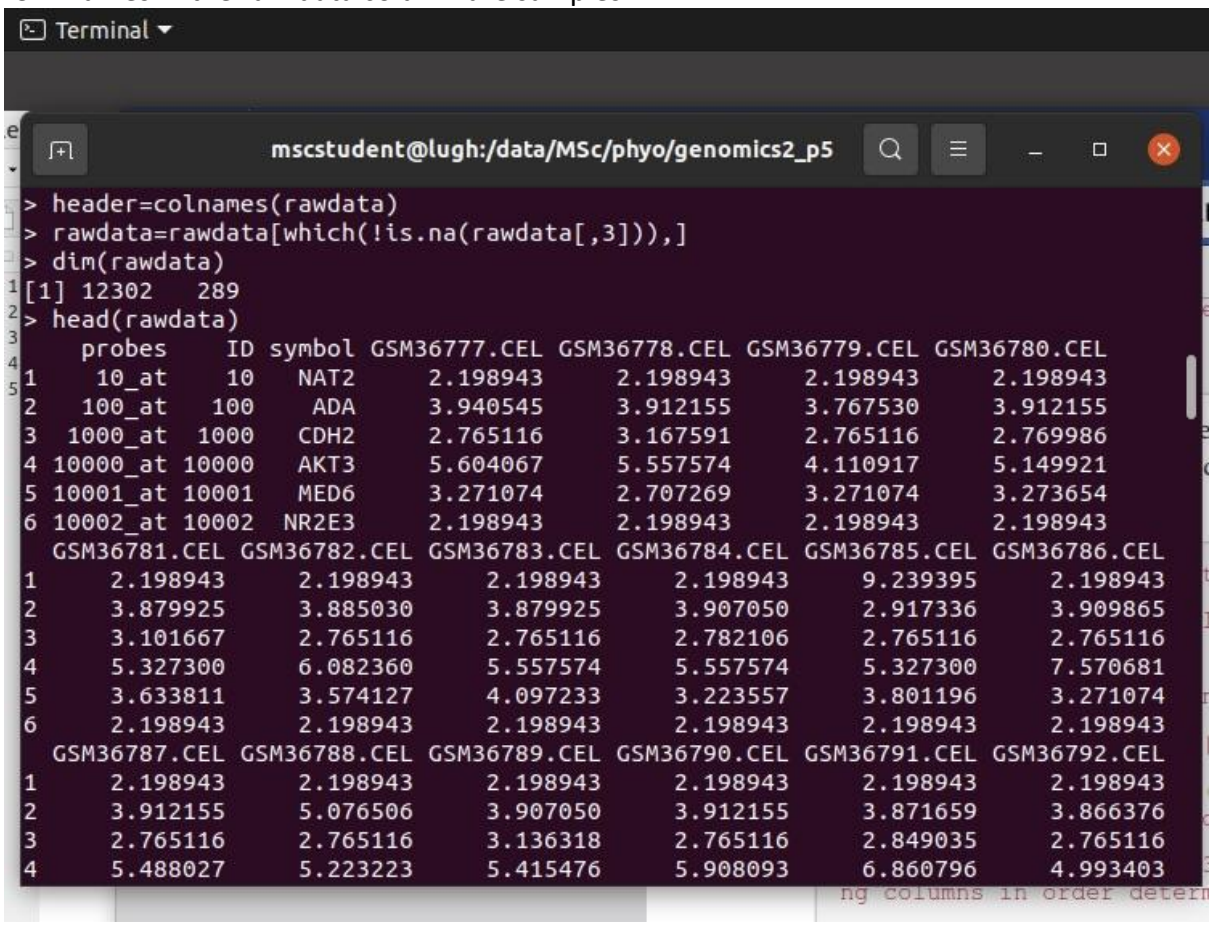
Aung Myat Phyo 22221184
Practical 5

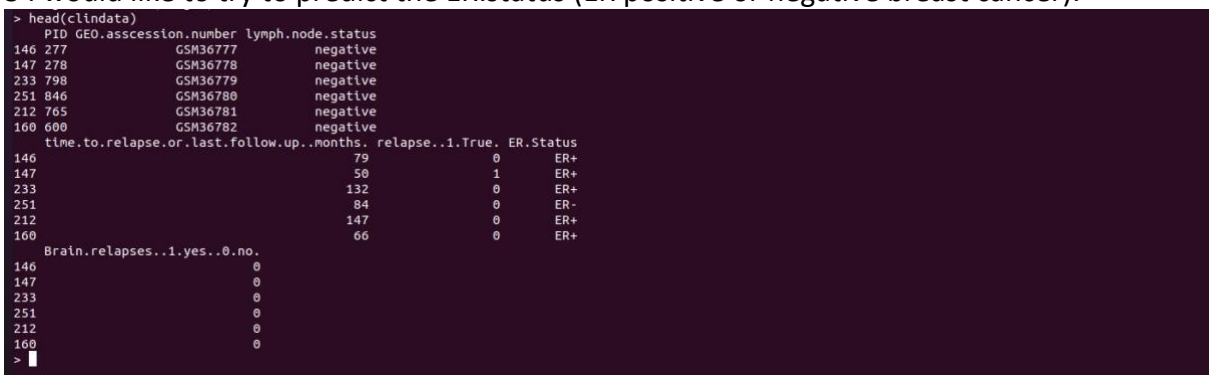
1

.CEL names in the raw data column are samples.

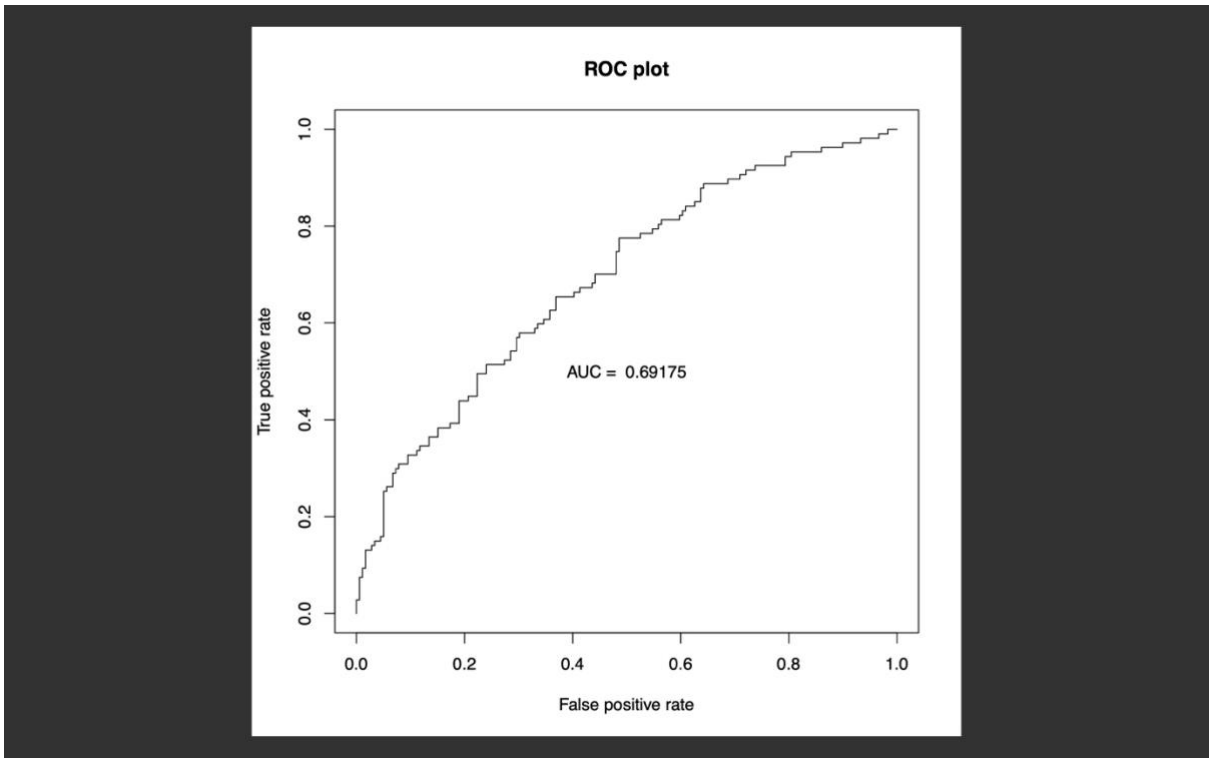


2 A CEL file contains gene expression raw data from microarray procedures. It is binary file format. The CEL file format is widely used in genomics field. For identifying differentially expressed genes, the CEL file data are preprocessed before next analysis (normalization and statistical testing).

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



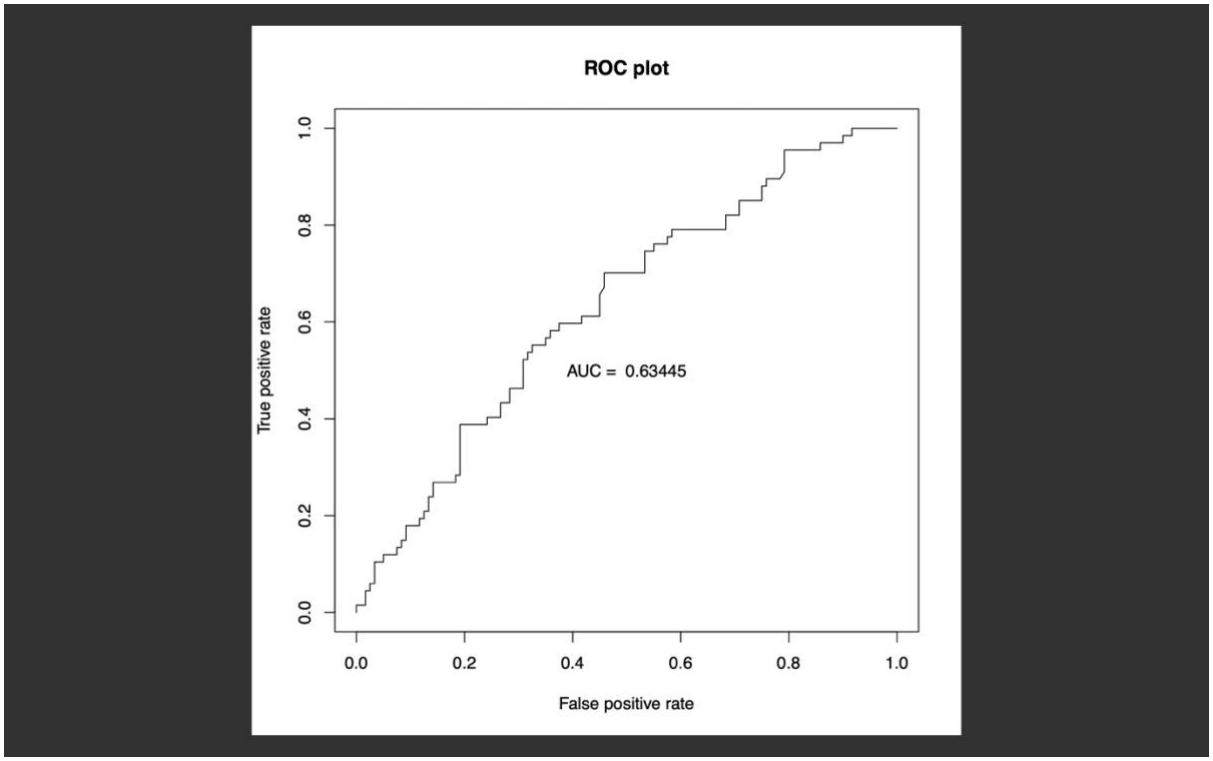
4 AUC value is 0.69175, it is good or acceptable AUC. In general, AUC value above 0.5 is desirable because it describes that model is better than random chance.



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

```
                PTBP3    KCNE2    DGCR2 CASP8AP2     SCO2
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 Sensitiviy value is 0.2835.

```
          1
> 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 replase 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.