??CanAI (??? Cancer Prediction Machine Learning Project)

George Karakatsoulis, Nikos Pechlivanis

Institute of Applied Biosciences, Centre for Research and Technology, Hellas

Contents

[Preprocessing of the input dataset 1](#_Toc172897446)

[Fast Unified Random Forests with randomForestSRC 4](#_Toc172897447)

[Death time prediction 4](#_Toc172897448)

[Experiment #1 5](#_Toc172897449)

[Experiment #2 5](#_Toc172897450)

[Experiment #3 6](#_Toc172897451)

# Preprocessing of the input dataset

1. Any numeric predictors that were stored as characters were converted to numeric.
2. \*\_pathogenic predictors labelled as “na” were converted to NA (Not Available)
3. \*\_pathogenic predictors labelled as “0” were converted to “NON-PATH”
4. Filtering out of near zero variance predictors:

**Table 1**: Zero Variance Predictors

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| predictor | freqRatio | percentUnique | zeroVar | nzv |
| p16\_score | 31.66666667 | 4.964539007 | FALSE | TRUE |
| ACVR1 | 60 | 1.418439716 | FALSE | TRUE |
| AKT1 | 29.5 | 1.418439716 | FALSE | TRUE |
| ARHGAP26 | 60 | 1.418439716 | FALSE | TRUE |
| ARID2 | 19.33333333 | 1.418439716 | FALSE | TRUE |
| ARID4B | 19.33333333 | 1.418439716 | FALSE | TRUE |
| ATG2A | 29.5 | 1.418439716 | FALSE | TRUE |
| ATG7 | 19.33333333 | 1.418439716 | FALSE | TRUE |
| ATRX | 29.5 | 1.418439716 | FALSE | TRUE |
| BRAF | 29.5 | 1.418439716 | FALSE | TRUE |
| CASP8 | 60 | 1.418439716 | FALSE | TRUE |
| CD247 | 60 | 1.418439716 | FALSE | TRUE |
| CDK4 | 19.33333333 | 1.418439716 | FALSE | TRUE |
| CRTC3 | 29.5 | 1.418439716 | FALSE | TRUE |
| DNMT3A | 19.33333333 | 1.418439716 | FALSE | TRUE |
| EBF1 | 60 | 1.418439716 | FALSE | TRUE |
| EZH2 | 19.33333333 | 1.418439716 | FALSE | TRUE |
| H3F3B | 60 | 1.418439716 | FALSE | TRUE |
| HOXA13 | 29.5 | 1.418439716 | FALSE | TRUE |
| KDM4A | 29.5 | 1.418439716 | FALSE | TRUE |
| KDR | 60 | 1.418439716 | FALSE | TRUE |
| LMO2 | 19.33333333 | 1.418439716 | FALSE | TRUE |
| MET | 19.33333333 | 1.418439716 | FALSE | TRUE |
| MSH3 | 19.33333333 | 1.418439716 | FALSE | TRUE |
| NCOR2 | 29.5 | 1.418439716 | FALSE | TRUE |
| NFKB2 | 60 | 1.418439716 | FALSE | TRUE |
| NOTCH1 | 29.5 | 1.418439716 | FALSE | TRUE |
| NOTCH2 | 29.5 | 1.418439716 | FALSE | TRUE |
| NRAS | 19.33333333 | 1.418439716 | FALSE | TRUE |
| PAX3 | 19.33333333 | 1.418439716 | FALSE | TRUE |
| PHOX2B | 19.33333333 | 1.418439716 | FALSE | TRUE |
| PTPRC | 19.33333333 | 1.418439716 | FALSE | TRUE |
| RAD21 | 29.5 | 1.418439716 | FALSE | TRUE |
| RAD50 | 19.33333333 | 1.418439716 | FALSE | TRUE |
| RSPO2 | 19.33333333 | 1.418439716 | FALSE | TRUE |
| SFPQ | 19.33333333 | 1.418439716 | FALSE | TRUE |
| SLC34A2 | 60 | 1.418439716 | FALSE | TRUE |
| SMARCC1 | 60 | 1.418439716 | FALSE | TRUE |
| SRGAP3 | 29.5 | 1.418439716 | FALSE | TRUE |
| TET3 | 19.33333333 | 1.418439716 | FALSE | TRUE |
| WHSC1 | 60 | 1.418439716 | FALSE | TRUE |
| ABL1\_pathogenic | 23.5 | 1.418439716 | FALSE | TRUE |
| ACVR1\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| AKT1\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| ARHGAP26\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| ARID1A\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| ARID2\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| ARID4B\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| ATG13\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| ATG2A\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| ATG7\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| ATRX\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| AXIN1\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| BAP1\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| BRAF\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| BRD4\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| CASP8\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| CD247\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| CD274\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| CD8A\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| CD8B\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| CDK4\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| CHD6\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| CHEK2\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| CRTC3\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| CTLA4\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| DNMT3A\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| EBF1\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| EP300\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| ERBB2\_pathogenic | 23.5 | 1.418439716 | FALSE | TRUE |
| ERBB3\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| EZH2\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| FGFR2\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| H3F3B\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| HOXA13\_pathogenic | 23.5 | 1.418439716 | FALSE | TRUE |
| JAK1\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| JAK2\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| KDM4A\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| KDR\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| KIT\_pathogenic | 23.5 | 1.418439716 | FALSE | TRUE |
| KMT2C\_pathogenic | 23.5 | 1.418439716 | FALSE | TRUE |
| KMT2D\_pathogenic | 23.5 | 1.418439716 | FALSE | TRUE |
| LMO2\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| MAP2K2\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| MET\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| MLH1\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| MSH2\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| MSH3\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| MSH6\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| NCOR2\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| NFKB2\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| NIN\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| NOTCH1\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| NOTCH2\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| NOTCH3\_pathogenic | 23.5 | 1.418439716 | FALSE | TRUE |
| NRAS\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| PALB2\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| PAX3\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| PDCD1\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| PHOX2B\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| PTCH1\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| PTPRC\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| RAD21\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| RAD50\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| RANBP2\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| RSPO2\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| SETD2\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| SFPQ\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| SLC34A2\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| SMARCC1\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| SRCAP\_pathogenic | 23.5 | 1.418439716 | FALSE | TRUE |
| SRGAP3\_pathogenic | 48 | 1.418439716 | FALSE | TRUE |
| TET1\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| TET2\_pathogenic | 23.5 | 1.418439716 | FALSE | TRUE |
| TET3\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| TSC1\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| TSHZ3\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| WHSC1\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |
| ZNF521\_pathogenic | 0 | 0.709219858 | TRUE | TRUE |

# Fast Unified Random Forests with randomForestSRC

Survival analysis for both prediction of death and progress time was computed with *randomForestSRC* R package and the *rfsrc* function

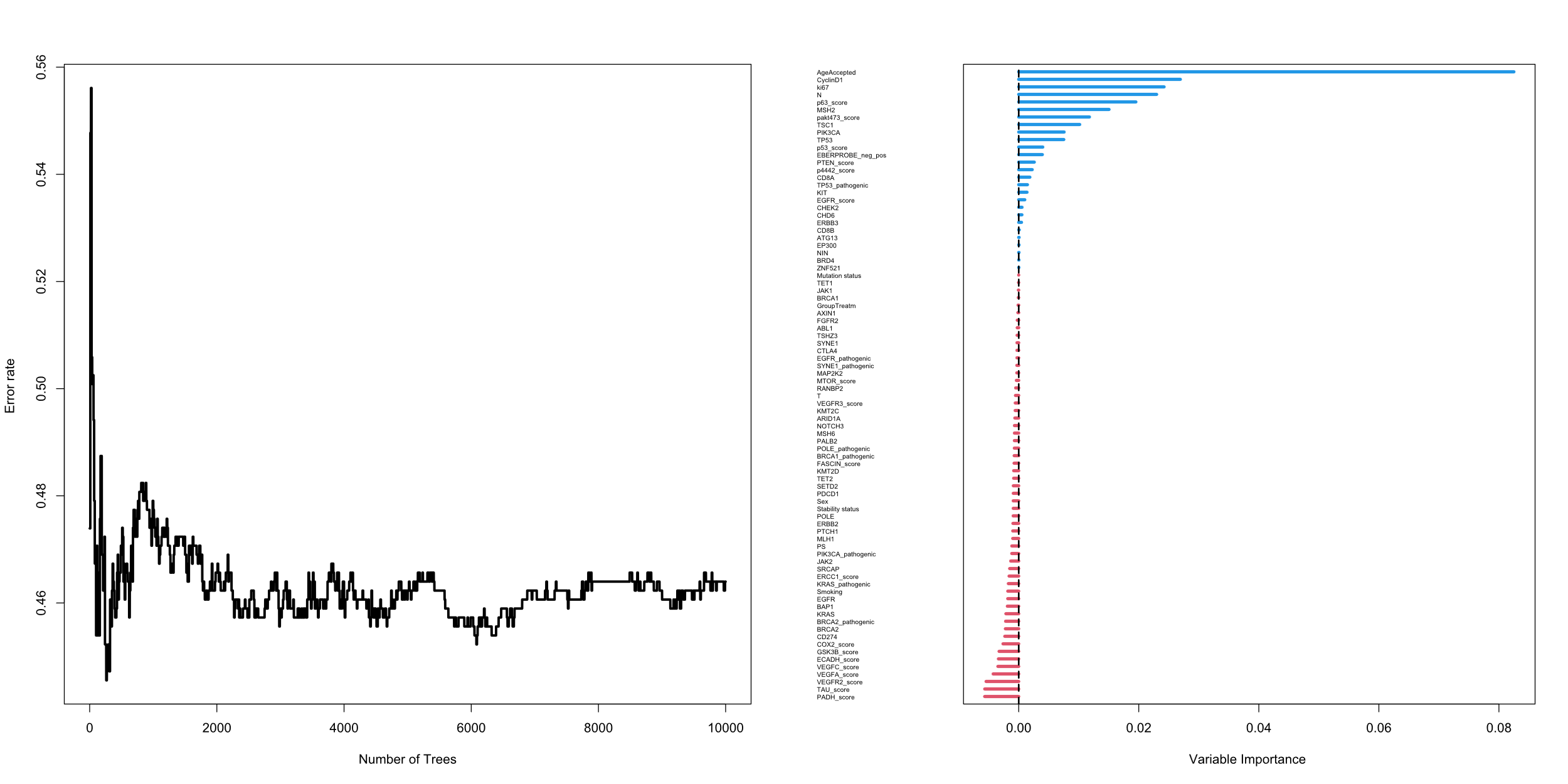
## Death time prediction

Analysis was conducted in four experiments:

* experiment #1: considering all predictors (as produced by previous preprocessing steps)
* experiment #2: considering only clinical (*"GroupTreatm", "AgeAccepted", "PS", "EBERPROBE\_neg\_pos",Smoking", "Sex", "T", "N", "CyclinD1", "ki67", "Mutation status", "Stability status"*) and \*\_score predictors (*"TAU\_score", "ERCC1\_score", "p63\_score", "ECADH\_score", "PADH\_score", "pakt473\_score", "MTOR\_score", "VEGFR2\_score", "VEGFA\_score", "VEGFR3\_score", "VEGFC\_score", "PTEN\_score", "COX2\_score", "FASCIN\_score", "p53\_score", "EGFR\_score", "GSK3B\_score", "p4442\_score"*)
* experiment #3: considering only gene NGS predictors (*"ABL1", "ARID1A", "ATG13", "AXIN1", "BAP1", "BRCA1", "BRCA2", "BRD4", "CD274", "CD8A", "CD8B", "CHD6", "CHEK2", "CTLA4", "EGFR", "EP300", "ERBB2", "ERBB3", "FGFR2", "JAK1", "JAK2", "KIT", "KMT2C", "KMT2D", "KRAS", "MAP2K2", "MLH1", "MSH2", "MSH6", "NIN", "NOTCH3", "PALB2", "PDCD1", "PIK3CA", "POLE", "PTCH1", "RANBP2", "SETD2", "SRCAP", "SYNE1", "TET1", "TET2", "TP53", "TSC1", "TSHZ3", "ZNF521"*)
* experiment #4: consideting only \*\_pathogenic predictors *("BRCA1\_pathogenic", "BRCA2\_pathogenic", "EGFR\_pathogenic", "KRAS\_pathogenic", "PIK3CA\_pathogenic", "POLE\_pathogenic", "SYNE1\_pathogenic", "TP53\_pathogenic"*)

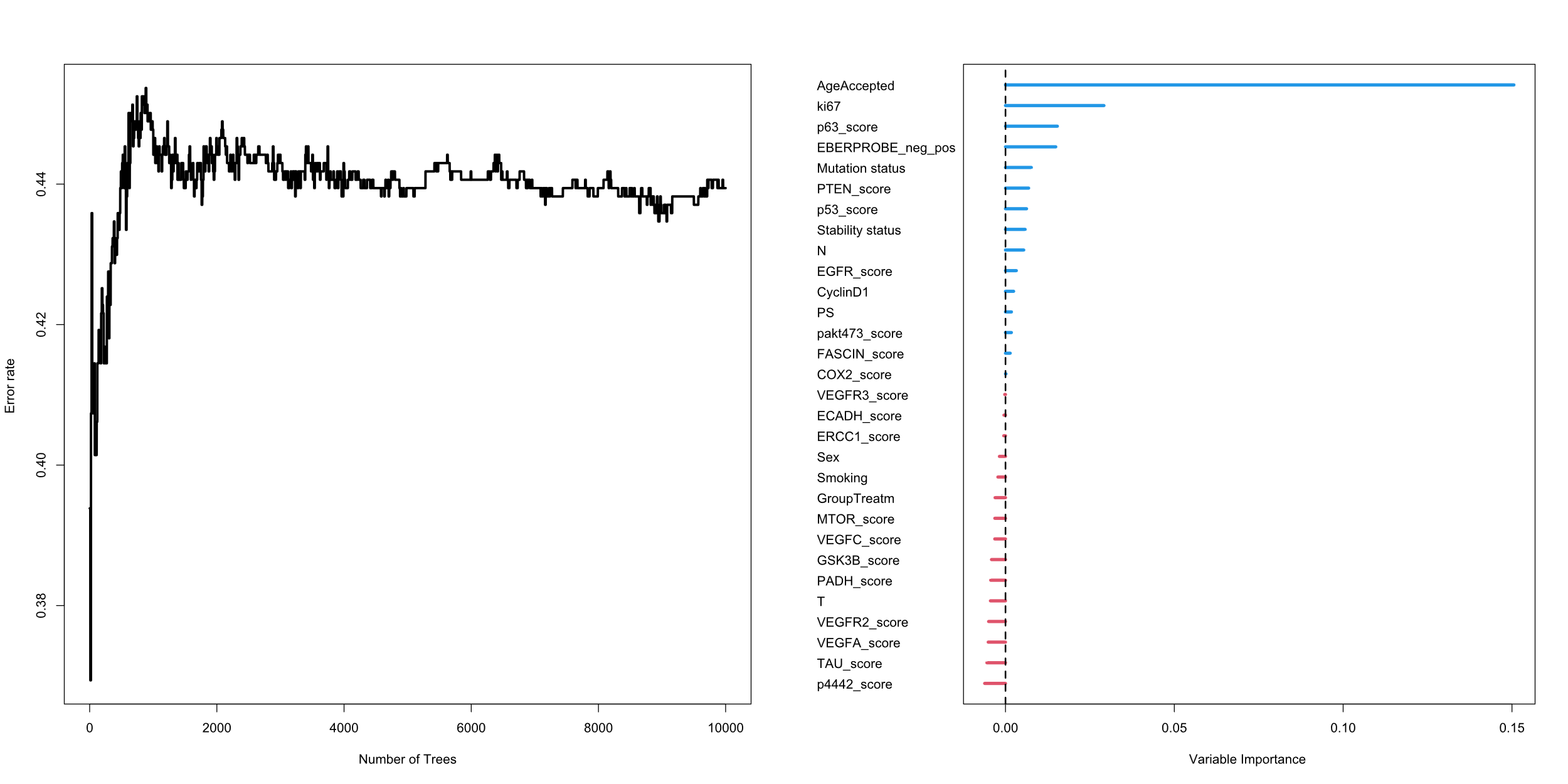
### Experiment #1

* Performance error: 0.4639866
* C-index: 0.23275618
* Most important variables (Importance > .01): *"AgeAccepted", "CyclinD1", "ki67", "N", "p63\_score", "MSH2", "pakt473\_score" "TSC1"*



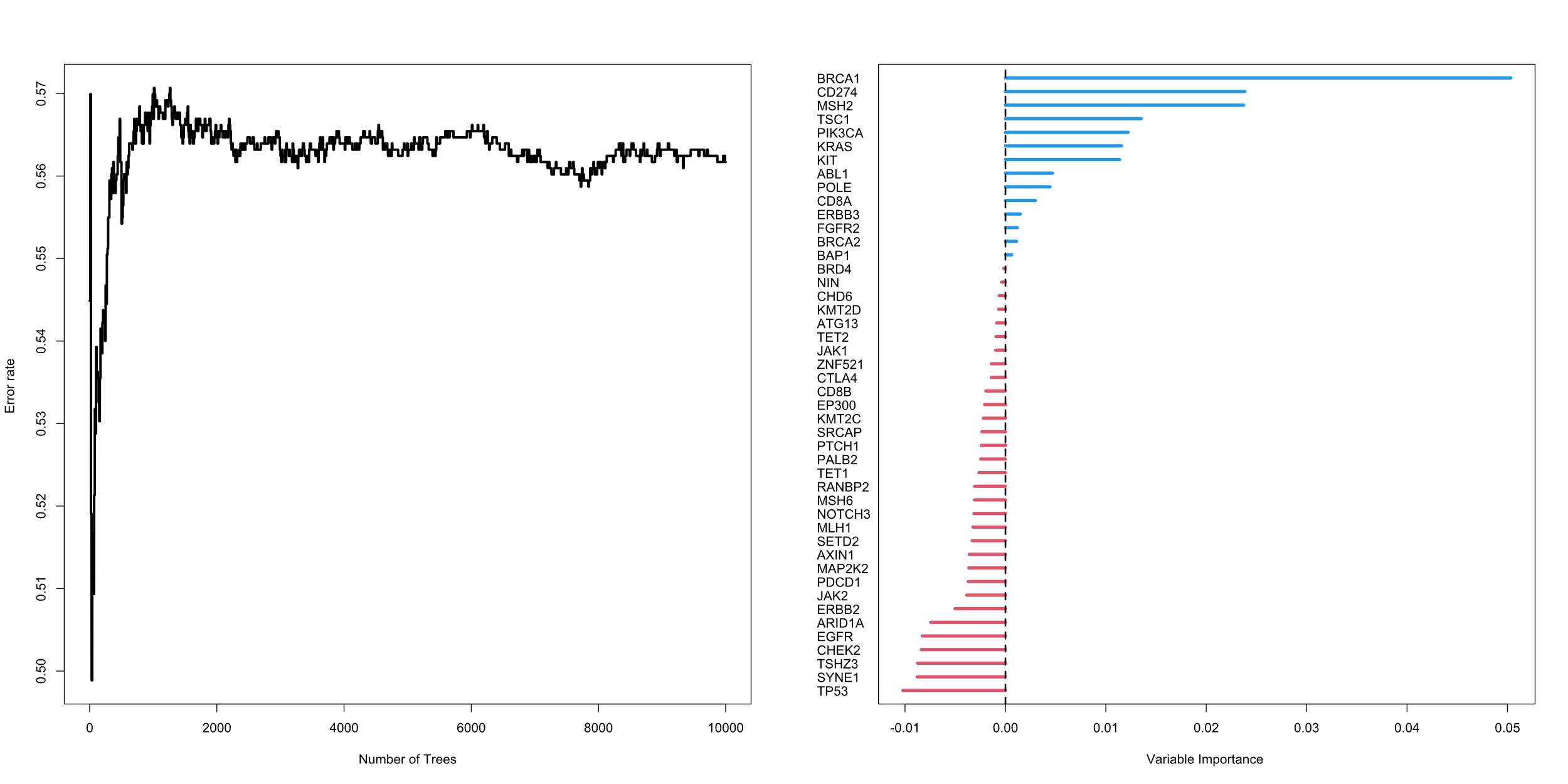
### Experiment #2

* Performance error: 0.43942993
* C-index: 0.21979704
* Most important variables (Importance > .01): *"AgeAccepted", "ki67", "p63\_score", "EBERPROBE\_neg\_pos"*



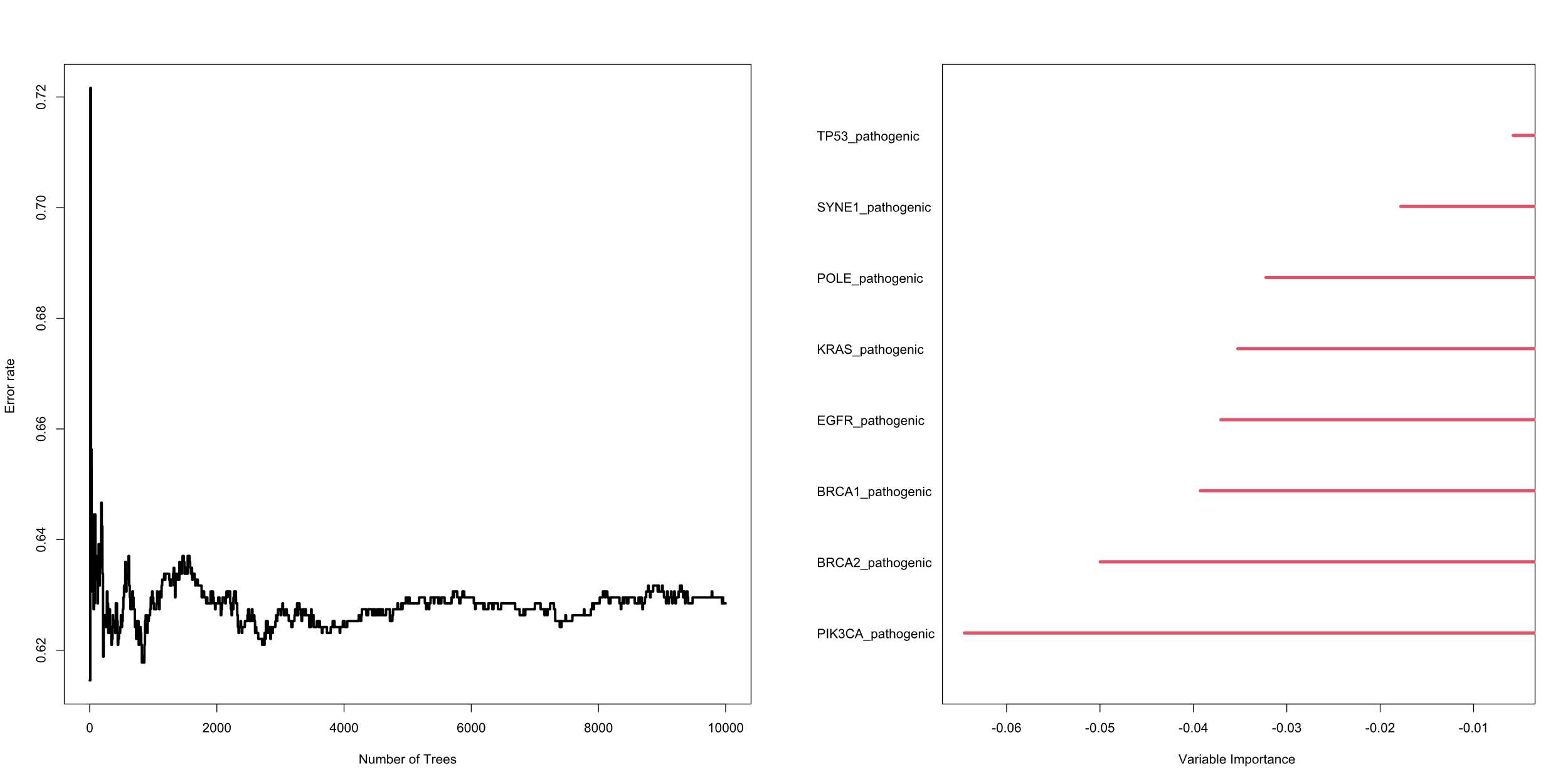
### Experiment #3

* Performance error: 0.56170531
* C-index: 0.24158321
* Most important variables (Importance > .01): *"BRCA1", "CD274", "MSH2", "TSC1", "PIK3CA" "KRAS", "KIT"*



### Experiment #4

* Performance error: 0.62847966
* C-index: 0.2481601
* Most important variables (Importance > .01): *Does not return any important variables*



## Progression time prediction

Analysis was conducted in four experiments as above.

### Experiment #1

* Performance error: 0.50558659
* C-index: 0.251254
* Most important variables (Importance > .01): *"CyclinD1" "MSH2" "AgeAccepted" "pakt473\_score" "p63\_score" "ki67" "EGFR\_score" "p53\_score"*