Análise da performance dos classificadores

Análise de correlação dos genes

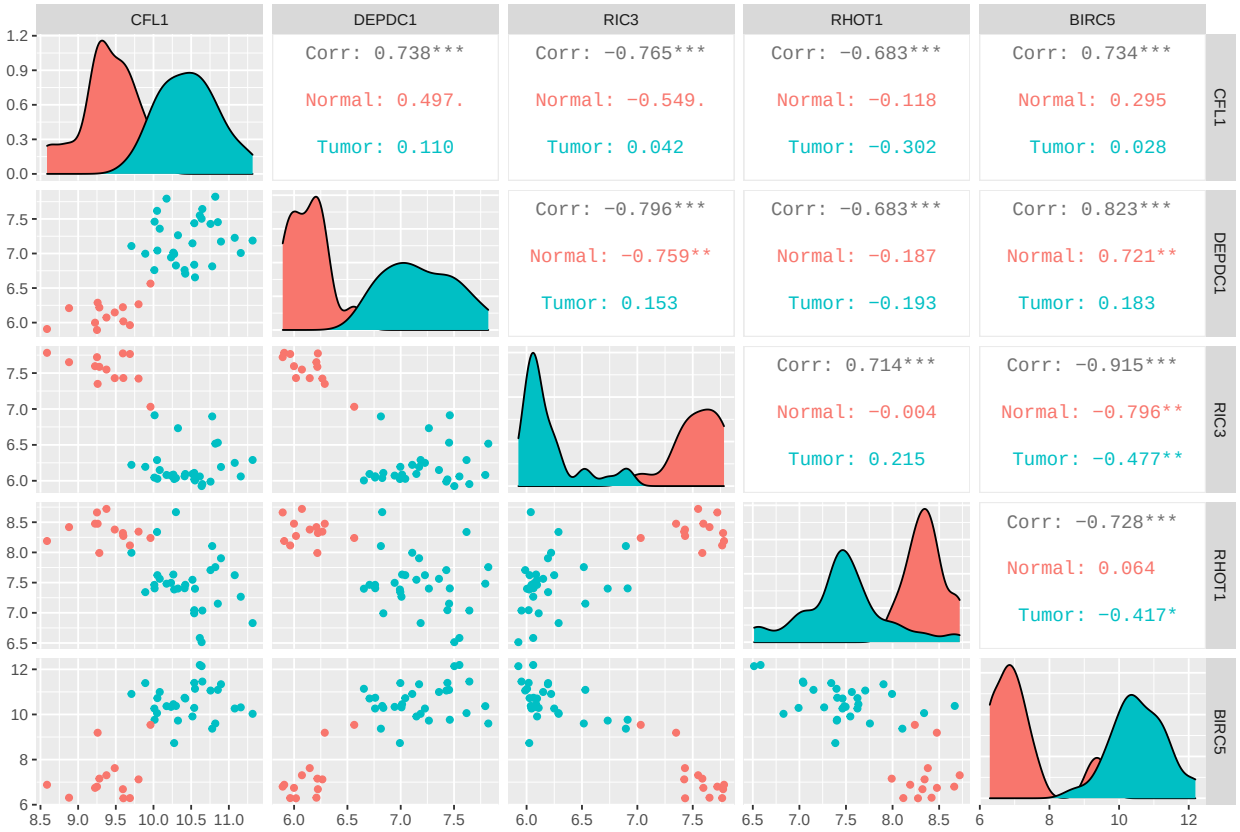
Correlação entre os 5 genes mais bem ranqueados de acordo com os seguintes parâmetros:

- **50 bags**
- **Information gain** as feature selector
- **Mean** as aggregation method

Os genes mais bem ranqueados, de fato, parecem separar bem os dados entre as duas classes com exceção do GSE71053. A questão é: esse conjunto de genes é o mais representativo em relação aos dados?

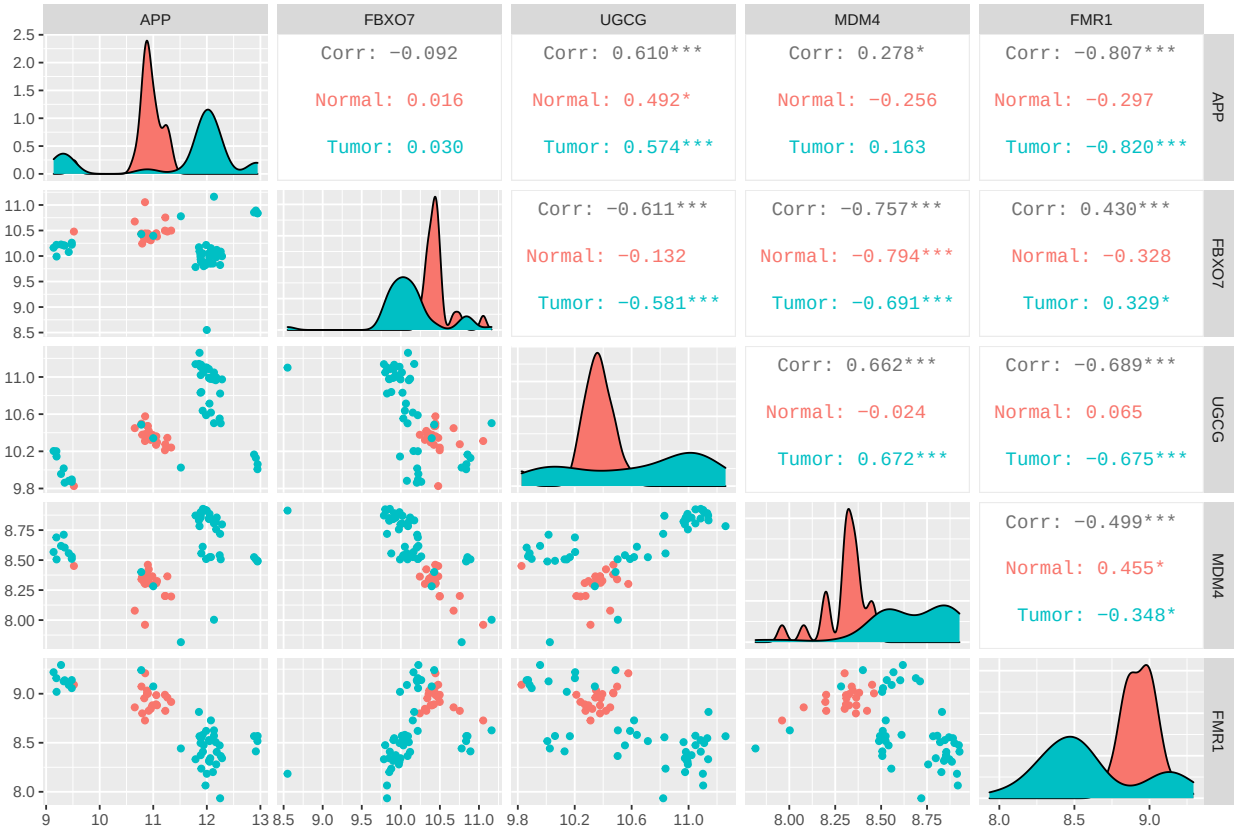
GSE10797, SVM, acc=1, recall=1

GSE38959, SVM, acc=1, recall=1

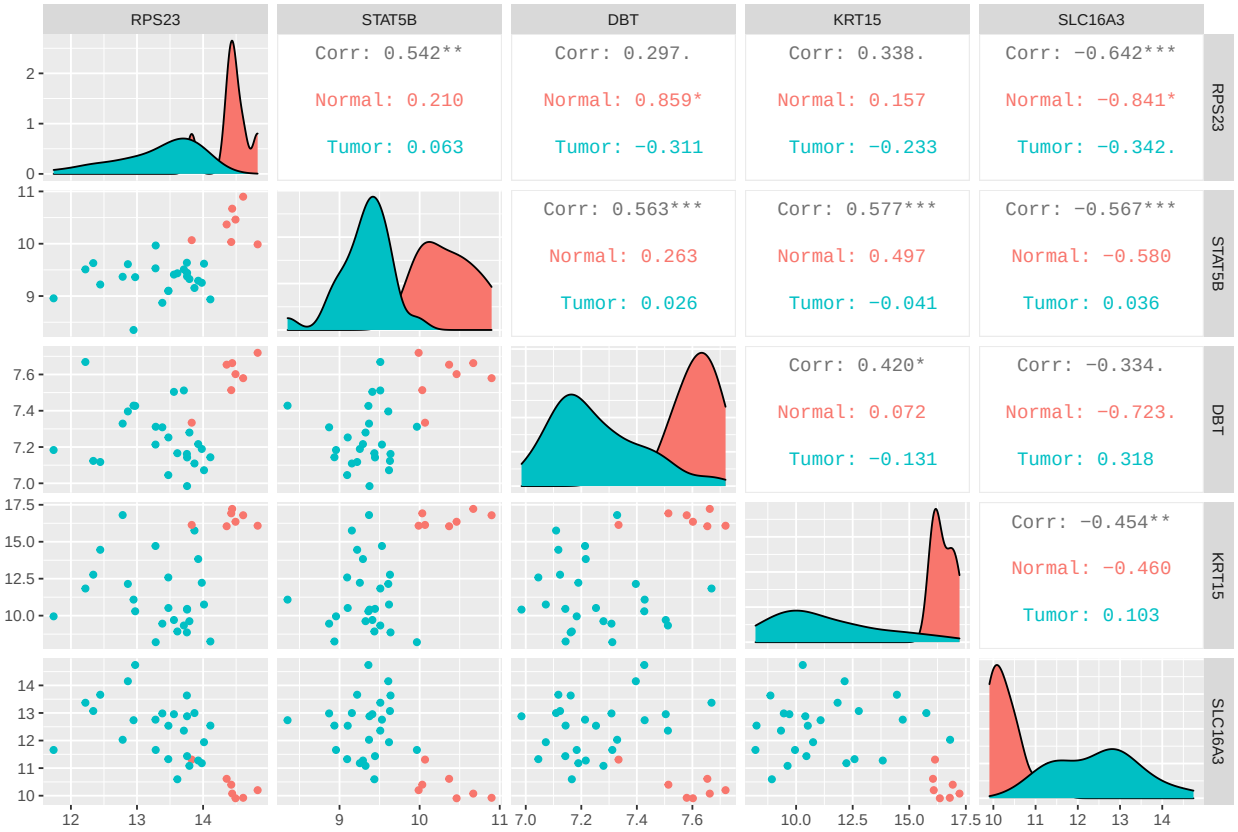


GSE45827, SVM, acc=0.987, recall=0.972

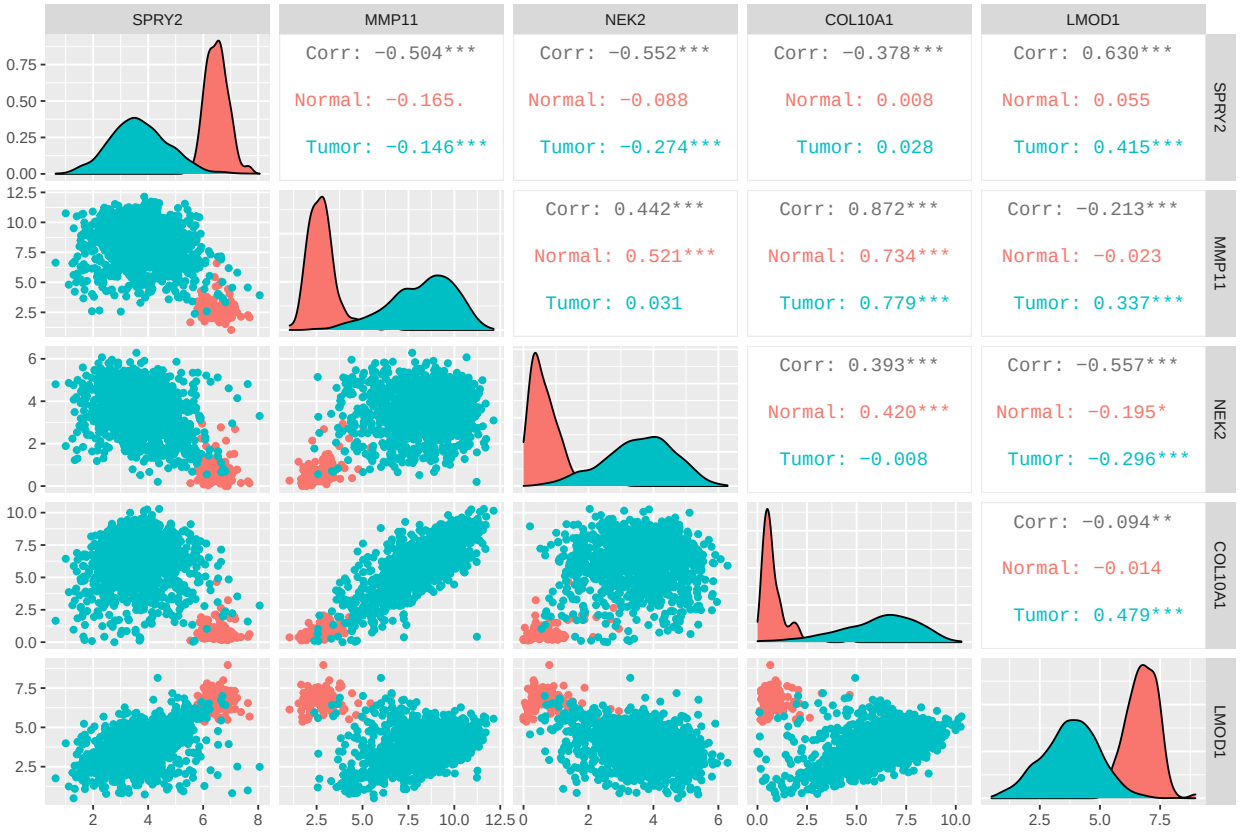
GSE53752, SVM, acc=0.912, recall=0.938



GSE57297, SVM, acc=1, recall=1

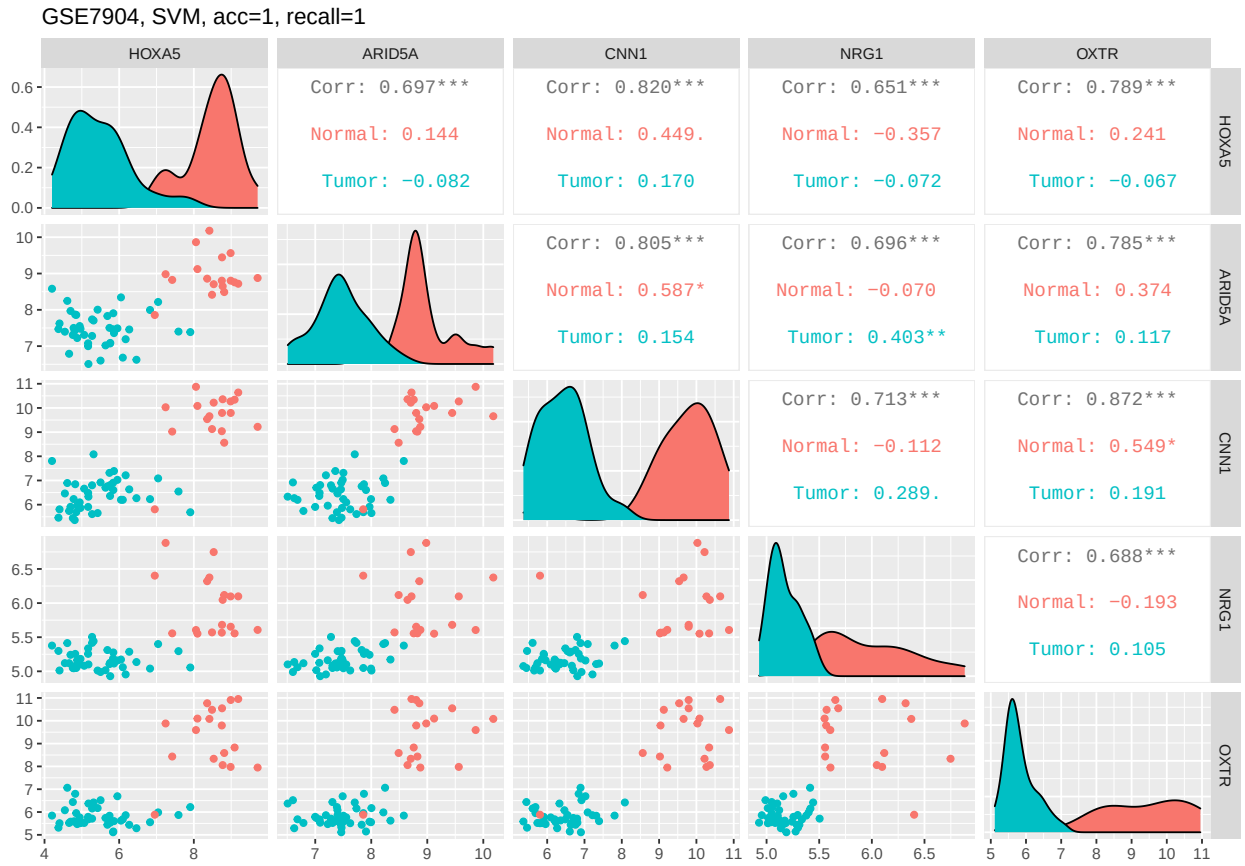


GSE62944, SVM, acc=0.996, recall=1



GSE70947, SVM, acc=0.873, recall=0.898

GSE71053, SVM, acc=0.875, recall=1



Análise da performance

Para responder a questão anterior, para cada dataset **5 genes** foram selecionados aleatoriamente e um modelo **SVM radial** foi treinado com os **dados de validação** (*dados originais + SMOTE*). Esse processo foi **repetido 100 vezes** para garantir a aleatoriedade dos genes. O resultado final é a **performance média** do modelo para cada dataset. Comparamos os resultados com a acurácia obtida pelos modelos treinados com os 5 genes selecionados utilizando **50 bags** e **information gain**.

```
##    DATASET    FS_ACC RANDOM_ACC    DIFF_ACC FS_RECALL RANDOM_RECALL
## 1 GSE10797 1.0000000  0.9485000 -0.05150000 1.0000000     0.9840000
## 2 GSE38959 1.0000000  0.8990909 -0.10090909 1.0000000     0.8810000
## 3 GSE45827 0.9866667  0.9644000 -0.02226667 0.9722222     0.9686111
## 4 GSE53752 0.9117647  0.8367647 -0.07500000 0.9375000     0.8231250
## 5 GSE57297 1.0000000  0.9685000 -0.03150000 1.0000000     1.0000000
## 6 GSE62944 0.9964789  0.9433685 -0.05311033 1.0000000     0.9759506
## 7 GSE70947 0.8728814  0.7783051 -0.09457627 0.8983051     0.8072881
## 8 GSE71053 0.8750000  0.8050000 -0.07000000 1.0000000     0.8600000
## 9  GSE7904 1.0000000  0.8886667 -0.11133333 1.0000000     0.9050000
##    DIFF_RECALL
## 1 -0.016000000
## 2 -0.119000000
## 3 -0.003611111
## 4 -0.114375000
## 5  0.000000000
```

```
## 6 -0.024049383
## 7 -0.091016949
## 8 -0.140000000
## 9 -0.095000000
```

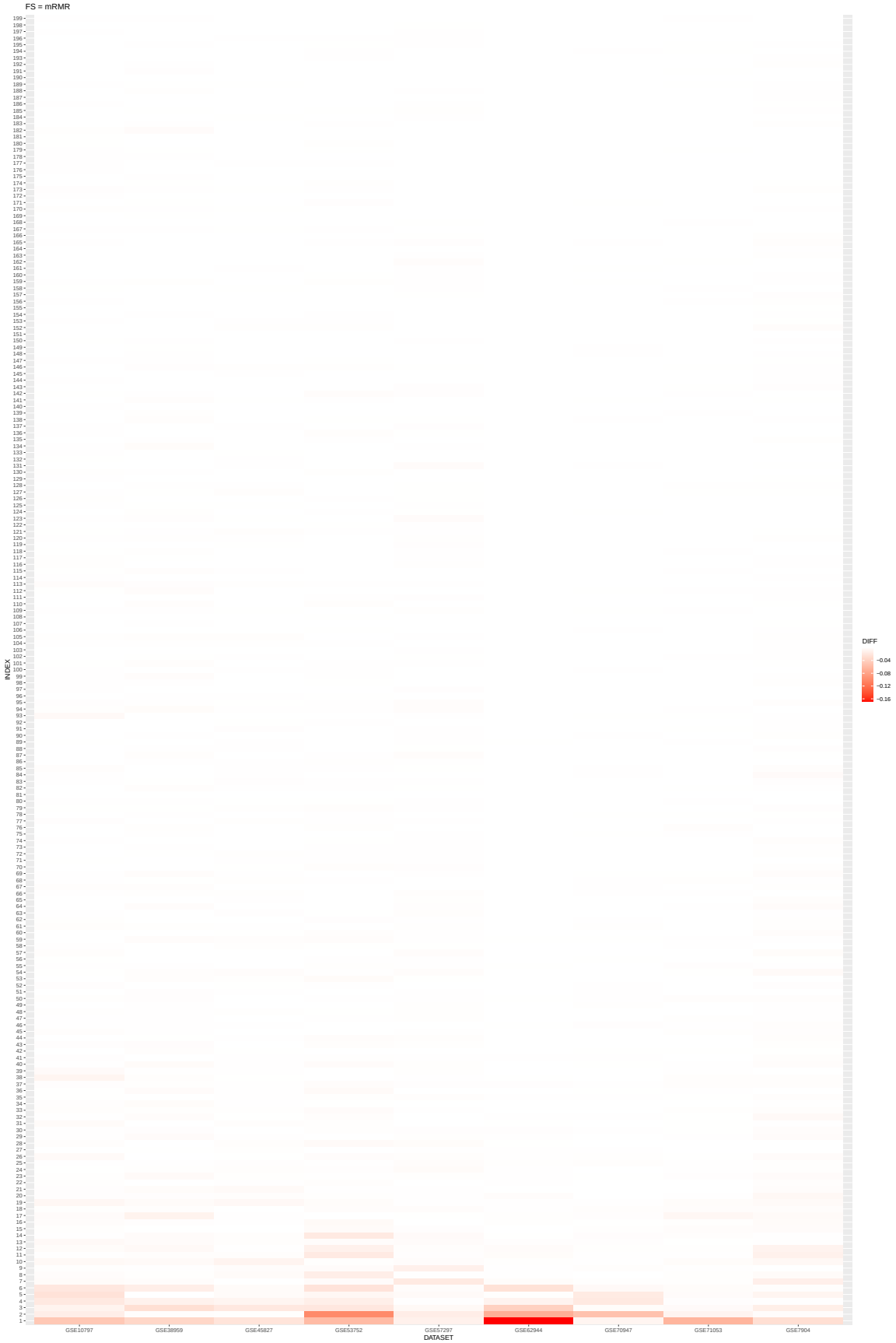
Estranhamente, a performance dos modelos treinados com genes aleatórios é satisfatória. Isso quer dizer que qualquer conjunto de 5 genes é representativo dos dados? Algumas performances mantiveram-se altas, como em GSE10797, outras cairam um pouco (GSE71053). A seguir, vamos analisar as diferenças dos *information gain* scores entre os genes de um ranking.

Análise dos rankings

Para cada dataset, carregamos o ranking final agrupado por média a partir de 50 rankings gerados por *information gain*. Calculamos a diferença de scores adjacentes no rankings da posição 1 até a posição 200. O gráfico abaixo apresenta essas diferenças como um *heatmap*.


FS = infoGain
INDEX
DATASET
GSE10797
GSE38959
GSE45827
GSE53752
GSE57297
GSE62944
GSE70947
GSE71053
GSE7904
DIFF
0.00
−0.02
−0.04
−0.06
−0.08

Os genes mais informativos – nas primeiras posições dos rankings – dos datasets GSE(10797, 57297, 71053, 7904) gerados pelo *information gain* e agregados por média não possuem diferença entre si. Isso significa que o information gain retornou o mesmo score para um grande conjunto de genes. Além disso, a diferença apresentada nos outros rankings não é significativa. A maior diferença identifica é **0.08**, o que explicaria o bom desempenho dos classificadores com genes aleatórios porque, sengundo o information gain, todos os genes são suficientemente informativos. Para investigar melhor os scores obtidos para cada gene, realizaremos a mesma análise para outros métodos de seleção de atributos.


FS = mRMR
INDEX
DATASET
GSE10797
GSE38959
GSE45827
GSE53752
GSE57297
GSE62944
GSE70947
GSE71053
GSE7904
DIFF
−0.04
−0.08
−0.12
−0.16

No caso

do mRMR, podemos perceber que as maiores diferenças concentram-se nas primeiras posições dos rankings. Isso significa algumas coisas:

1. Pelo mRMR, há genes que distinguem-se pela representatividade nos dados principalmente em GSE62944 e GSE53752;
2. O bom desempenho dos modelos treinados com genes aleatórios não pode ser explicado pelo argumento mencionado anteriormente;
3. As diferenças nas primeiras posições dos rankings são maiores comparadas com os resultados do info-Gain;
4. Por outro lado, as diferenças apresentadas pelo mRMR ainda são pequenas, o que pode corroborar com o bom desempenho mencionado no item 2.


FS = anova
INDEX
DATASET
GSE10797
GSE38959
GSE45827
GSE53752
GSE57297
DIFF
−1
−2

O método anova, assim como o mRMR, apresenta as maiores diferenças no topo do ranking. O que poderia indicar que, de fato, existem genes mais representativos dos dados do que outros. Próximos passos é entender melhor como cada método funciona, o que cada um leva em consideração para selecionar os genes. Isso vai nos ajudar a entender as diferenças nos resultados apresentados.