Evaluation(Stability):

Data collection：
We collected the three datasets GSE51675, GSE1563 and GSE46474 respectively because they all have three columns of gene, age and gender.

Data cleaning：

We need to get the correct ID in gse1, gse2 and gse3, we first collect the column REFSEQ, then we split the ID we need with the strsplit() function, last we use duplicated() to delete the duplicate ID.

We use the exprs() function to extract the expression matrix information, the pData() function to extract the clinical information, then determine the outcome, and use the model.matrix() function to construct the experimental design matrix. Given a series of sequences, a linear model and Bayesian test were fitted to each gene, and the top 1500 gene tables were extracted to obtain gene1, gene2, and gene3, respectively.

From the Venn diagram, it can be seen that there are four intersecting genes in the three gene tables, namely NM_014654, NM_012474, NM_005526 and NM_006634.

Data modelling：

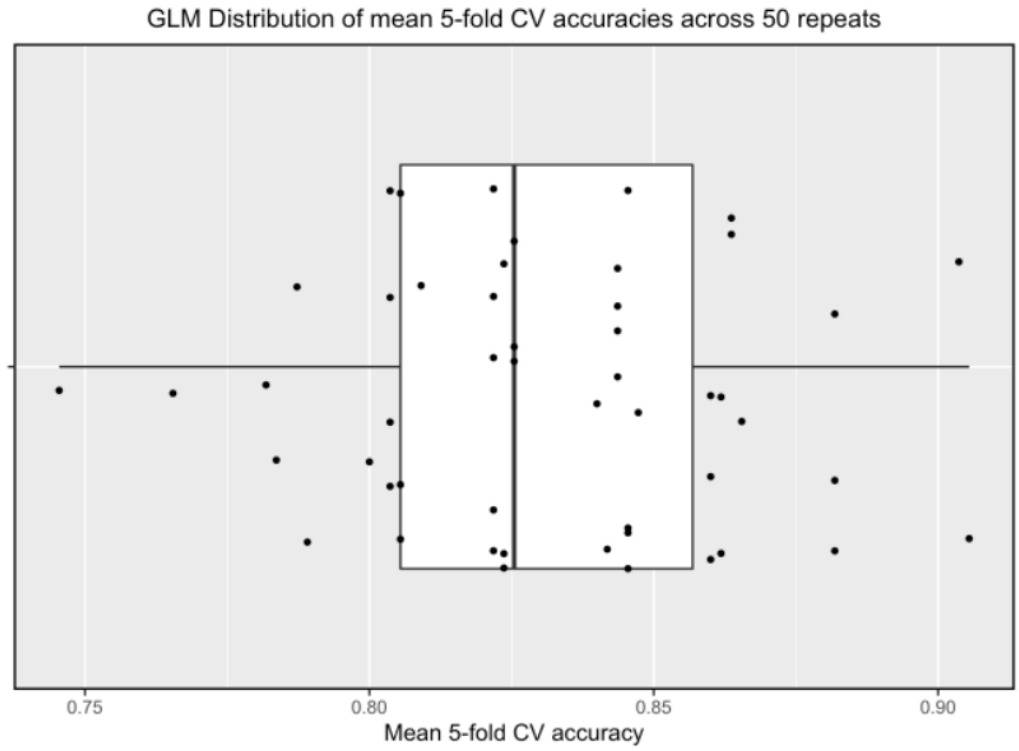
1. GLM(general linear model)
    1. impact of parameters

```
    Null deviance: 57.901  on 50  degrees of freedom
Residual deviance: 11.411  on 44  degrees of freedom
AIC: 25.411

Number of Fisher Scoring iterations: 11
```

    2. outcome

    The accuracy of the random forest model is more than 0.825, which means that it is accurate. Also, the stability is good in this model. The identification of system parameters in the general or multi-parametric linear model can give rise to serious difficulties, particularly when only a small amount of data is available. The problem is centred on the degree of collinearity in the input to the system. A high degree of collinearity causes ill-conditioning, and small errors in the input—output data result in severe oscillations in the derived kernel function. This effect is examined by analysing the system both in the time domain and in the transform domain, and in each case an index is derived which measures the degree of collinearity present.



2. RF（Random forest）
    1. impact of parameters

```
        OOB estimate of  error rate: 17.65%
Confusion matrix:
          Rejection Stable class.error
Rejection         6      7  0.53846154
Stable            2     36  0.05263158
```

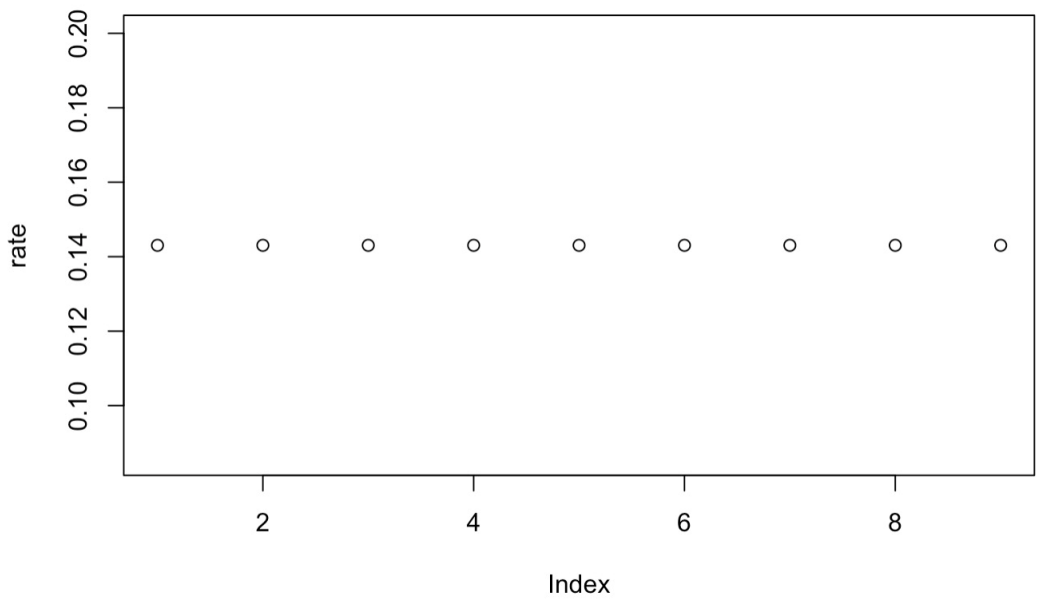
We also use the importance() function to determine the importance of the random forest model. It is obviously that NM_014654, NM_012474 and NM_005526 are more important variables, and NM_006634 is not such important in this random forest model.

```r
importance(rf_fit, type = 2)
```

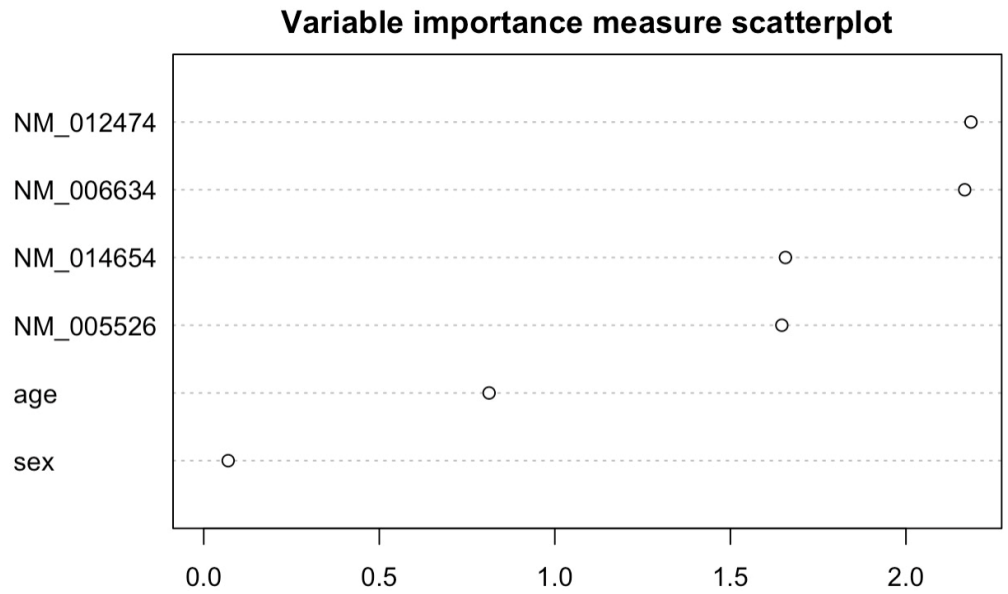
```
          MeanDecreaseGini
age              2.6430262
sex              0.1988364
NM_014654        4.1459962
NM_012474        4.1470728
NM_005526        5.4547362
NM_006634        2.8853321
```

Last, we calculate the mean misjudgment rate of the display model, which is 0.143.

```{r}
for(i in 1:(n-1)){
rate[i]<-mean(rf_fit$err.rate)}
plot(rate)
```




```
varImpPlot(x=rf_fit,sort=TRUE,n.var=nrow(rf_fit$importance),main="Variable importance measure scatterplot")
```



2. outcome

The accuracy of the random forest model is more than 0.84, which means that it is accurate. Also, the stability is good in this model. The prevalence of intrinsic stability of the RF demonstrates that the instability of RF not only comes from data perturbations or parameter variations, but also stems from the intrinsic randomness of RF. This finding gives a better understanding of RF

stability, and may help reduce the instability of RF.