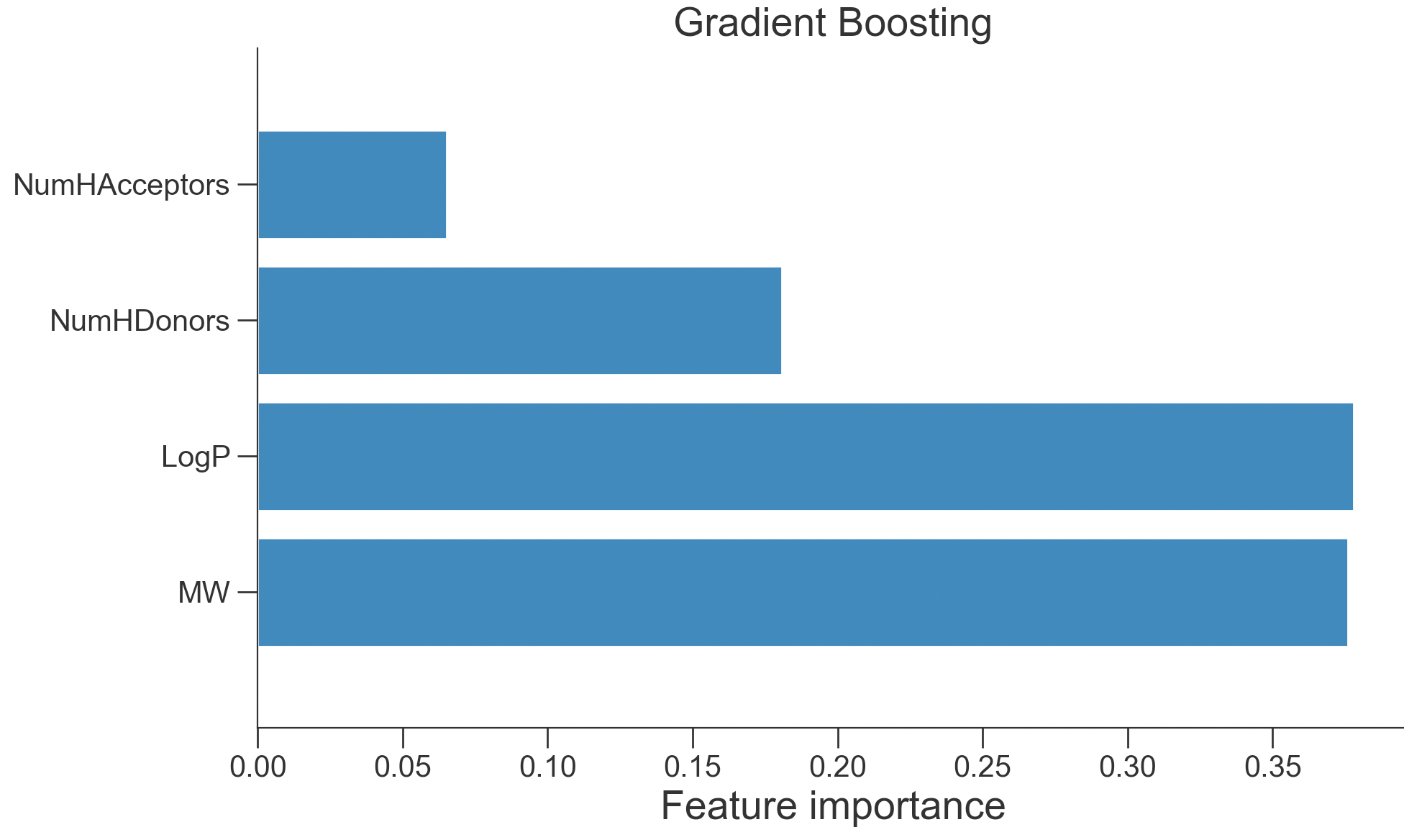
**Figure 8** Confusion matrix of Decision tree



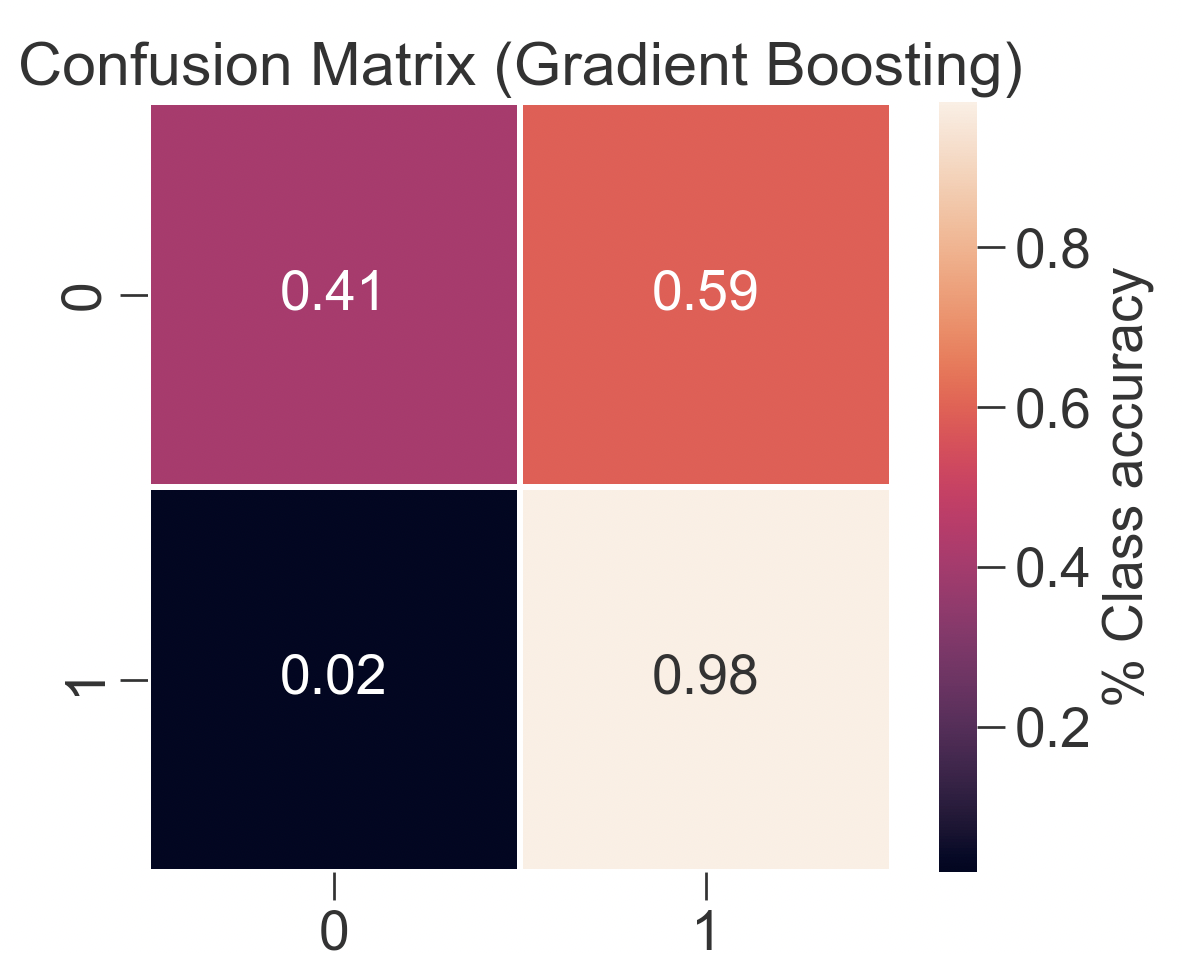
**Figure 9** Feature determination of random forest

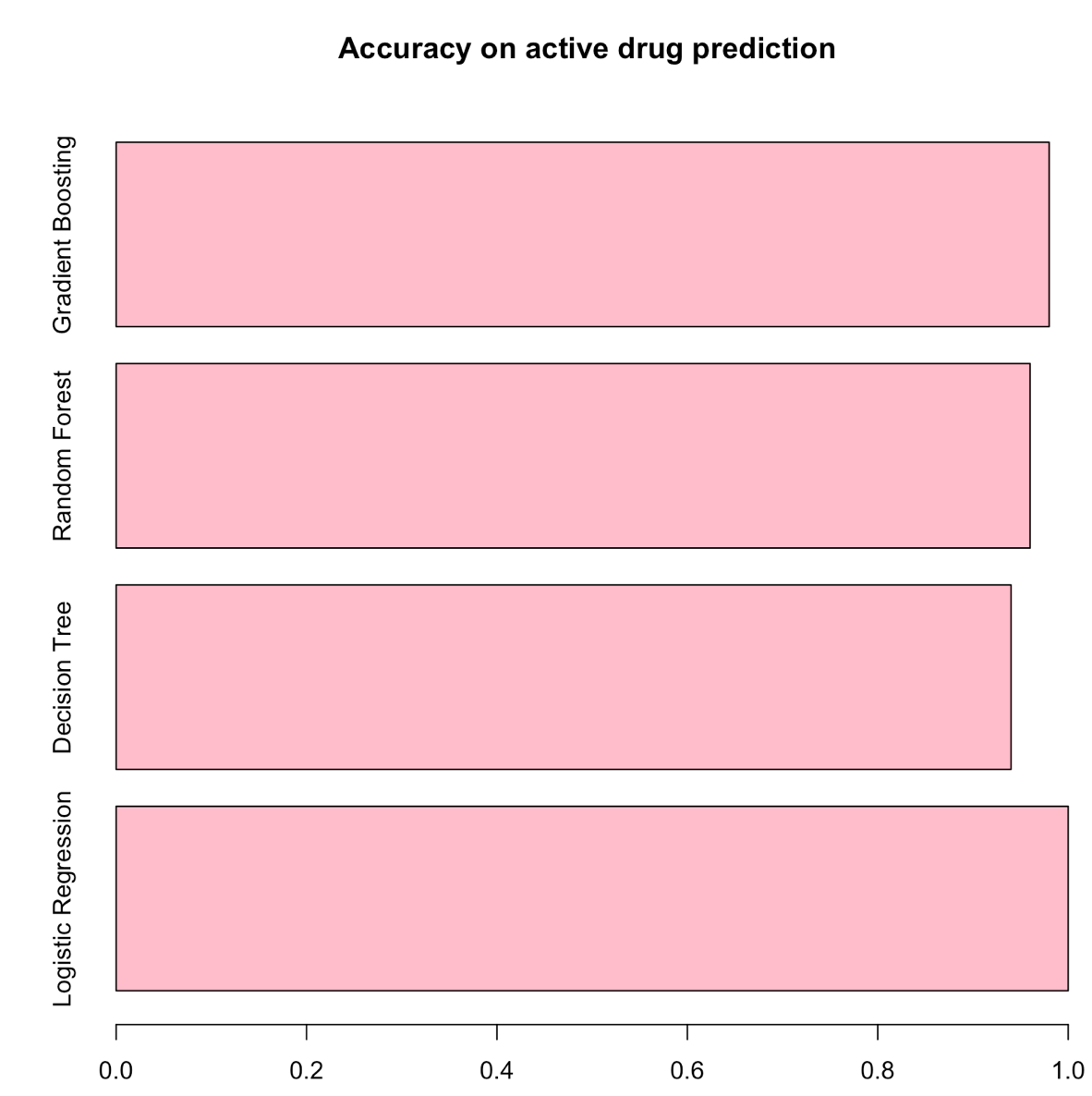


**Figure 10** Confusion matrix of random forest

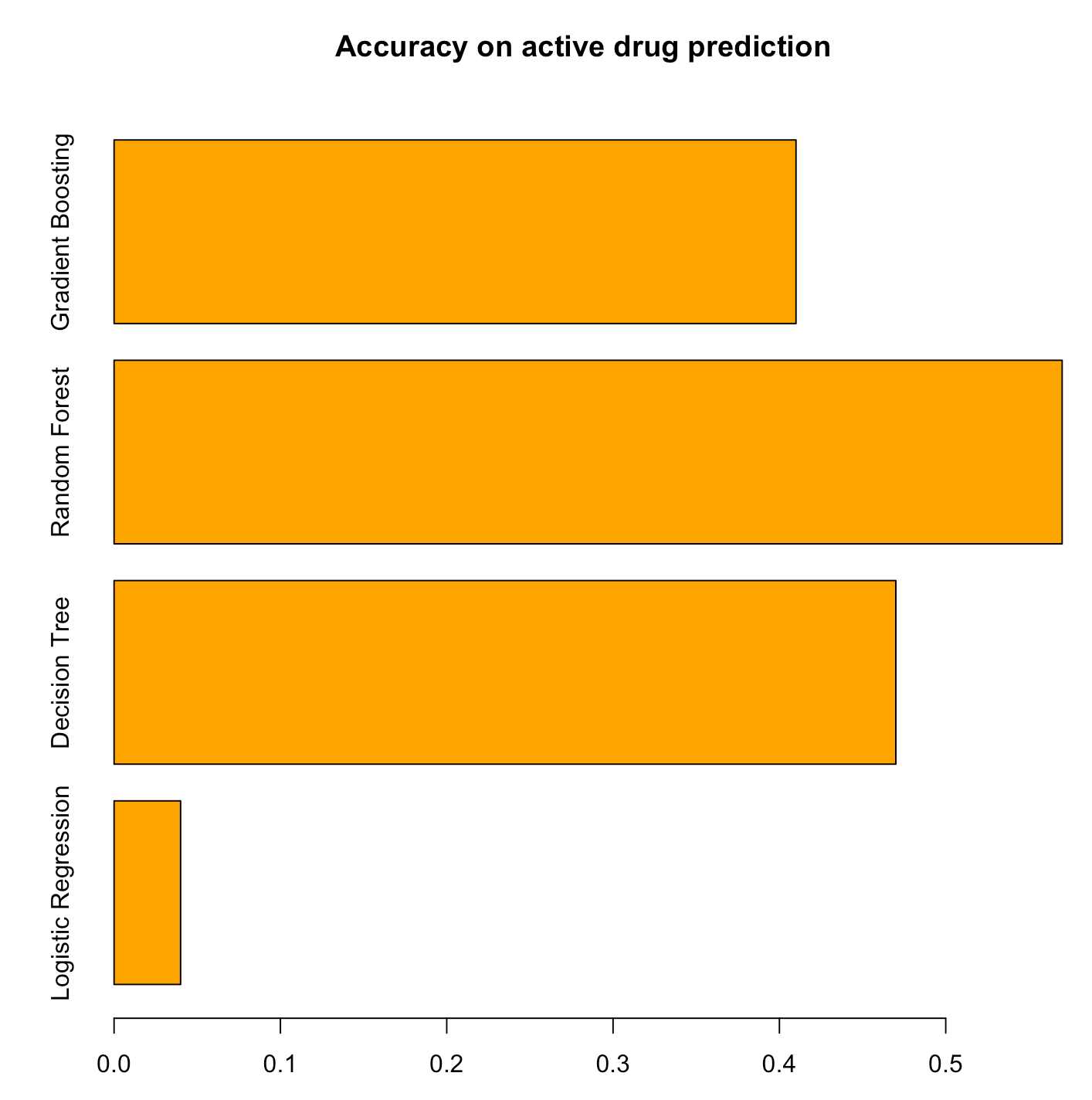


**Figure 11** Feature determination of gradient boosting

**Figure 12** Confusion matrix of gradient boosting



**Figure 13** Accuracy on active drug prediction



**Figure 14** Accuracy on inactive drug prediction