

1

.CEL names in the raw data column are samples.

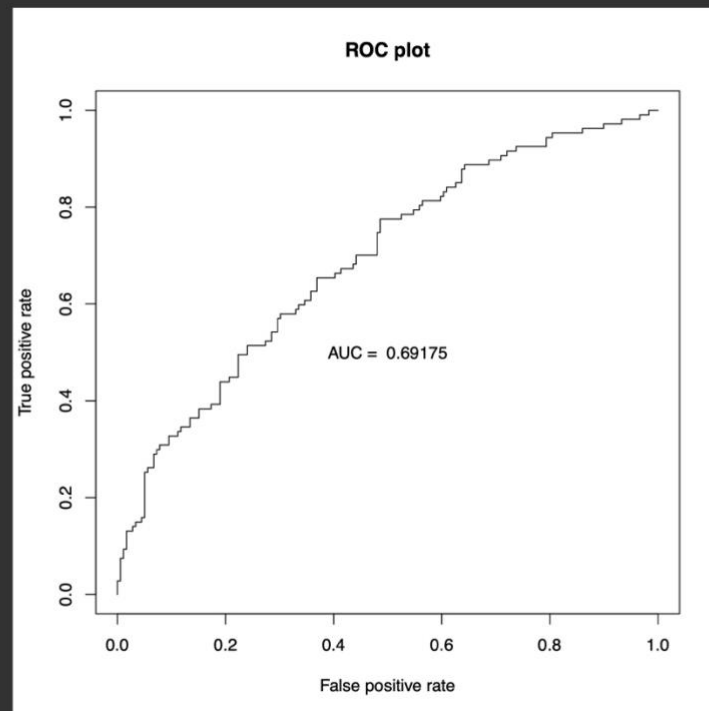
```
Terminal
mscstudent@lugh:/data/MSc/phyto/genomics2_p5
> header=colnames(rawdata)
> rawdata=rawdata[which(!is.na(rawdata[,3])),]
> dim(rawdata)
[1] 12302 289
> head(rawdata)
  probes ID symbol GSM36777.CEL GSM36778.CEL GSM36779.CEL GSM36780.CEL
1  10_at  10  NAT2    2.198943    2.198943    2.198943    2.198943
2  100_at 100   ADA    3.940545    3.912155    3.767530    3.912155
3 1000_at 1000  CDH2    2.765116    3.167591    2.765116    2.769986
4 10000_at 10000 AKT3    5.604067    5.557574    4.110917    5.149921
5 10001_at 10001 MED6    3.271074    2.707269    3.271074    3.273654
6 10002_at 10002 NR2E3    2.198943    2.198943    2.198943    2.198943
  GSM36781.CEL GSM36782.CEL GSM36783.CEL GSM36784.CEL GSM36785.CEL GSM36786.CEL
1    2.198943    2.198943    2.198943    2.198943    9.239395    2.198943
2    3.879925    3.885030    3.879925    3.907050    2.917336    3.909865
3    3.101667    2.765116    2.765116    2.782106    2.765116    2.765116
4    5.327300    6.082360    5.557574    5.557574    5.327300    7.570681
5    3.633811    3.574127    4.097233    3.223557    3.801196    3.271074
6    2.198943    2.198943    2.198943    2.198943    2.198943    2.198943
  GSM36787.CEL GSM36788.CEL GSM36789.CEL GSM36790.CEL GSM36791.CEL GSM36792.CEL
1    2.198943    2.198943    2.198943    2.198943    2.198943    2.198943
2    3.912155    5.076506    3.907050    3.912155    3.871659    3.866376
3    2.765116    2.765116    3.136318    2.765116    2.849035    2.765116
4    5.488027    5.223223    5.415476    5.908093    6.860796    4.993403
```

2 A CEL file contains gene expression raw data from microarray procedures. It is binary file format. The CEL file format is widely used in genomics field. For identifying differentially expressed genes, the CEL file data are preprocessed before next analysis (normalization and statistical testing).

3 I would like to try to predict the ER.status (ER positive or negative breast cancer).

```
> head(clindata)
  PID GEO.accession.number lymph.node.status
146 277 GSM36777 negative
147 278 GSM36778 negative
233 798 GSM36779 negative
251 846 GSM36780 negative
212 765 GSM36781 negative
160 600 GSM36782 negative
  time.to.relapse.or.last.follow.up..months relapse..1.True ER.Status
146 79 0 ER+
147 50 1 ER+
233 132 0 ER+
251 84 0 ER-
212 147 0 ER+
160 66 0 ER+
  Brain.relapses..1.yes..0.no.
146 0
147 0
233 0
251 0
212 0
160 0
```

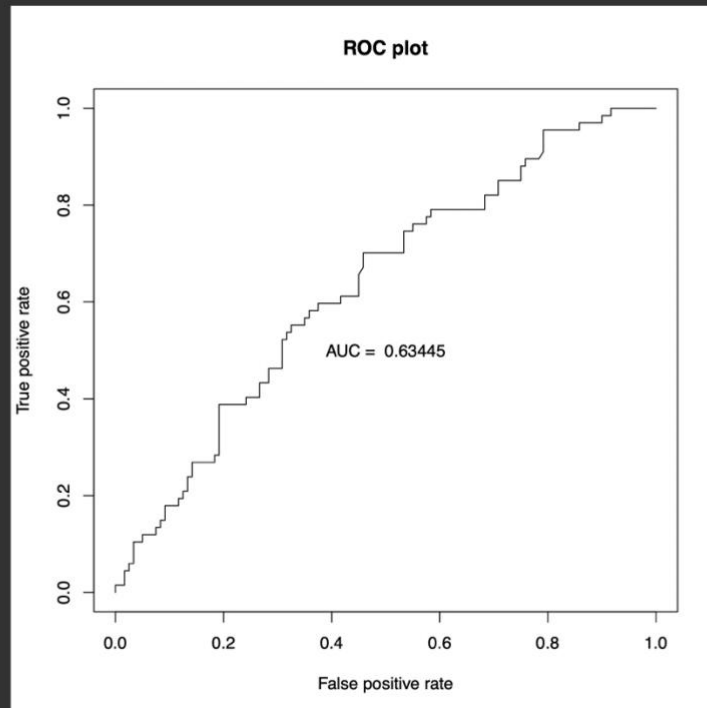
4 AUC value is 0.69175, it is good or acceptable AUC. In general, AUC value above 0.5 is desirable because it describes that model is better than random chance.



5 The dimensions of the predictor data is 189 and 12302.

```
gsm65321.cel 2.951880 2.165573 5.654798 7.422966 5.384437
PTBP3 KCNE2 DGCR2 CASP8AP2 SC02
gsm65316.cel 2.951880 2.165573 5.654798 7.422966 5.384437
gsm65317.cel 3.160646 2.165573 5.698952 6.660930 7.801790
gsm65318.cel 2.949961 2.165573 5.701854 5.751829 6.612244
gsm65319.cel 2.945755 2.165573 5.770306 6.296617 7.550900
gsm65320.cel 2.854822 2.165573 6.025985 6.835045 7.436133
gsm65321.cel 2.951880 2.165573 5.701854 7.118841 8.693071
> dim(predictor_data)
[1] 189 12302
>
```

6 AUC is 0.63445.



7 Sensitivity value is 0.2835.

```

> str(perf_ROC)
Formal class 'performance' [package "ROCR"] with 6 slots
 ..@ x.name      : chr "False positive rate"
 ..@ y.name      : chr "True positive rate"
 ..@ alpha.name  : chr "Cutoff"
 ..@ x.values    : List of 1
 .. ..$ : num [1:180] 0 0 0.00833 0.01667 0.01667 ...
 ..@ y.values    : List of 1
 .. ..$ : num [1:180] 0 0.0149 0.0149 0.0149 0.0299 ...
 ..@ alpha.values: List of 1
 .. ..$ : num [1:180] Inf 0.638 0.603 0.603 0.579 ...
> print(sens_out)
[1] "sensitivity=28.3582089552239"

```

8 Sensitivity value (0.2835) of this model is low. Therefore, when the model is intended for predicting relapse cases, it may be missing a huge number of positive relapse cases.

9 For predicting relapse, this random forest classifier is not recommended because of low sensitivity value. It may miss the patients with high relapse potential.