

SML_project

Mohammadreza shahriyarkeshe & Alireza Vashegani farahani

2026-02-15

Part 1. Data Exploration and Feature Selections

Load neccesary libraries

```
# install.packages(c("randomForest", "e1071", "class", "tidyverse", "ggplot2", "caret", "glmnet", "rpart"))
set.seed(123)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.5.2
## Warning: package 'tibble' was built under R version 4.5.2
## Warning: package 'tidyr' was built under R version 4.5.2
## Warning: package 'readr' was built under R version 4.5.2
## Warning: package 'purrr' was built under R version 4.5.2
## Warning: package 'dplyr' was built under R version 4.5.2
## Warning: package 'stringr' was built under R version 4.5.2
## Warning: package 'forcats' was built under R version 4.5.2
## Warning: package 'lubridate' was built under R version 4.5.2

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.6
## v forcats    1.0.1      v stringr    1.6.0
## v ggplot2    4.0.0      v tibble     3.3.0
## v lubridate  1.9.4      v tidyr      1.3.1
## v purrr      1.2.0
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(ggplot2)
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.5.2
## Loading required package: lattice
## Warning: package 'lattice' was built under R version 4.5.2
##
## Attaching package: 'caret'
##
```

```

## The following object is masked from 'package:purrr':
##
## lift
library(glmnet)

## Warning: package 'glmnet' was built under R version 4.5.2
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
## expand, pack, unpack
##
## Loaded glmnet 4.1-10
library(ggplot2)
library(class)

## Warning: package 'class' was built under R version 4.5.2
library(pROC)

## Warning: package 'pROC' was built under R version 4.5.2
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
##
## cov, smooth, var
library(randomForest)

## Warning: package 'randomForest' was built under R version 4.5.2
## randomForest 4.7-1.2
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:dplyr':
##
## combine
##
## The following object is masked from 'package:ggplot2':
##
## margin
library(e1071)

## Warning: package 'e1071' was built under R version 4.5.2
##
## Attaching package: 'e1071'
##
## The following object is masked from 'package:ggplot2':

```

```
##
##      element
```

Load the dataset

```
df <- read.csv("METABRIC_RNA_Mutation.csv")
head(df)
```

```
##      patient_id age_at_diagnosis type_of_breast_surgery  cancer_type
## 1             0           75.65      MASTECTOMY Breast Cancer
## 2             2           43.19  BREAST CONSERVING Breast Cancer
## 3             5           48.87      MASTECTOMY Breast Cancer
## 4             6           47.68      MASTECTOMY Breast Cancer
## 5             8           76.97      MASTECTOMY Breast Cancer
## 6            10           78.77      MASTECTOMY Breast Cancer
##      cancer_type_detailed cellularity chemotherapy
## 1      Breast Invasive Ductal Carcinoma              0
## 2      Breast Invasive Ductal Carcinoma          High      0
## 3      Breast Invasive Ductal Carcinoma          High      1
## 4 Breast Mixed Ductal and Lobular Carcinoma      Moderate      1
## 5 Breast Mixed Ductal and Lobular Carcinoma          High      1
## 6      Breast Invasive Ductal Carcinoma      Moderate      0
##      pam50_._claudin.low_subtype cohort er_status_measured_by_ihc er_status
## 1              claudin-low          1          Positive Positive
## 2              LumA              1          Positive Positive
## 3              LumB              1          Positive Positive
## 4              LumB              1          Positive Positive
## 5              LumB              1          Positive Positive
## 6              LumB              1          Positive Positive
##      neoplasm_histologic_grade her2_status_measured_by_snp6 her2_status
## 1              3              NEUTRAL      Negative
## 2              3              NEUTRAL      Negative
## 3              2              NEUTRAL      Negative
## 4              2              NEUTRAL      Negative
## 5              3              NEUTRAL      Negative
## 6              3              NEUTRAL      Negative
##      tumor_other_histologic_subtype hormone_therapy inferred menopausal_state
## 1              Ductal/NST              1              Post
## 2              Ductal/NST              1              Pre
## 3              Ductal/NST              1              Pre
## 4              Mixed              1              Pre
## 5              Mixed              1              Post
## 6              Ductal/NST              1              Post
##      integrative_cluster primary_tumor_laterality lymph_nodes_examined_positive
## 1              4ER+              Right              10
## 2              4ER+              Right              0
## 3              3              Right              1
## 4              9              Right              3
## 5              9              Right              8
## 6              7              Left              0
##      mutation_count nottingham_prognostic_index oncotree_code
## 1              NA              6.044              IDC
## 2              2              4.020              IDC
## 3              2              4.030              IDC
```

## 4	1	4.050	MDLC
## 5	2	6.080	MDLC
## 6	4	4.062	IDC
##	overall_survival_months	overall_survival	pr_status radio_therapy
## 1	140.50000	1	Negative 1
## 2	84.63333	1	Positive 1
## 3	163.70000	0	Positive 0
## 4	164.93333	1	Positive 1
## 5	41.36667	0	Positive 1
## 6	7.80000	0	Positive 1
##	X3.gene_classifier_subtype	tumor_size	tumor_stage death_from_cancer brca1
## 1	ER-/HER2-	22	2 Living -1.3990
## 2	ER+/HER2- High Prolif	10	1 Living -1.3800
## 3		15	2 Died of Disease 0.0670
## 4		25	2 Living 0.6744
## 5	ER+/HER2- High Prolif	40	2 Died of Disease 1.2932
## 6	ER+/HER2- High Prolif	31	4 Died of Disease -0.4341
##	brca2	palb2	pten tp53 atm cdh1 chek2 nbn nf1
## 1	-0.5738	-1.6217	1.4524 0.3504 1.1517 0.0348 0.1266 -0.8361 -0.8578
## 2	0.2777	-1.2154	0.5296 -0.0136 -0.2659 1.3594 0.7961 0.5419 -2.6059
## 3	-0.8426	0.2114	-0.3326 0.5141 -0.0803 1.1398 0.4187 -0.4030 -1.1305
## 4	-0.5428	-1.6592	0.6369 1.6708 -0.8880 1.2491 -1.1889 -0.4174 -0.6165
## 5	-0.9039	-0.7219	0.2168 0.3484 0.3897 0.9131 0.9356 0.7675 -0.2940
## 6	0.6931	0.6909	1.0840 -1.9371 0.9252 1.1520 0.7951 -0.9677 -0.7750
##	stk11	bard1	mlh1 msh2 msh6 pms2 epcam rad51c rad51d
## 1	-0.4294	-1.1201	-0.4844 -0.7483 -1.6660 -0.1250 -0.3721 -0.6508 -0.1278
## 2	0.5120	0.4390	1.2266 0.7612 0.1821 1.0104 0.5600 -0.4018 -0.2909
## 3	0.2362	-0.1721	-1.7910 3.0955 0.6608 2.6127 2.5553 -0.0391 -0.4421
## 4	1.0078	-0.4010	-1.3905 4.8798 0.0615 2.9414 4.1161 -0.3098 -1.3470
## 5	-0.2961	0.6320	-0.3582 0.3032 0.8747 0.6323 0.3349 -0.2652 -0.1541
## 6	-0.3532	1.0599	-0.2754 0.5496 1.1321 -0.4786 0.3220 0.2841 -0.3295
##	rad50	rb1	rb1l1 rb1l2 ccna1 ccnb1 cdk1 ccne1 cdk2
## 1	1.7330	-0.2770	-1.0673 0.0615 0.1034 -1.6635 -2.0649 -1.3388 -0.6208
## 2	0.7440	-1.7488	0.4602 0.7835 -0.2690 -0.2715 0.3895 -0.6273 0.1813
## 3	1.4528	1.4142	0.3143 -0.0662 -0.5558 0.6967 1.2949 -0.1245 -0.1109
## 4	1.2102	1.5345	-0.6347 0.7279 0.0219 1.8801 0.0476 -0.6651 -1.4457
## 5	-0.7631	1.0158	1.8996 0.6618 -0.7454 0.2862 1.1035 -1.0514 1.5423
## 6	-0.0611	0.8727	0.8561 0.5396 -0.0412 1.5735 1.5887 -0.4022 0.9184
##	cdc25a	ccnd1	cdk4 cdk6 ccnd2 cdkn2a cdkn2b myc cdkn1a
## 1	-1.7309	-0.8059	-1.1037 0.0744 1.4313 -0.5785 1.2180 2.5602 2.3152
## 2	-1.2643	1.1057	0.4108 -0.6040 0.0546 0.2754 -0.0952 0.7248 0.4389
## 3	-0.3923	-1.2747	-0.5312 0.2585 0.8958 -0.2270 1.1899 -0.7207 1.4829
## 4	-0.3827	-0.4537	-1.5248 0.3689 1.1455 0.0432 1.0868 -0.5237 0.2119
## 5	1.2169	0.1169	1.7648 0.8892 -1.0195 0.3664 0.7145 0.5223 1.5638
## 6	0.4784	-0.8358	1.1628 -0.2269 -0.7485 -0.4693 -0.1114 -0.8320 -0.8937
##	cdkn1b	e2f1	e2f2 e2f3 e2f4 e2f5 e2f6 e2f7 e2f8
## 1	-0.2487	0.1601	-1.8799 -1.3629 0.1992 -0.0907 -0.1167 -1.3917 -0.8178
## 2	1.8759	-1.8984	-0.3327 -1.4560 1.0011 0.0519 -0.3793 -0.4915 -1.3693
## 3	3.0696	2.2248	-1.4859 -0.4127 -0.8333 0.2885 -0.3496 -0.3549 0.0908
## 4	3.4617	0.4610	-1.5404 -0.2119 -0.8333 1.8323 0.3869 1.3434 0.0409
## 5	0.1562	-0.2666	0.4509 -0.4449 -0.3181 2.2129 1.1063 -0.0650 1.4216
## 6	0.3760	1.4738	1.0299 0.2396 -1.0892 -0.6725 0.1387 0.5118 0.1876
##	src	jak1	jak2 stat1 stat2 stat3 stat5a stat5b mdm2
## 1	-1.0737	1.1097	1.6186 0.0413 -0.7556 0.9765 3.9189 2.5773 0.3126

```

## 2 -0.3297 0.9804 -0.0104 -1.1940 -0.7991 -0.8428 -0.4983 -1.3251 0.8803
## 3 -0.7181 1.5835 0.7379 -0.8346 -1.3487 -0.2243 0.0434 -1.0832 -0.8822
## 4 -0.3804 0.6194 0.8552 0.3317 -0.5512 -1.2401 -0.8535 -0.0195 0.2551
## 5 0.9720 0.0461 0.1645 0.2677 0.3308 -0.5049 -0.3295 -0.4278 -0.0063
## 6 -0.5544 0.7215 0.0223 -0.7983 -0.0821 -1.2582 -0.3493 -1.4171 -0.3743
## tp53bp1 adam10 adam17 aph1a aph1b arrdc1 cir1 ctbp1 ctbp2
## 1 -0.6271 -0.5319 0.8721 0.0020 -0.0098 -1.4652 -0.2752 -1.1021 0.8414
## 2 -1.5924 -1.2672 -0.2006 1.3135 0.6880 1.3366 1.5911 1.1926 0.3010
## 3 0.7590 2.0364 -1.1229 3.1331 -0.4959 -1.0596 1.6744 -2.2695 0.2397
## 4 -1.0064 1.4659 -1.0490 2.3645 -1.2458 1.4919 2.6438 -0.7268 -0.4622
## 5 0.1063 0.9378 -0.2383 0.1228 0.3690 0.9222 0.2017 -0.8767 1.8501
## 6 -1.0313 1.4207 -0.6227 1.0733 -0.3213 0.3046 -0.4826 -2.2878 0.3841
## cul1 dll1 dll3 dll4 dtx1 dtx2 dtx3 dtx4 ep300
## 1 0.4469 0.6307 -0.5067 1.6931 1.7100 -1.8994 0.6028 0.3299 1.5090
## 2 -0.4113 -1.5476 -0.1581 -0.1005 -0.9584 -0.1775 -0.7991 -0.3891 -0.4294
## 3 0.1616 -1.0439 0.9493 -1.7455 -0.8876 0.2312 -1.2906 1.1141 -1.5670
## 4 0.1933 0.4481 1.6109 -0.5916 -0.4730 -0.2765 -1.1845 1.1098 -1.7947
## 5 -0.6839 -1.2165 -0.8212 0.2763 0.1702 -0.7525 -2.4018 0.3444 -0.5688
## 6 -1.1976 -0.8885 -0.7506 1.2604 -1.2258 -0.2350 -1.2827 0.7424 -0.7230
## fbxw7 hdac1 hdac2 hes1 hes5 hey1 itch jag1 jag2
## 1 0.3038 -1.6822 0.1622 0.8322 0.9870 -1.9028 -0.4093 -0.0858 0.2445
## 2 0.6385 1.6267 -0.9130 0.2582 0.0066 -0.5588 0.7234 0.2604 -0.6473
## 3 -1.6157 -1.8103 0.8620 1.1837 -0.1830 -0.4287 0.0269 0.9807 -1.2545
## 4 -1.3703 -0.3412 1.0671 -0.1616 -0.5077 -1.8552 0.4120 -1.4933 -0.8333
## 5 -0.5238 0.4816 -1.1730 -1.2850 -0.3400 -1.9064 1.1496 -1.1192 -1.0050
## 6 0.2296 0.9504 0.6525 -0.3005 0.0483 0.4208 1.4359 -0.2027 -1.3310
## kdm5a lfng mam11 mam12 mam13 ncor2 ncstn notch1 notch2
## 1 -0.9667 -0.8598 -0.1030 3.2415 -1.5365 0.4684 -2.0007 -0.3395 -0.2029
## 2 0.7122 -0.1773 -0.1606 -1.4582 -0.1651 -0.0066 -0.9120 -1.4125 0.3260
## 3 -1.6218 -1.2785 -1.9686 1.4406 -0.6574 -1.9701 -1.4975 -1.0102 -0.9168
## 4 -2.2726 1.1498 -1.5455 -0.2630 -0.8255 -0.9941 -2.5452 -1.5746 -1.1520
## 5 -0.0353 -1.4281 -2.7196 -0.0878 -2.1710 -2.1310 -0.7509 -0.0075 0.6149
## 6 -0.0372 1.0007 -0.1913 0.6408 0.2829 -1.0606 -0.7447 -0.3868 1.7182
## notch3 nrarp numb numbl psen1 psen2 psenen rbpj rbpjl
## 1 -0.8725 -0.6873 -0.4592 -0.3334 0.4056 -1.0686 -0.3936 1.1433 -0.0714
## 2 -1.4752 1.0167 -0.4651 0.8040 0.6743 -0.6104 0.6839 -0.4825 1.4183
## 3 0.3654 -0.6829 2.3143 -0.0641 1.3108 -1.0017 2.7082 0.0557 -1.6488
## 4 -0.5382 -0.8165 2.3152 1.2117 1.1638 -1.5603 0.7708 1.1797 -0.6543
## 5 -1.4361 -1.6582 -0.7700 -0.3525 2.1816 -0.4056 0.7125 -0.4021 0.3117
## 6 -0.2060 -0.1295 1.2164 -0.8778 1.9355 -0.7217 0.5640 0.2732 -0.2292
## rfng snw1 spen hes2 hes4 hes7 hey1 hey2 acvr1
## 1 0.2433 0.0513 0.2550 -0.3139 0.0938 -1.3190 -0.2490 2.0761 0.4710
## 2 -0.0538 2.1712 1.2776 -0.0013 1.2396 -2.1458 0.3879 -0.4597 0.8386
## 3 -0.0451 3.8545 -2.5468 -0.7339 -0.1023 0.6490 1.4153 2.0196 2.2031
## 4 1.4984 4.1430 -2.8894 1.3690 -0.4797 0.3536 -0.2092 0.4415 1.0455
## 5 0.0706 -0.5673 -0.6666 -0.5381 0.4906 -0.1500 -0.4913 -0.8568 -0.7534
## 6 -0.1417 0.6209 -1.1095 -0.3023 -1.0802 -0.0626 -0.0632 -0.7204 0.6530
## acvr1b acvr1c acvr2a acvr2b acvr1l akt1 akt1s1 akt2 apaf1
## 1 -0.9114 1.8079 -0.9273 -0.7363 2.9004 -1.1280 -0.4574 0.2391 -1.8513
## 2 0.6737 0.7279 1.0905 0.5358 -1.8416 0.0312 -0.1056 0.6519 -0.0742
## 3 -1.7817 0.6777 1.5452 -0.5908 -1.2039 0.9269 0.1185 1.2371 0.1704
## 4 -1.2848 -0.4277 2.9200 0.5155 -1.7602 1.2375 -1.0408 -0.1938 0.2536
## 5 0.6626 0.2038 0.4134 -0.0238 -0.8343 -0.9541 -0.8911 2.1021 -0.6142
## 6 2.0786 0.0892 0.5916 -0.3711 -0.2839 0.8486 0.2940 -1.0773 -0.8040

```

##	arl11	atr	aurka	bad	bcl2	bcl2l1	bmp10	bmp15	bmp2
## 1	-0.0124	-0.6790	-1.5733	-0.3367	0.5371	-0.1719	0.0308	-0.8320	4.7741
## 2	-0.5111	0.5127	-0.3052	1.8906	0.3243	-0.5759	2.6618	3.3397	-1.1385
## 3	0.0771	0.0012	-0.2248	1.3090	-1.0989	0.3435	0.3679	1.5678	-0.7861
## 4	-0.5253	-0.5074	2.0447	-0.1371	0.2252	0.7383	0.2514	0.7842	-0.6000
## 5	0.0657	-0.3322	0.5882	0.5063	-1.2477	1.2839	-0.6159	-0.1525	-0.5262
## 6	-1.8359	-0.0956	0.9681	0.6301	-1.0655	-0.3556	-0.7085	-0.9338	0.1218
##	bmp3	bmp4	bmp5	bmp6	bmp7	bmpria	bmprib	bmpr2	braf
## 1	1.8812	0.7940	0.2512	2.5915	-0.3928	1.1934	-0.8505	0.9883	-0.6387
## 2	-0.3217	0.7861	-0.1874	-0.3346	0.2365	-0.3413	0.5569	0.1759	0.5886
## 3	0.2980	2.0460	1.3173	-0.3617	-0.4294	0.4681	0.4623	1.7894	0.7275
## 4	0.9754	-0.3231	-0.0546	-0.1192	4.1957	1.0486	-0.4867	2.3596	0.2123
## 5	0.6704	0.1790	-0.0792	0.5098	2.4848	0.8654	-1.0298	-0.7718	-0.2628
## 6	0.0182	0.9278	-0.3303	-0.2204	-0.4885	-0.0481	-0.3995	0.7021	1.3157
##	casp10	casp3	casp6	casp7	casp8	casp9	chek1	csf1	csfir
## 1	0.3033	-2.1257	-0.8434	0.6977	0.1816	-0.4332	-1.6332	-0.0424	0.4810
## 2	0.1039	0.5775	1.7652	-0.1495	1.8819	-0.4484	-0.9604	0.2664	-2.1839
## 3	-0.8205	-0.9291	1.7003	1.0673	0.6432	-1.2527	0.7585	-0.6927	-0.5916
## 4	1.2658	0.5207	0.0513	1.5423	-1.1012	0.6762	-0.8414	0.2440	-0.7753
## 5	0.1754	-0.5207	-0.5943	0.9725	1.8931	-0.3756	0.4420	-0.0896	-0.3344
## 6	-0.3328	0.7524	0.4298	0.6657	0.1365	0.7313	0.5122	1.4540	-0.0239
##	cxcl8	cxcr1	cxcr2	dab2	diras3	dlec1	dph1	egfr	EIF4E
## 1	0.1226	0.9278	-0.0070	1.2766	0.2883	-0.4517	1.1179	2.0696	-0.2314
## 2	-1.1673	-1.2491	0.4389	-1.7476	0.6600	0.4479	1.6732	-0.3609	0.9246
## 3	1.2092	0.0089	-1.8753	-0.7936	-0.1572	-0.6188	0.2521	-0.0991	3.1838
## 4	0.0309	-0.0706	1.8137	0.1723	0.2875	-0.7166	1.4561	-0.7250	2.9343
## 5	-0.3125	-0.9502	0.7129	-0.3024	-0.5542	0.0629	1.3411	-0.5042	0.4861
## 6	1.3514	0.1979	-1.5995	0.7715	-0.6152	-0.8360	2.0360	-0.3820	1.0809
##	EIF4EBP1	EIF5A2	ERBB2	ERBB3	ERBB4	FAS	FGF1	FGFR1	FOLR1
## 1	-0.2265	0.5513	-1.0548	-1.7207	0.2415	-0.1829	2.2525	1.7000	1.6797
## 2	-0.3173	-0.5307	-0.7633	0.5803	0.0059	-0.4012	-0.8545	0.3062	-0.7914
## 3	-0.5358	0.2586	-0.7661	-1.3607	-0.4383	2.1470	0.9773	-0.4567	-0.4299
## 4	-0.2353	0.4649	-0.3172	-0.6405	-0.5122	1.8836	0.4925	-0.7820	-0.5803
## 5	1.2588	-1.0460	-0.5962	1.4865	-1.0922	0.5010	-1.0715	0.8701	0.5115
## 6	-0.8255	0.1909	-0.7556	1.8514	-0.2987	0.5731	1.1448	-0.3321	-0.7140
##	FOLR2	FOLR3	FOXO1	FOXO3	GDF11	GDF2	GSK3B	HIF1A	HLAG
## 1	1.0893	0.4606	2.5507	0.9213	-0.6499	-1.5386	-0.7982	0.4191	0.1048
## 2	-0.9169	-0.0141	-0.7180	-1.6144	0.0169	1.9384	-0.0094	-0.8038	-0.9355
## 3	0.5313	0.3959	-0.4464	-1.2823	-1.2771	0.5162	-1.7768	3.5336	1.9768
## 4	1.3413	0.6259	0.4124	-1.2511	-0.9396	-2.8998	-1.6895	3.3403	2.0772
## 5	0.6556	0.3455	-1.1731	-1.5713	-0.3794	0.5430	-0.3001	1.3211	0.0246
## 6	0.5147	0.5408	-1.1890	0.5518	-0.1169	-1.5779	0.6956	0.3223	-0.4904
##	HRAS	IGF1	IGF1R	INHA	INHBA	INHBC	ITGAV	ITGB3	IZUMO1R
## 1	-0.9923	0.9947	-0.4445	0.0883	1.1190	-0.4613	-0.4337	-0.2304	0.7302
## 2	0.9679	0.8191	0.0638	-0.2797	-0.6347	-0.2217	-0.7036	0.2577	-0.9707
## 3	-0.6058	1.5676	-0.6426	-0.4539	-0.3512	-1.4372	2.2121	0.6469	-1.1507
## 4	0.9831	2.8491	-0.7844	1.9837	0.6968	0.1198	1.3652	-0.8502	-0.8229
## 5	0.2881	-0.2109	-0.5785	-0.1702	-0.5076	0.3660	-1.2198	-0.3194	1.0331
## 6	-0.2939	0.3345	0.1098	-0.2980	-0.7263	0.0017	1.0856	-0.0188	0.7347
##	KDR	KIT	KRAS	MAP2K1	MAP2K2	MAP2K3	MAP2K4	MAP2K5	MAP3K1
## 1	2.7183	2.9336	-2.1663	0.4229	-0.0078	2.6548	-0.0279	-0.2537	-1.0723
## 2	-0.9864	-0.4866	-0.6278	-2.0640	-0.5408	-1.9500	-1.1421	-1.0739	1.3129
## 3	-0.4649	0.4321	0.5189	0.0169	-0.8582	-2.2248	0.6299	-2.8788	1.6581
## 4	-1.2923	-0.5447	0.5582	-0.2496	-1.4272	-0.0413	1.0026	-2.2113	1.0801

```

## 5 -1.7323 -0.1758 -0.6360 0.4079 -2.2363 -0.1867 -0.3795 -0.4151 -1.2170
## 6 0.8371 0.0673 0.2856 -0.2006 -1.3852 -0.5597 0.6858 -1.9100 0.5212
## map3k3 map3k4 map3k5 mapk1 mapk12 mapk14 mapk3 mapk4 mapk6
## 1 0.9728 -0.5303 1.1210 -0.7856 -0.4231 1.4091 0.9809 0.5448 -0.4627
## 2 -1.6056 -0.7564 0.4997 -0.4899 -0.3840 1.7598 -0.6298 0.2111 -0.9976
## 3 -2.2545 -0.6332 0.1242 0.3567 -0.4056 1.0982 -0.8799 -0.6478 1.8587
## 4 -1.3444 -0.1964 -0.3292 -0.3793 0.2475 -0.2556 1.3314 1.6936 0.2379
## 5 -0.2842 -2.3884 -0.3605 1.2317 0.3536 3.9901 0.5779 -1.0071 0.9238
## 6 -0.5480 -0.4851 1.2085 0.0673 -1.3031 1.1976 -0.7768 -0.0952 0.4993
## mapk7 mapk8 mapk9 mdc1 mlst8 mmp1 mmp10 mmp11 mmp12
## 1 0.3205 -2.4995 -0.8403 -0.6932 -1.0871 -0.8340 -0.7603 -3.2039 -0.5300
## 2 -0.6786 2.3882 0.5150 -1.1688 -0.1174 -0.4467 -0.7450 0.5243 -0.4845
## 3 -1.6654 0.1825 1.0738 -1.5809 -1.7957 2.2321 -0.2289 0.7216 -0.0249
## 4 -1.2308 -0.1847 0.2743 -0.8201 -1.2531 1.2883 -0.3216 0.1539 -0.2261
## 5 -1.9312 -0.4494 -0.1982 0.3260 0.9414 -0.1305 -0.6322 0.0268 -0.5826
## 6 -0.8710 0.9069 0.7342 -0.2527 2.6402 0.4307 -0.7828 0.6495 -0.4895
## mmp13 mmp14 mmp15 mmp16 mmp17 mmp19 mmp2 mmp21 mmp23b
## 1 -0.8768 -0.5866 -0.8050 1.1284 0.0194 1.7351 -0.1627 0.4474 0.7726
## 2 -0.7357 0.1100 -0.0552 0.9059 0.7236 -0.0589 -0.0365 1.1748 0.6258
## 3 1.9421 2.7550 0.0417 -0.5516 0.0935 0.2619 0.2149 0.0325 0.0978
## 4 2.8566 0.6631 1.7173 0.9451 0.4783 0.6355 -0.0788 -1.1357 0.1764
## 5 -0.2464 -1.7046 -0.2091 0.0024 -0.0867 0.1040 -0.6777 0.2603 1.9485
## 6 4.5023 0.3540 0.5598 -0.4017 -0.1714 0.0062 0.5728 0.7628 0.9354
## mmp24 mmp25 mmp26 mmp27 mmp28 mmp3 mmp7 mmp9 mtor
## 1 0.2549 -0.2840 0.7557 0.6495 3.9945 -1.3277 1.1844 -2.3771 -0.5469
## 2 1.1908 2.8068 0.0684 0.3887 0.2456 -1.1408 -1.5162 -1.3469 -0.3186
## 3 -0.1158 -1.2076 -0.0915 0.4895 0.3810 -0.1280 1.3559 -0.1297 -1.7472
## 4 0.2230 0.2569 -0.3969 -0.2373 0.8236 -0.4842 -0.1366 -0.7623 -1.9265
## 5 -0.9704 0.6494 -0.7281 0.6346 -0.3306 -0.5542 0.0062 -0.7551 -0.6149
## 6 -0.1423 -1.1207 0.2957 0.2306 0.2345 2.1229 -1.3101 0.8590 -1.8715
## nfkb1 nfkb2 opcm1 pdgfa pdgfb pdgfra pdgfrb pdpk1 peg3
## 1 -0.8838 -0.3837 -1.5136 1.0340 -0.0349 1.6642 0.1510 -0.2005 0.3824
## 2 -0.8477 -0.8394 0.8983 -0.3446 -0.3739 0.5016 -0.7094 0.2841 1.7849
## 3 -1.2457 -0.8461 0.2619 -1.3937 -1.6093 1.8002 0.0673 -0.4960 -0.9138
## 4 -2.0421 -0.7812 -0.4338 -0.4175 -0.4251 2.1803 -0.2670 -0.7175 -0.5943
## 5 -1.3880 -0.3511 -0.3835 -1.2038 -1.0203 -0.3866 -1.6346 1.4198 0.6252
## 6 0.4924 -0.5438 0.0600 -1.4420 0.0464 1.1543 0.4245 1.4094 -0.3538
## pik3ca pik3r1 pik3r2 plagl1 ptk2 rab25 rad51 raf1 rassf1
## 1 -0.7545 2.0932 -3.1626 2.5054 -1.5906 -1.6133 -0.7447 -0.4185 1.4167
## 2 -0.6025 -0.3072 0.2887 -0.4624 -0.7651 0.4303 -0.4151 -0.1043 1.5891
## 3 2.2172 1.7297 -0.9902 -0.6465 -1.0594 0.3867 -0.1498 -1.8437 1.2840
## 4 3.5440 -0.6674 -1.3752 -0.8595 2.6289 0.2140 0.8658 -0.6233 -0.2534
## 5 -0.4320 -1.1464 -0.2831 -0.3823 3.0912 -0.6780 1.8685 -1.2441 1.1045
## 6 0.4340 1.4169 0.6127 -0.4005 0.5229 0.3893 0.8858 -0.3513 0.8479
## rheb rictor rps6 rps6ka1 rps6ka2 rps6kb1 rps6kb2 raptor slc19a1
## 1 0.7390 -1.1872 0.8191 -1.7290 0.6942 -0.5645 -1.8694 0.0165 -0.0677
## 2 0.7496 -0.0622 1.4198 -0.9011 0.1972 -1.4849 2.3000 -1.3098 -1.1252
## 3 1.0751 1.5229 1.3971 -2.2621 -0.1481 -1.2594 -1.0724 -1.5115 -0.7975
## 4 0.7494 1.2288 1.6454 -2.5622 -0.7248 1.4816 -1.5166 0.4326 -0.4880
## 5 -0.8958 -0.4307 -1.3988 -0.5631 -2.0748 -0.6419 1.4917 -0.3344 2.3240
## 6 -0.2092 0.2306 0.4615 -0.1899 -0.9189 0.0843 0.7960 -0.3891 -0.5554
## smad1 smad2 smad3 smad4 smad5 smad6 smad7 smad9 sptbn1
## 1 -0.0464 0.7905 0.6308 0.2348 -0.3647 0.0523 1.2775 -0.7776 2.9551
## 2 0.9060 1.1109 -0.5117 0.8578 -0.5399 -0.2627 0.3427 0.1040 -1.1895

```

```

## 3  0.6871 -1.0038 -1.7931  1.4349  3.4869 -0.1044 -0.2511 -0.8081 -1.0472
## 4 -0.5883  0.4072 -0.2181  0.8580  7.3425 -1.3747 -0.6002 -1.8264 -0.9305
## 5  0.5721 -0.7546 -0.7921  0.1512 -0.7187  0.3484 -0.5915 -0.7889 -1.0334
## 6 -0.8846  0.5649 -0.2695 -0.1633  0.6546  1.5221  0.1917 -1.9749 -1.1020
##      terc      tert      tgfb1      tgfb2      tgfb3      tgfb1r1      tgfb1r2      tgfb1r3      tsc1
## 1  0.1389  2.0141 -0.6235  0.1020 -1.4828 -0.6998  2.2907  2.0023  0.0623
## 2  0.2867  1.0530  0.0613 -0.1426  0.4945 -1.6412 -0.6814 -1.5001  0.8054
## 3 -1.0066  1.0893  0.4069  2.5760  1.1768  0.8615  0.2776 -1.0694 -0.9652
## 4 -0.0247  0.3395  0.9068 -0.4965  1.0085  0.3283  1.0795 -0.8085 -0.6103
## 5 -0.0315 -0.9543 -1.0100  0.5378 -0.9729 -0.2201 -0.2670 -0.3679 -0.7420
## 6 -0.6549 -0.2850  1.3148  0.7296 -0.0217  0.9350  0.6334 -0.2339 -1.0450
##      tsc2      vegfa      vegfb      wfdc2      wwox      zfyve9      arid1a      arid1b      cbfb
## 1 -0.7012 -1.1656  1.4530 -0.0897 -0.4464  0.7291 -1.2999 -0.2649 -0.2706
## 2 -0.6919 -0.4946  1.6321  0.9657 -0.8286 -0.3056  1.9157 -0.9599  0.3507
## 3 -1.3444 -1.0330 -1.0909  0.8227 -0.3171  0.0015 -0.9847 -0.3794  1.1649
## 4 -0.9989 -1.1922 -0.9584 -0.5884 -0.4584 -1.5983 -1.5992 -0.8519  1.4578
## 5  0.4206  0.6806  2.0137 -0.6005  0.1638  1.3898  0.5634 -0.4830 -0.3686
## 6  1.7267  0.1574  0.7314  0.4464  1.6318 -1.0152 -0.5369  0.1070  0.3577
##      gata3      kmt2c      kmt2d      myh9      ncor1      pde4dip      ptpd      ros1      runx1
## 1 -1.7113 -0.9045  2.6008  0.2852  0.3529  1.3600 -0.7693 -1.0379 -1.9710
## 2  1.1638 -0.0208  0.9370 -1.8283  0.1916 -1.7667  0.5273  0.4946  0.6746
## 3 -0.1419 -0.5063  0.1020  0.4317 -1.7078 -1.6482  1.4359  0.2610 -1.0401
## 4 -0.5560 -1.2840  1.4290  0.7069 -1.2370 -2.0525  1.3493  1.5894 -0.8902
## 5  0.1444 -1.1026  1.4702 -1.7583 -0.2024 -0.2026 -0.7522  0.3038 -2.0765
## 6  0.1892  0.3166 -1.6740 -0.4642 -0.0344 -1.0594  1.6772  0.0028 -0.1756
##      tbx3      abcb1      abcb11      abcc1      abcc10      bbc3      bmf      cyp2c8      cyp3a4
## 1 -0.1164  1.6758 -0.4071 -1.0213 -0.1948 -0.2122 -2.3395  0.2242 -0.0572
## 2  1.5592 -1.1793 -0.0931  0.4261 -0.9788 -0.3817 -0.3431  0.4754  0.9464
## 3 -1.0888 -0.4457 -1.1208 -0.5168 -1.5536 -0.0325 -0.2389  1.3501 -1.3978
## 4 -1.6030 -0.8191 -0.7736 -1.4145 -0.9771  2.5351 -1.0042 -0.4249  1.1476
## 5 -0.8439 -0.3463  0.7826 -0.8794  0.0819 -0.9266 -0.0444 -0.5863 -0.1807
## 6 -0.6413 -1.1718 -0.2498 -0.8899 -1.0814 -0.4160 -0.8803  0.1493 -0.3404
##      fgf2      fn1      map2      map4      mapt      nr1i2      slcolb3      tubb1      tubb4a
## 1  2.4435 -1.1291 -0.2147 -0.3917 -0.1692  0.6645 -0.1634  0.7029 -0.0250
## 2 -1.1548  0.6923 -0.5615  0.9038  1.2644 -0.6135 -0.5180  0.8972 -0.1003
## 3  0.5059  0.3415 -0.8314 -0.6845 -0.0460  1.1309  0.6809  0.3157  1.2084
## 4  0.0401 -0.0785 -0.3800 -1.0618 -0.7920  2.5443 -0.2881  8.4721  0.3142
## 5 -0.0446  0.8398 -0.1281 -0.2336 -0.4751 -1.1240 -0.0012  1.0188 -0.6606
## 6 -0.5561  0.6853 -0.8649  0.5824  0.3931  0.1498 -0.7615  1.1460 -0.2737
##      tubb4b      twist1      adgra2      afdn      aff2      agmo      agtr2      ahnak      ahnak2
## 1 -0.4113  2.8096  2.8014 -0.0004  0.9673  0.3011 -0.8436  1.8227  0.3066
## 2  0.7791 -0.2273 -0.4462 -1.9854  0.5022 -0.9526 -1.8435  1.6662  1.4577
## 3 -0.6572  0.1984 -1.0721 -0.9729  0.0515  0.1109  0.9874 -0.0154  0.8589
## 4 -0.4413  0.1932 -1.0215  0.4553 -0.2354  0.4003  1.4839  0.3101 -0.6734
## 5 -1.4697  0.4128 -1.5326 -0.4795  1.0052  0.9739  0.8825 -0.7598  0.5698
## 6  0.3531  0.2687 -0.1279 -0.1615  0.4695 -1.5671 -1.1999  0.3160  0.6014
##      akap9      alk      apc      arid2      arid5b      asxl1      asxl2      bap1      bcas3
## 1 -0.7988  0.4355  0.5815 -1.4317  2.2128  0.6381 -1.4209 -0.9975 -0.6185
## 2  0.4849 -1.9242  1.3903  2.8106 -0.8652 -0.2488  0.7743  0.3675 -1.1052
## 3  1.7950  0.9890  1.5604  2.0526  0.8695 -1.8879 -2.2948 -2.6732 -0.5876
## 4  1.5143 -0.4072  1.6567  2.6831 -0.2603 -1.6951 -1.7671 -3.8521 -0.6948
## 5 -0.9309  0.7014  0.3155  1.0845 -0.9843 -0.7672 -1.0214  0.2162 -0.1630
## 6 -0.1376 -0.3880  1.9626  2.2988 -0.0683 -0.3980 -0.2695 -1.1370 -1.4179
##      birc6      cacna2d3      ccnd3      chd1      clk3      clrn2      col12a1      col22a1      col6a3

```



```

## 1 -1.0493  0.0055  0.2307  0.2223 -0.1376  0.2433 -0.6096 -0.3107 -0.3426
## 2  1.3807 -0.4037 -0.9088  0.0454 -0.7820 -0.8355  1.0075 -0.2229  0.2785
## 3 -0.3068  1.8081 -2.6859  0.6366 -1.6504 -0.6838  0.8384 -0.2396  0.9698
## 4  1.9645 -0.0030 -0.6588  2.0459 -1.4345  0.3367  0.7001 -0.4587  0.6020
## 5 -0.7862 -0.3866  1.1366 -0.2469 -0.0795  0.2905 -0.9057 -0.9781 -1.1143
## 6  0.1823 -0.2400  0.4169  0.6997  1.3162  0.6249  1.0053 -0.4554  0.9192
##      ctcf  cttna1  cttna3  dnah11  dnah2  dnah5  dtwd2  fam20c  fanca
## 1 -0.4098 -0.0482  0.2311 -0.0721 -1.0962 -0.9616 -1.2864  1.4018  0.7203
## 2  0.2358 -1.2060 -0.8015 -0.9145 -1.1824  1.9868 -1.3597 -0.4388 -0.8625
## 3 -0.8201  0.7313 -0.0799 -1.4797 -0.7075 -0.3000  0.8004 -0.4371 -0.0681
## 4 -0.6187  0.9967 -0.8569 -0.7878 -1.1014 -0.6013 -0.1297 -0.0893  0.6354
## 5 -0.1674  0.5049  0.2764 -0.7933 -1.1111 -1.1666 -0.1833 -1.7009 -0.5370
## 6 -0.8262  0.2651 -0.8396 -1.2011  0.2288 -0.1805  1.0086 -0.9360 -1.1738
##      fancd2  flt3  foxp1  frmd3  gh1  gldc  gpr32  gps2  hdac9
## 1 -2.0976 -0.5296 -0.3045  1.3213  1.0185 -0.2863  0.2566  0.7920  1.5660
## 2 -0.7109  1.7385 -0.1447 -0.4348  1.9129  3.9959 -0.0545 -1.7629 -0.1561
## 3 -0.6161  0.7884  2.0214 -1.0083 -0.2739 -0.7664 -1.0429 -0.7136 -0.0906
## 4 -0.2885 -0.1103  1.5297 -0.2226  0.5911  1.6744 -0.5610 -0.1025  0.3307
## 5 -0.5824 -1.1332  0.6269 -0.7415 -0.5916 -0.2990  2.7664 -0.7149  0.3153
## 6 -0.2206 -0.8089  1.6298 -0.7298  0.3492 -0.7283 -0.7031 -1.6201  0.2974
##      herc2  hist1h2bc  kdm3a  kdm6a  klrg1  l1cam  lama2  lamb3  large1
## 1  0.0066 -1.1763 -0.6926  0.3471  1.1316  0.2378  2.6466  1.8590 -0.4848
## 2 -1.7101  1.0178  1.3896 -0.2019 -1.1526 -0.4051  0.3336 -1.2531  0.7805
## 3 -2.9315  0.7477  2.6560  1.6417 -0.8215 -0.5621  0.1738  1.2534 -0.0103
## 4 -2.4937  0.0737  1.4758  0.5496 -0.8541  1.0439 -0.5435 -1.0418  0.4836
## 5 -1.2114  2.5406  0.5260 -0.3067  0.0272 -0.5308 -0.7579 -0.8603  0.6386
## 6 -1.4627  0.9576 -0.2366  0.7672 -0.9405  0.0090  0.0893 -0.6230  0.5194
##      ldrlap1  lifr  lipi  magea8  map3k10  map3k13  men1  mtap  muc16
## 1 -0.7400  4.0518 -0.5201 -0.2491 -1.0700 -0.4401 -1.7847  0.4840 -0.1399
## 2 -0.3074 -0.2627  0.5183  0.4693  0.4975 -0.1830  2.0132  0.6991 -0.7662
## 3 -1.4649  1.4443  0.2173 -0.2498  0.2100 -0.4972 -0.6047 -1.0802 -0.3453
## 4 -0.2056  0.6378  0.6124 -0.4743  1.2555 -1.2748 -1.8707 -0.6785  2.2705
## 5 -0.8198 -0.4067  0.0563 11.3981 -0.5294 -0.4843 -0.6125  0.1886 -0.5968
## 6  0.2304 -0.0888 -0.6343 -0.5654  0.3499  0.9730  0.1819  0.4657  0.2256
##      myo1a  myo3a  ncoa3  nek1  nf2  npnt  nr2f1  nr3c1  nras
## 1 -1.3288 -0.1871 -0.2318 -0.7678 -0.5124 -1.1541  0.9129  1.5544 -0.3113
## 2 -0.9760  1.4785 -0.0219  0.2494 -0.8002  1.1126 -0.4449  0.1142  1.1293
## 3 -1.1474 -0.1573 -0.2147  3.0991  0.6707  0.4893  2.0438 -0.4219  1.7783
## 4 -0.2572 -0.2659  3.2845  1.6192 -0.4457  0.6028 -0.8912  2.1985  0.5188
## 5 -0.1359 -1.2848  0.1167 -1.8565  0.8564 -0.5695 -0.4462 -0.2680  0.9580
## 6 -1.3054 -1.2631  0.5862  1.4648  0.5299  1.0635 -0.8682 -0.2192  1.3277
##      nrg3  nt5e  or6a2  palld  pbrm1  ppp2cb  ppp2r2a  prkacg  prkce
## 1 -0.3845  1.7567  1.3489  0.1345  0.1833  0.3977  1.3159 -0.8578  0.5443
## 2  0.4071  0.3615 -0.0552  1.0493 -1.0334  1.8291  0.8355 -0.4891  2.0953
## 3  1.5159  0.5243  0.7059  1.8646  1.7470  4.3730 -1.5965 -0.3600 -0.6999
## 4  0.1731  2.1346  0.3979  1.1317  1.4847  4.9811 -1.0110  0.3507 -0.5121
## 5 -0.6391 -0.0071 -0.9548 -0.7848  0.4775 -0.2867 -2.5448 -0.9880 -0.3974
## 6  0.0070  2.1647  0.8452  0.9250 -0.6904  0.2598 -0.4024  0.7375 -0.9735
##      prkcq  prkcz  prkg1  prps2  prr16  ptpn22  ptpm  rasgef1b  rpgr
## 1  0.3109 -1.8952  0.8269 -1.4296  0.6743 -1.2848  1.7244  0.6254  2.1977
## 2 -0.8368  1.2026 -0.9437  0.2955 -0.1490 -0.9459  0.2569 -0.5997 -0.2163
## 3  1.4239 -1.8042 -1.3318  0.2297 -0.5960  2.0742 -0.3079  0.5588  1.6148
## 4  0.1865  0.2039  1.2360  0.2784 -0.2526  3.3501 -0.6313  0.3029  1.3755
## 5  0.6268 -0.1897 -0.0193 -2.6080  0.2641 -0.0903 -0.9172  0.6662 -0.8037

```

```

## 6 0.8967 -0.8874 0.1783 1.3370 0.0451 0.0035 -0.0864 -0.7041 -0.0620
##      ryr2      sbno1      setd1a      setd2      setdb1      sf3b1      sgcd      shank2      siah1
## 1 -1.4373 -0.6340 0.1715 -0.7240 -1.3426 0.0103 0.3185 -0.1317 0.6696
## 2 -0.1650 -1.1499 -1.4204 1.0934 -1.0891 2.5210 0.7896 -0.4393 0.6933
## 3 1.1700 1.3883 -1.5638 -2.7355 -1.9798 -0.6588 2.0649 1.0082 0.7468
## 4 -0.5139 1.2652 -2.2722 -3.6153 -2.7271 -0.6197 0.5162 0.8421 1.4606
## 5 -1.5695 0.5120 -0.1065 0.5528 -1.0253 1.0402 -0.1148 -0.1429 -0.1627
## 6 0.3629 0.4572 -1.1197 -0.4740 -0.8448 0.6623 1.8478 0.5896 1.6416
##      sik1      sik2      smarcb1      smarcc1      smarcc2      smarcd1      spaca1      stab2      stmn2
## 1 3.3718 5.7407 -0.2313 -0.2688 1.3924 -0.8209 -0.0393 1.5674 0.6176
## 2 -0.0651 -1.9196 0.3384 -1.3411 0.0250 -0.2070 -0.6883 -0.0915 1.0232
## 3 0.4396 0.0909 -0.5603 0.2472 -2.3825 -2.6745 0.0532 -0.0486 0.4847
## 4 -0.9108 -0.3750 0.0969 -0.6907 -1.8951 -1.8433 0.2254 -0.9919 1.4694
## 5 -0.1813 -0.3797 -1.6013 0.8841 0.2401 -1.2328 0.1099 0.2682 -0.5840
## 6 -0.9837 -0.0076 -0.7990 0.0162 -0.2341 -0.4519 -1.1549 -0.8653 -0.0367
##      syne1      taf1      taf4b      tbl1xr1      tg      thada      thsd7a      ttyh1      ubr5
## 1 3.5103 0.0980 -0.4821 -0.2392 -1.2637 0.8521 4.0274 2.0126 -1.6059
## 2 0.7871 3.5023 0.2932 -0.0918 -0.5201 -0.5883 -0.6051 -0.7237 0.5109
## 3 1.6215 -1.8384 0.7701 -0.4117 -0.8488 -1.4498 -1.3763 -0.3872 0.4933
## 4 1.3334 -1.9635 3.1831 -0.4433 -0.2497 -1.2420 0.2360 -0.3080 1.6069
## 5 -0.6586 0.0868 -0.5795 -0.3344 -0.3768 -0.3361 -0.2675 -0.1509 3.1012
## 6 0.4099 0.0511 -0.3753 0.3371 0.7847 0.7389 -0.5280 -0.4994 0.3742
##      ush2a      usp9x      utrnl      zfp361l      ackr3      akr1c1      akr1c2      akr1c3      akr1c4
## 1 0.9776 -0.0678 1.6410 0.3492 1.9395 0.0707 2.7077 1.9038 3.1926
## 2 -0.7401 0.3202 0.1828 1.4963 1.4264 1.3935 -0.8240 -0.7537 -0.4241
## 3 0.8019 0.0390 -0.5732 2.8160 1.1780 -0.1044 -0.3182 -0.5528 -0.3453
## 4 0.1163 0.3602 -0.1134 2.0714 2.4700 0.5877 -0.9752 -0.7100 -0.8394
## 5 -0.5005 0.6024 -0.7561 -0.0912 1.3426 -0.7445 0.6585 -0.0784 0.6210
## 6 -0.4287 -0.1533 0.0460 0.3209 1.0843 0.1077 -0.5853 0.0042 -0.3056
##      akt3      ar      bche      cdk8      cdkn2c      cyb5a      cyp11a1      cyp11b2      cyp17a1
## 1 0.8128 -0.3513 1.2200 -0.1165 6.4965 0.6802 2.3311 0.5169 1.4722
## 2 -0.0711 0.4480 -0.0890 -1.0050 0.5189 1.2116 -0.2053 -2.1169 -0.2640
## 3 1.5564 -0.0780 0.3554 -0.6829 1.1287 0.9132 -0.7445 -0.3764 -0.6322
## 4 0.7053 -1.7853 0.6284 0.5619 0.2592 -0.6110 -0.5875 0.4810 0.1082
## 5 -0.5713 -0.6808 -0.0286 -0.0587 2.6149 -0.1401 -1.1262 1.2664 1.1137
## 6 -0.6313 0.3439 0.6035 -0.6751 -0.1684 0.3139 -0.3901 0.9760 -0.1454
##      cyp19a1      cyp21a2      cyp3a43      cyp3a5      cyp3a7      ddc      hes6      hsd17b1      hsd17b10
## 1 0.3968 -0.7765 -0.9321 1.4916 -0.1297 -0.3049 -1.6573 0.2581 -0.0487
## 2 -0.8541 -1.2046 -0.4826 -0.7198 -0.5679 -0.1981 0.0306 -0.4467 -0.0693
## 3 -0.3206 0.9513 -0.0321 -0.3119 0.5541 -0.3686 1.1041 -0.6253 0.0082
## 4 -0.6410 -1.4589 0.1806 -0.4545 1.0612 -0.6271 0.9094 4.8189 0.0960
## 5 0.0991 0.2388 -0.9798 -0.4957 1.0556 -0.9671 2.2482 -0.0590 0.2796
## 6 -0.4500 0.5214 -2.1118 -0.5430 -0.2316 -0.1529 0.6200 -0.4705 0.0219
##      hsd17b11      hsd17b12      hsd17b13      hsd17b14      hsd17b2      hsd17b3      hsd17b4      hsd17b6      hsd17b7
## 1 1.6822 1.7409 8.8731 -0.6494 0.4931 1.0860 0.2650 -1.8717 -1.7598
## 2 -0.7837 0.0117 1.5355 0.7590 -0.5652 0.0623 -1.2693 1.0729 1.1973
## 3 1.6822 0.8981 -0.0943 1.1599 0.1442 0.1493 0.1117 1.6262 2.2685
## 4 1.4099 -0.6707 -0.3589 -0.8998 -0.5063 0.2546 -0.6586 1.7024 0.1617
## 5 0.0493 -0.7074 0.0696 -0.4491 -0.5634 -0.7627 -0.7051 0.6065 -0.0141
## 6 0.8457 0.2548 -0.4682 -0.7314 -0.3240 -0.3282 1.1360 1.5011 -0.4428
##      hsd17b8      hsd3b1      hsd3b2      hsd3b7      mecom      met      ncoa2      nrp1      pik3r3
## 1 0.5445 0.5018 0.7345 -0.9899 1.1211 0.1674 -0.3421 0.4076 -0.6199
## 2 -2.9847 -1.9199 0.6433 -0.4801 -1.8732 -1.0840 -0.7220 -0.1878 -1.0623
## 3 -0.9910 0.1767 -0.9128 -0.0006 0.7079 -0.7401 1.2279 -0.2882 -0.1727

```

```

## 4  1.4208  0.1320  1.8754 -1.8850  0.1474  0.5097 -0.0393 -0.2049 -0.3290
## 5  0.7040 -2.0938 -0.1260 -0.6658  0.6451  0.5497  4.1999  0.2832  0.4018
## 6  0.6873  0.1311 -0.4149 -0.9105 -0.8526 -0.7718  0.0466 -0.1657  1.0143
##      prkci  prkd1      ran      rdh5      sdc4  serpini1      shbg  slc29a1      sox9
## 1 -1.2582  1.8407  0.6630  5.0109  0.5008  0.0052 -0.4357  0.9859  1.2386
## 2  1.7653  0.3500 -0.2505 -0.6337 -0.1047  0.0222 -0.2938 -1.0821 -1.3206
## 3  1.1257  0.8306  0.2707 -0.7554 -0.3559 -0.7735 -0.1387 -0.9122  1.2552
## 4  3.2140  2.4162  1.7962 -1.2505 -0.1742 -0.2858 -0.7305 -1.0178 -0.7887
## 5  0.1308 -0.5351  1.2930  1.2971 -0.8885 -0.5545  0.0266  0.5328  0.1858
## 6  1.3033  1.3506  0.2875 -1.0378  1.1967 -0.3843  0.4306  0.0314  0.6884
##      spry2  srd5a1  srd5a2  srd5a3      st7      star      tnk2      tulp4  ugt2b15
## 1  2.8796 -1.1877 -0.0194 -1.6345 -0.2142 -0.5698 -1.1741 -1.4779 -0.5954
## 2  0.2446 -0.4412  0.4534  0.4068  0.7634  0.0231  0.9121 -0.9538 -0.2264
## 3  0.4593 -0.5381  0.0668  0.8344  1.7227  0.4024 -3.7172 -1.5538  1.3701
## 4  1.3361 -0.5630 -0.7078  0.8228  0.6819 -0.1948 -2.3286 -0.9924 -0.3154
## 5 -0.3201 -0.5845 -0.3544 -1.0150  2.2961  0.1817 -0.1572  0.0427  5.0048
## 6 -0.4448 -0.3910  0.0395  0.8191  0.1977 -0.2302 -0.5092 -0.3021  0.9357
##      ugt2b17  ugt2b7  pik3ca_mut  tp53_mut  muc16_mut  ahnak2_mut  kmt2c_mut  syne1_mut
## 1 -0.8847 -0.3354          0          0          0          0          0          0
## 2  0.5398 -0.8920          0      H178P          0          0          0          0
## 3 -0.1078  0.3655      H1047R          0          0          0          0          0
## 4  0.2320 -0.4828      E542K          0          0          0          0          0
## 5  3.8476  1.3223          0      S241F          0          0          0          0
## 6 -0.4217  0.5340      H1047R  P67Qfs*56          0          0          0          0
##      gata3_mut  map3k1_mut  ahnak_mut  dnah11_mut  cdh1_mut  dnah2_mut  kmt2d_mut
## 1          0          0          0          0          0          0          0
## 2          0          0          0          0          0          0          0
## 3          0          0          0          0          0          0          0
## 4          0          0          0          0          0          0          0
## 5          0          0          0          0          0          0          0
## 6      A318T          0          0          0          0          0          0
##      ush2a_mut  ryr2_mut  dnah5_mut  herc2_mut  pde4dip_mut  akap9_mut  tg_mut  birc6_mut
## 1          0          0          0          0          0          0          0          0
## 2          0          0          0          0          0          0          0          0
## 3          0          0          0          0          0          0          0          0
## 4          0          0          0          0          0          0          0          0
## 5          0          0          0          0          0          0  F2321Y          0
## 6          0          0          0          0          0          0          0          0
##      utrn_mut  tbx3_mut  col6a3_mut  arid1a_mut  lama2_mut  notch1_mut  cbfb_mut
## 1          0          0          0          0          0          0          0
## 2          0          0          0          0          0          0          0
## 3          0          0          0          0          0          0          0
## 4          0          0          0          0          0          0          0
## 5          0          0          0          0          0          0          0
## 6          0          0          0          0          0          0          0
##      ncor2_mut  col12a1_mut  col22a1_mut  pten_mut  akt1_mut  atr_mut  thada_mut
## 1          0          0          0          0          0          0          0
## 2          0          0          0          0          0          0          0
## 3          0          0          0          0          0          0          0
## 4          0          0          0          0          0          0          0
## 5          0          0          0          0          0          0          0
## 6          0          0          0          0          0          0          0
##      ncor1_mut  stab2_mut  myh9_mut  runx1_mut  nf1_mut  map2k4_mut  ros1_mut  lamb3_mut
## 1          0          0          0          0          0          0          0          0

```

## 2	0	0	0	0	0	0	0	0
## 3	L1463P	0	0	0	0	0	0	0
## 4	0	0	0	0	0	0	0	0
## 5	0	0	0	0	0	0	0	0
## 6	Q661*	0	0	0	0	0	0	0
##	arid1b_mut	erbb2_mut	sf3b1_mut	shank2_mut	ep300_mut	ptprd_mut	usp9x_mut	
## 1	0	0	0	0	0	0	0	
## 2	0	0	0	0	0	0	0	
## 3	0	0	0	0	0	0	0	
## 4	0	0	0	0	0	0	0	
## 5	0	0	0	0	0	0	0	
## 6	0	0	0	0	0	0	0	
##	setd2_mut	setd1a_mut	thsd7a_mut	afdn_mut	erbb3_mut	rb1_mut	myo1a_mut	alk_mut
## 1	0	0	0	0	0	0	0	0
## 2	0	0	0	0	0	0	0	0
## 3	0	0	0	0	0	0	0	0
## 4	0	0	0	0	0	0	0	0
## 5	0	0	0	0	0	0	0	0
## 6	0	0	0	0	0	0	0	0
##	fanca_mut	adgra2_mut	ubr5_mut	pik3r1_mut	myo3a_mut	asxl2_mut	apc_mut	ctcf_mut
## 1	0	0	0	0	0	0	0	0
## 2	0	0	0	0	0	0	0	0
## 3	0	0	0	0	0	0	0	0
## 4	0	0	0	0	0	0	0	0
## 5	0	0	0	0	0	0	0	0
## 6	0	0	0	0	0	0	0	0
##	asxl1_mut	fancd2_mut	taf1_mut	kdm6a_mut	ctnna3_mut	brca1_mut	ptprm_mut	
## 1	0	0	0	0	0	0	0	
## 2	0	0	0	0	0	0	0	
## 3	0	0	0	0	0	0	0	
## 4	0	0	0	0	0	0	0	
## 5	0	0	0	0	0	0	0	
## 6	0	0	0	0	0	0	0	
##	foxo3_mut	usp28_mut	glc3c_mut	brca2_mut	cacna2d3_mut	arid2_mut	aff2_mut	
## 1	0	0	0	0	0	0	0	
## 2	A499Gfs*16	0	0	0	0	0	0	
## 3	0	0	0	0	0	0	0	
## 4	0	0	0	0	0	0	0	
## 5	0	0	0	0	0	0	0	
## 6	0	0	0	0	0	0	0	
##	lifr_mut	sbno1_mut	kdm3a_mut	ncoa3_mut	bap1_mut	l1cam_mut	pbrm1_mut	chd1_mut
## 1	0	0	0	0	0	0	0	0
## 2	0	0	0	0	0	0	0	0
## 3	0	0	0	0	0	0	0	0
## 4	0	0	0	0	0	0	0	0
## 5	0	0	0	0	0	0	0	0
## 6	0	0	0	0	0	0	0	0
##	jak1_mut	setdb1_mut	fam20c_mut	arid5b_mut	egfr_mut	map3k10_mut	smarcc2_mut	
## 1	0	0	0	0	0	0	0	
## 2	0	0	0	0	0	0	0	
## 3	0	0	0	0	0	0	0	
## 4	0	0	0	0	0	0	0	
## 5	0	0	0	0	0	0	0	
## 6	0	0	0	0	0	0	0	

##	erbb4_mut	npnt_mut	nek1_mut	agmo_mut	zfp3611_mut	smad4_mut	sik1_mut	casp8_mut
## 1	0	0	0	0	0	0	0	0
## 2	0	0	0	0	0	0	0	0
## 3	0	0	0	0	0	0	0	0
## 4	0	0	0	0	0	0	0	0
## 5	0	0	0	0	0	0	0	0
## 6	0	0	0	0	0	0	0	0
##	prkcq_mut	smarcc1_mut	palld_mut	dcaf4l2_mut	bcas3_mut	cdkn1b_mut	gps2_mut	
## 1	0	0	0	0	0	0	0	
## 2	0	0	0	0	0	0	0	
## 3	0	0	0	0	0	0	0	
## 4	0	0	0	0	0	0	0	
## 5	0	0	0	0	0	0	0	
## 6	0	0	0	0	0	0	0	
##	men1_mut	stk11_mut	sik2_mut	ptpn22_mut	brip1_mut	flt3_mut	nrg3_mut	fbxw7_mut
## 1	0	0	0	0	0	0	0	0
## 2	0	0	0	0	0	0	0	0
## 3	0	0	0	0	0	0	0	0
## 4	0	0	0	0	0	0	0	0
## 5	0	0	0	0	0	0	0	0
## 6	0	0	0	0	0	0	0	0
##	ttyh1_mut	taf4b_mut	or6a2_mut	map3k13_mut	hdac9_mut	prkacg_mut	rpgr_mut	
## 1	0	0	0	0	0	0	0	
## 2	0	0	0	0	0	0	0	
## 3	0	0	0	0	0	0	0	
## 4	0	0	0	0	0	0	0	
## 5	0	0	0	0	0	0	0	
## 6	0	0	0	0	0	0	0	
##	large1_mut	foxp1_mut	clk3_mut	prkcz_mut	lipi_mut	ppp2r2a_mut	prkce_mut	
## 1	0	0	0	0	0	0	0	
## 2	0	0	0	0	0	0	0	
## 3	0	0	0	0	0	0	0	
## 4	0	0	0	0	0	0	0	
## 5	0	0	0	0	0	0	0	
## 6	0	0	0	0	0	0	0	
##	gh1_mut	gpr32_mut	kras_mut	nf2_mut	chek2_mut	ldlrap1_mut	clrn2_mut	acvr11_mut
## 1	0	0	0	0	0	0	0	0
## 2	0	0	0	0	0	0	0	0
## 3	0	0	0	0	0	0	0	0
## 4	0	0	0	0	0	0	0	0
## 5	0	0	0	0	0	0	0	0
## 6	0	0	0	0	0	0	0	0
##	agtr2_mut	cdkn2a_mut	ctnna1_mut	magea8_mut	prr16_mut	dtwd2_mut	akt2_mut	
## 1	0	0	0	0	0	0	0	
## 2	0	0	0	0	0	0	0	
## 3	0	0	0	0	0	0	0	
## 4	0	0	0	0	0	0	0	
## 5	0	0	0	0	0	0	0	
## 6	0	0	0	0	0	0	0	
##	braf_mut	foxo1_mut	nt5e_mut	ccnd3_mut	nr3c1_mut	prkg1_mut	tbl1xr1_mut	
## 1	0	0	0	0	0	0	0	
## 2	0	0	0	0	0	0	0	
## 3	0	0	0	0	0	0	0	
## 4	0	0	0	0	0	0	0	

```
## 5      0      0      0      0      0      0      0
## 6      0      0      0      0      0      0      0
##   frmd3_mut smad2_mut sgcd_mut spaca1_mut rasgef1b_mut hist1h2bc_mut nr2f1_mut
## 1      0      0      0      0      0      0      0
## 2      0      0      0      0      0      0      0
## 3      0      0      0      0      0      0      0
## 4      0      0      0      0      0      0      0
## 5      0      0      0      0      0      0      0
## 6      0      0      0      0      0      0      0
##   klrg1_mut mbl2_mut mtap_mut ppp2cb_mut smarcd1_mut nras_mut ndfip1_mut
## 1      0      0      0      0      0      0      0
## 2      0      0      0      0      0      0      0
## 3      0      0      0      0      0      0      0
## 4      0      0      0      0      0      0      0
## 5      0      0      0      0      0      0      0
## 6      0      0      0      0      0      0      0
##   hras_mut prps2_mut smarcb1_mut stmn2_mut siah1_mut
## 1      0      0      0      0      0
## 2      0      0      0      0      0
## 3      0      0      0      0      0
## 4      0      0      0      0      0
## 5      0      0      0      0      0
## 6      0      0      0      0      0
```

Perform exploratory data analysis:

Clean the data, handle missing values, and explore the data structure using dimensionality reduction techniques.

Question requirements: Pre-processing: remove incomplete rows, perform imputation(for handle missing values), and apply PCA.

Visualize missing values

```
missing_columns=which(colSums(is.na(df))!=0)
print(missing_columns)
```

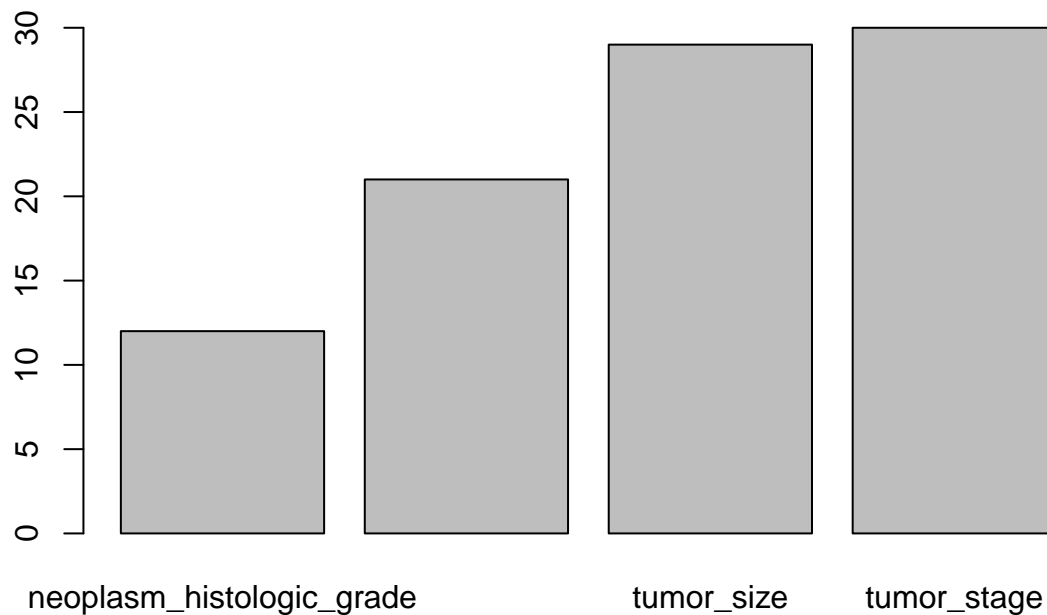
```
## neoplasm_histologic_grade      mutation_count      tumor_size
##                12                21                29
##      tumor_stage
##                30
```

```
#Missing value per column
sum(is.na(df))
```

```
## [1] 638
```

```
#Overall
```

```
barplot(missing_columns)
```



```
df_clean <- df %>% filter(death_from_cancer != "" & !is.na(death_from_cancer))
#Remove rows which don't have target value
```

```
df_clean$patient_id <- NULL
#Remove id column
```

```
summary(df_clean[, 1:10])
```

```
## age_at_diagnosis type_of_breast_surgery cancer_type
## Min. :21.93 Length:1903 Length:1903
## 1st Qu.:51.37 Class :character Class :character
## Median :61.78 Mode :character Mode :character
## Mean :61.09
## 3rd Qu.:70.59
## Max. :96.29
## cancer_type_detailed cellularity chemotherapy
## Length:1903 Length:1903 Min. :0.0000
## Class :character Class :character 1st Qu.:0.0000
## Mode :character Mode :character Median :0.0000
## Mean :0.2081
## 3rd Qu.:0.0000
## Max. :1.0000
## pam50_._claudin.low_subtype cohort er_status_measured_by_ihc
## Length:1903 Min. :1.000 Length:1903
## Class :character 1st Qu.:1.000 Class :character
## Mode :character Median :3.000 Mode :character
## Mean :2.644
## 3rd Qu.:3.000
## Max. :5.000
## er_status
## Length:1903
## Class :character
## Mode :character
```

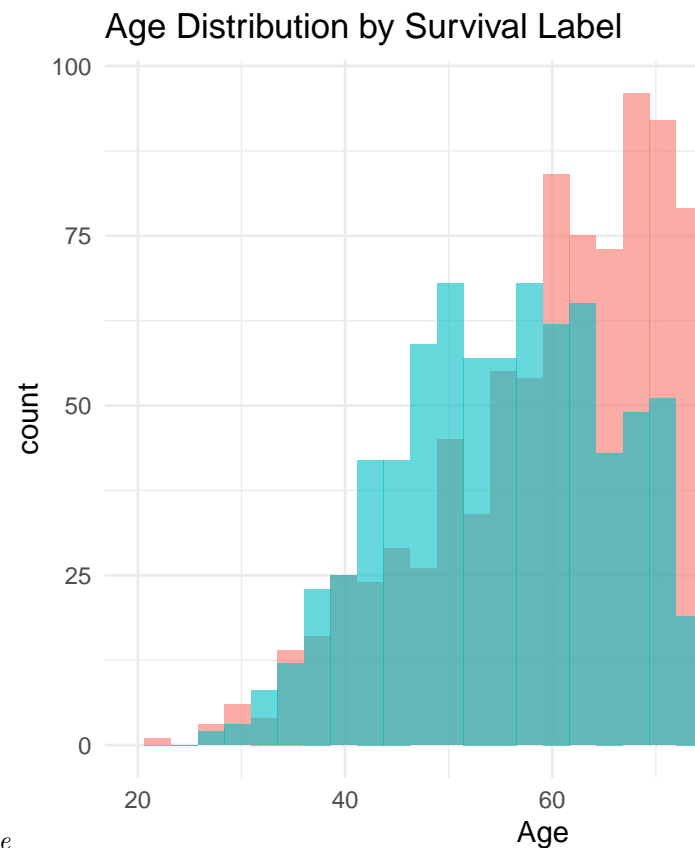
```
##
##
##
```

Separate numeric columns for co analysis and pca then impute missing values with mean

```
numeric_cols <- df_clean %>% select_if(is.numeric)
numeric_cols_imputed <- as.data.frame(lapply(numeric_cols, function(x) {
  if(any(is.na(x))) { x[is.na(x)] <- mean(x, na.rm = TRUE) }
  return(x)
})))
```

Factorize target column

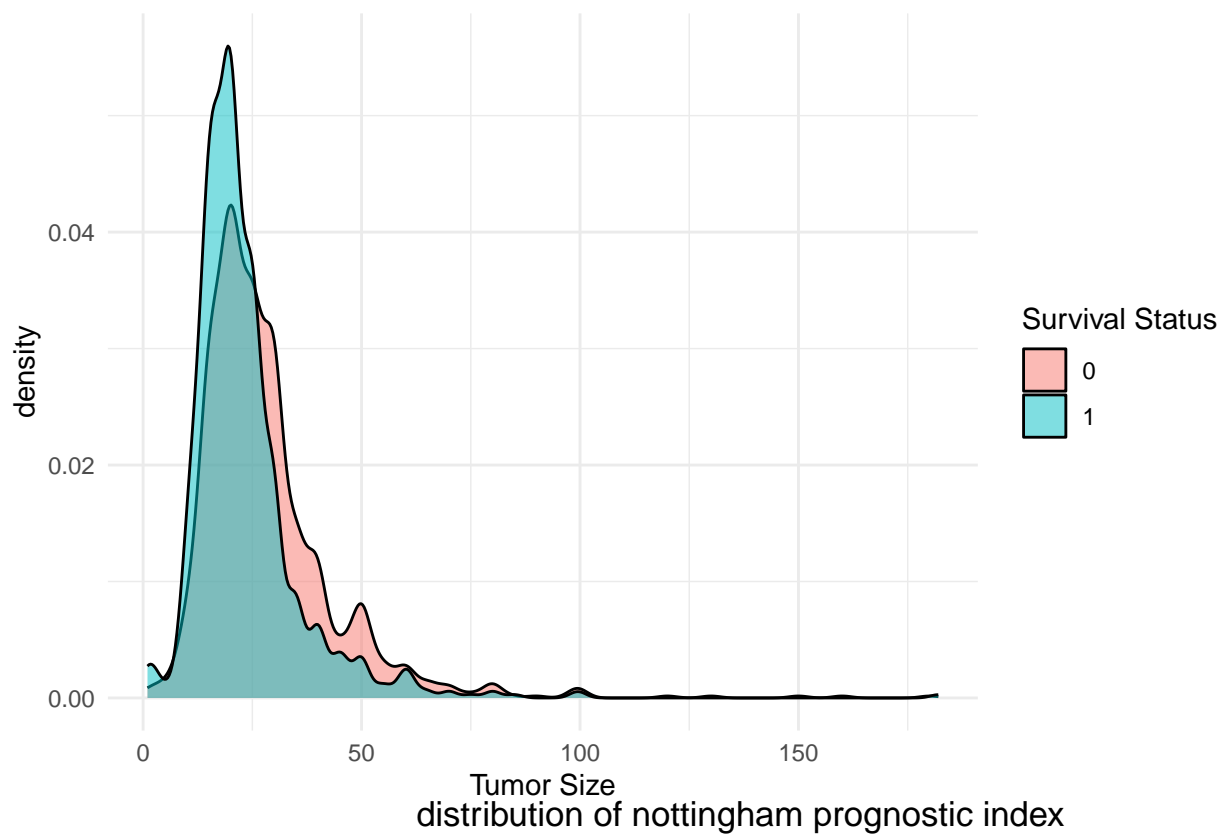
```
df_plot <- df_clean %>%
  filter(!is.na(overall_survival) & overall_survival != "") %>%
  mutate(overall_survival = as.factor(overall_survival))
```



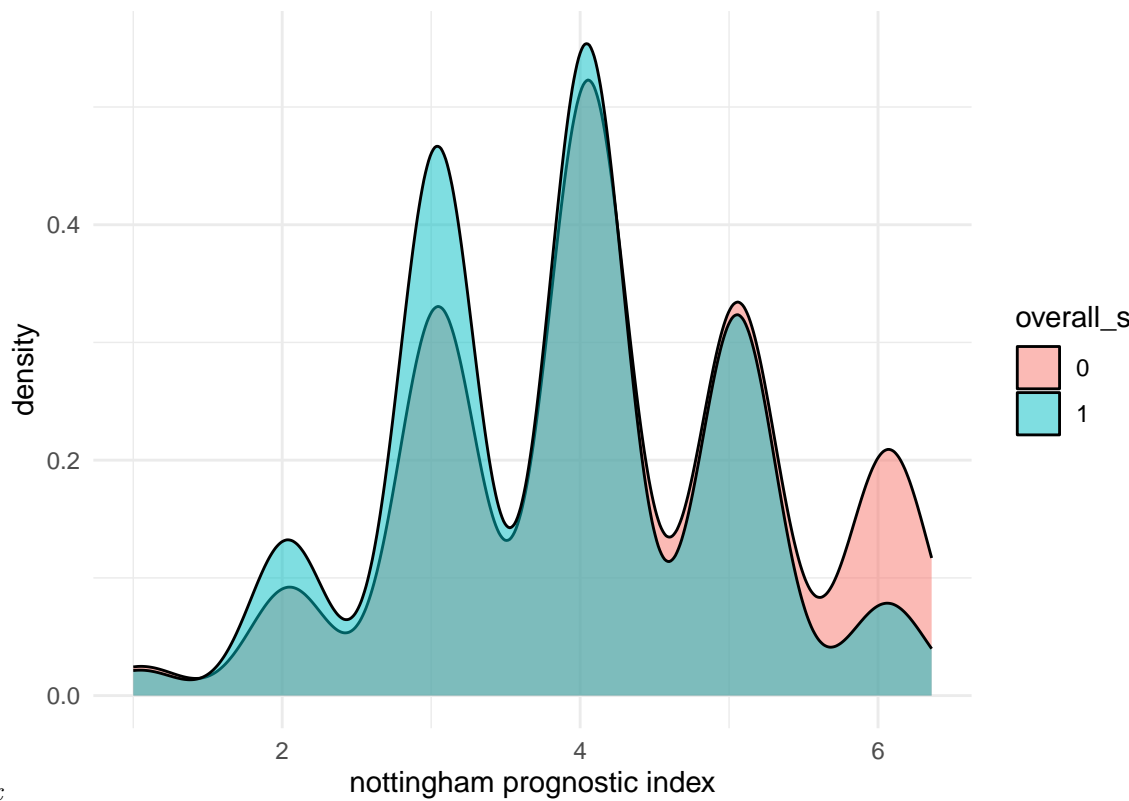
Distribution of some features colored by overall_survival(0 or 1) *age*
Tumor Size

```
## Warning: Removed 20 rows containing non-finite outside the scale range
## (`stat_density()`).
```


Tumor Size Distribution by Survival Label

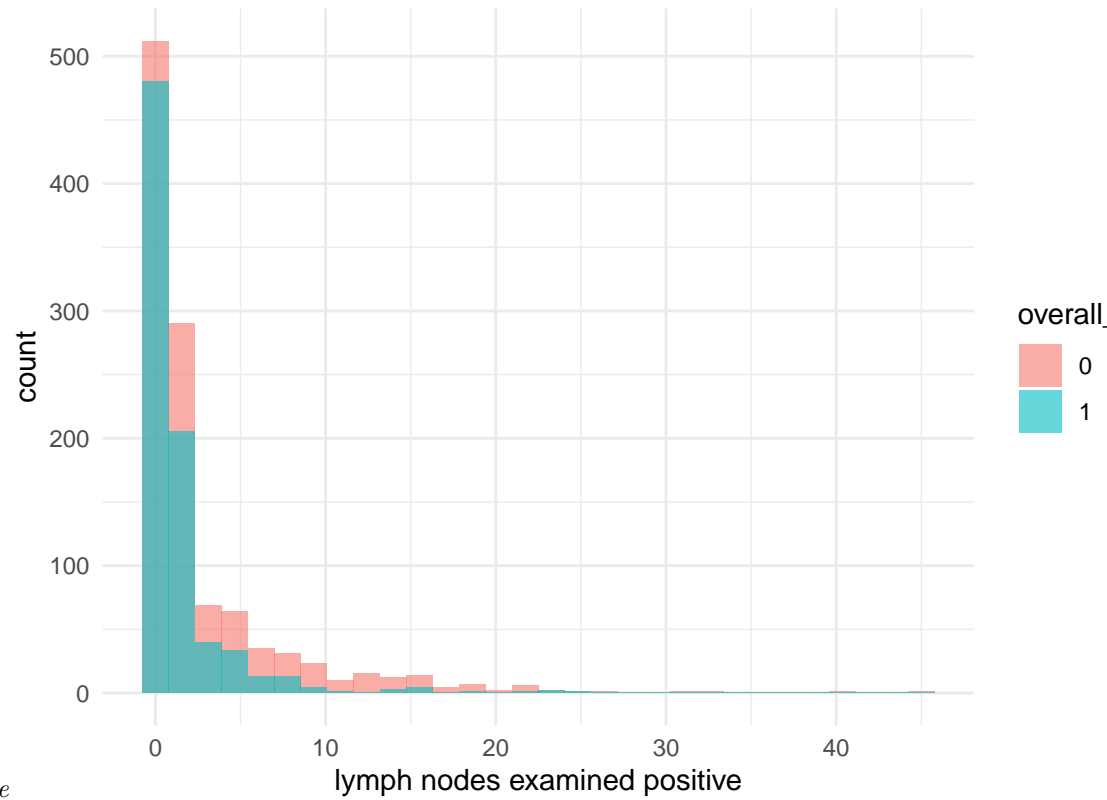


distribution of nottingham prognostic index



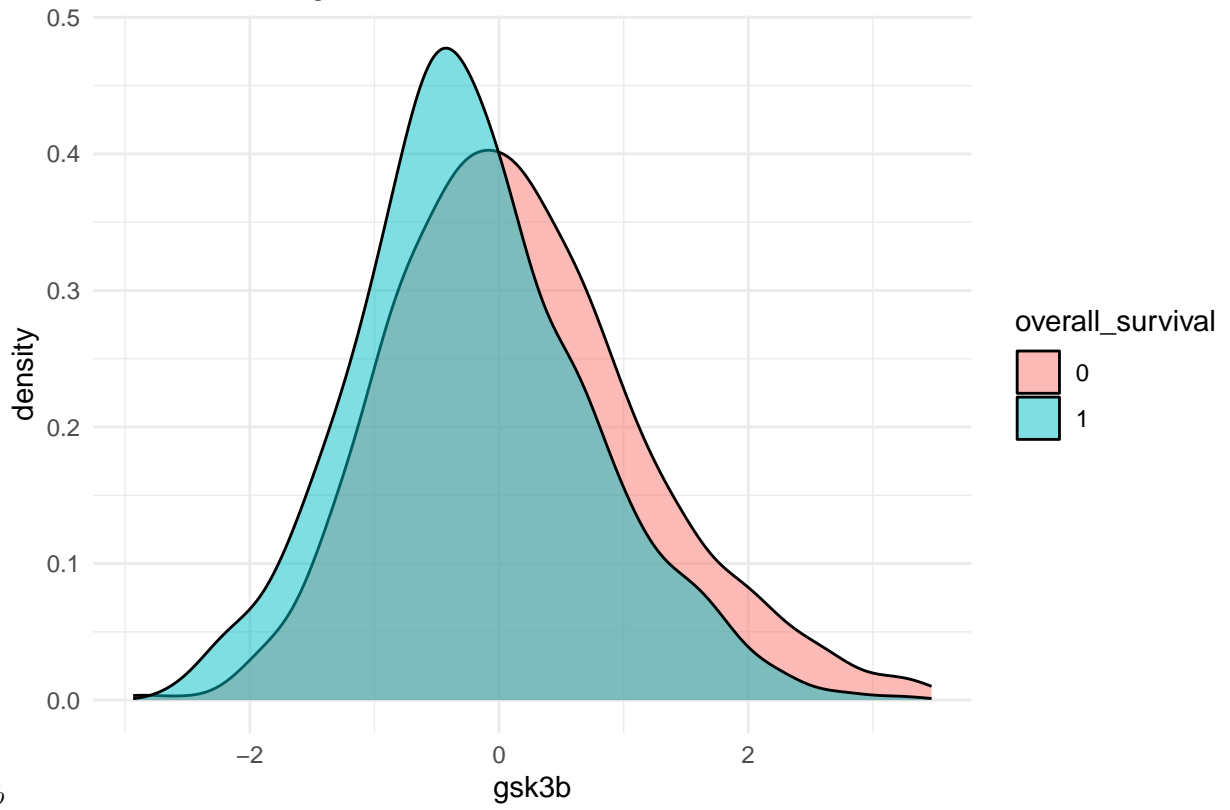
nottingham prognostic index

distribution of lymph nodes examined positive

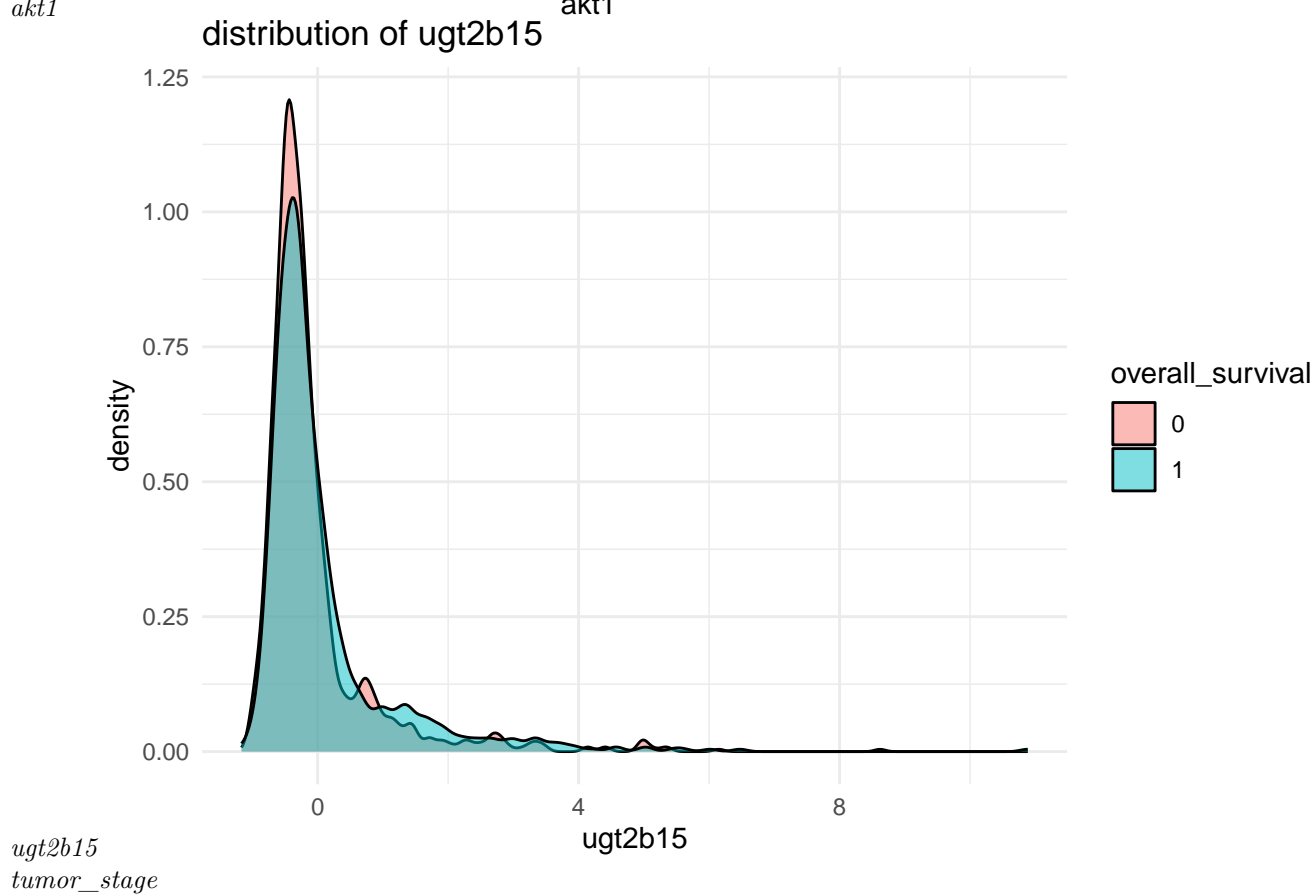
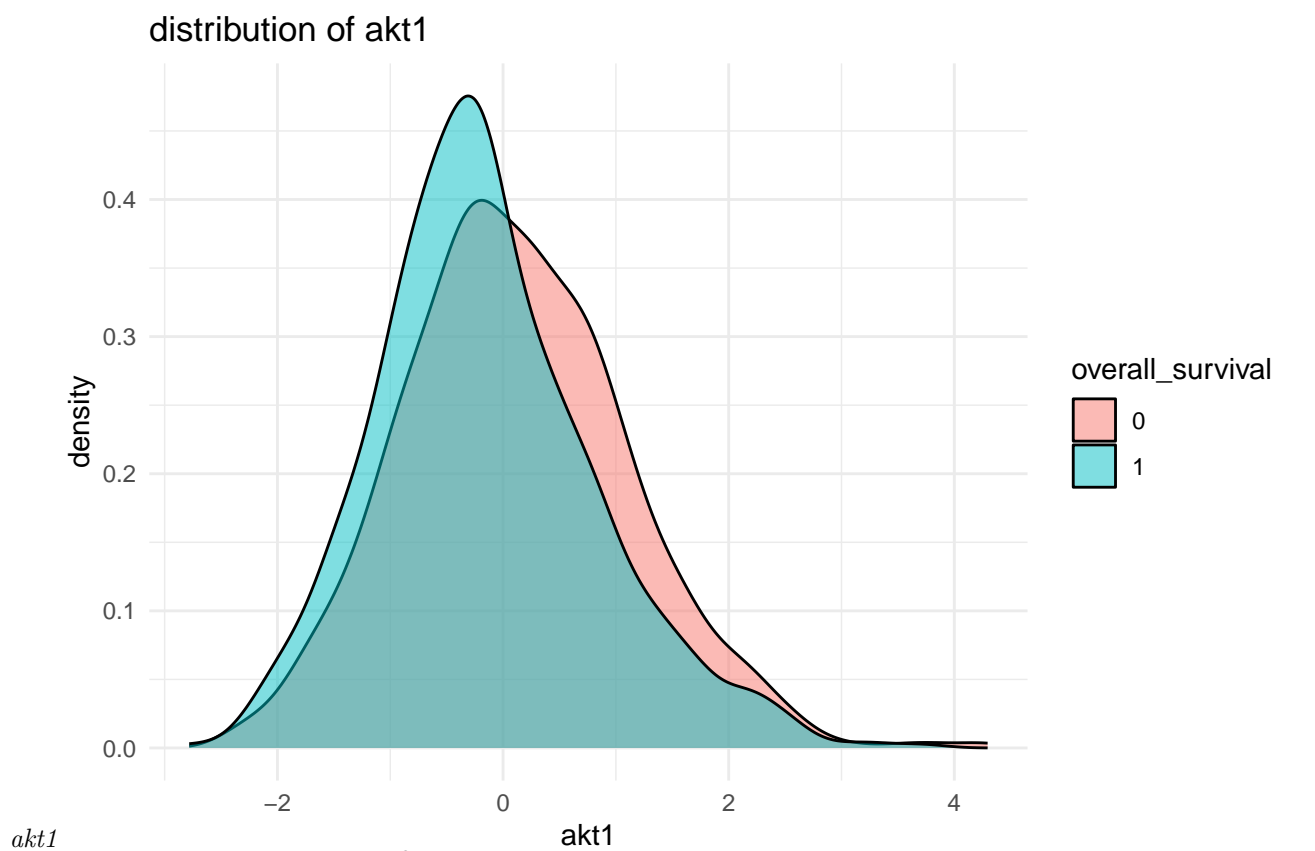


lymph nodes examined positive

distribution of gsk3b



gsk3b

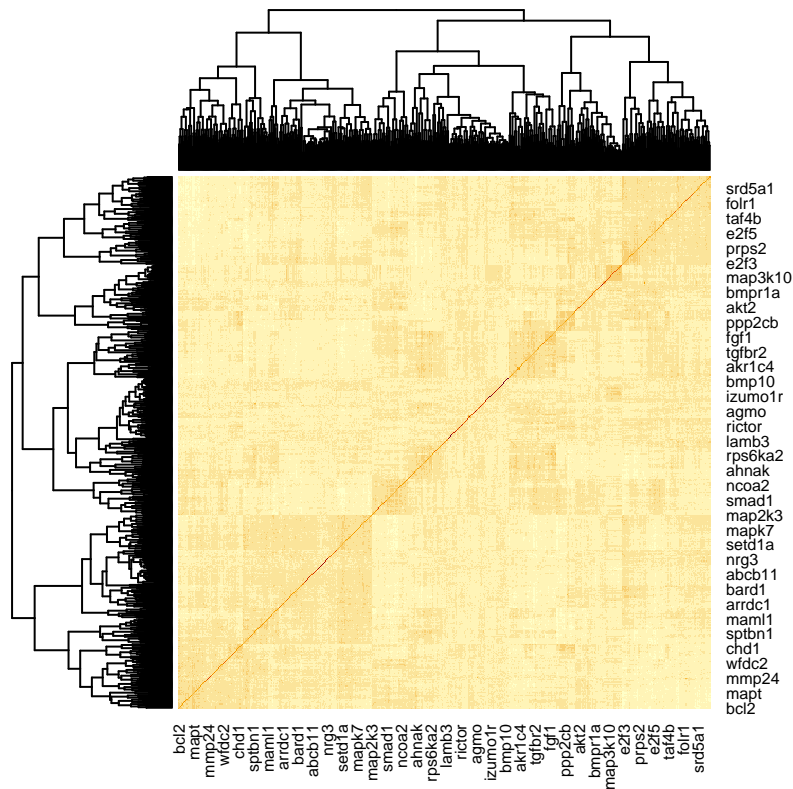


```
## Warning: Removed 501 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



Correlation Analysis

```
cor_matrix <- cor(numeric_cols_imputed)
highly_correlated <- findCorrelation(cor_matrix, cutoff = 0.9)
heatmap(cor_matrix)
```



Remove highly correlated features.

```
if(length(highly_correlated) > 0){
  df_final_numeric <- numeric_cols_imputed[, -highly_correlated]
  print(paste("Number of features removed due to high correlation (>0.9):", length(highly_correlated)))
} else {
  df_final_numeric <- numeric_cols_imputed
  print("No highly correlated features found.")
}
```

```
## [1] "No highly correlated features found."
```

Number of features removed due to high correlation (>0.9):

```
highly_correlated
```

```
## numeric(0)
```

```
length(highly_correlated)
```

```
## [1] 0
```

```
ncol(df)
```

```
## [1] 693
```

```
ncol(df_final_numeric)
```

```
## [1] 502
```

normalize dataframe then apply PCA

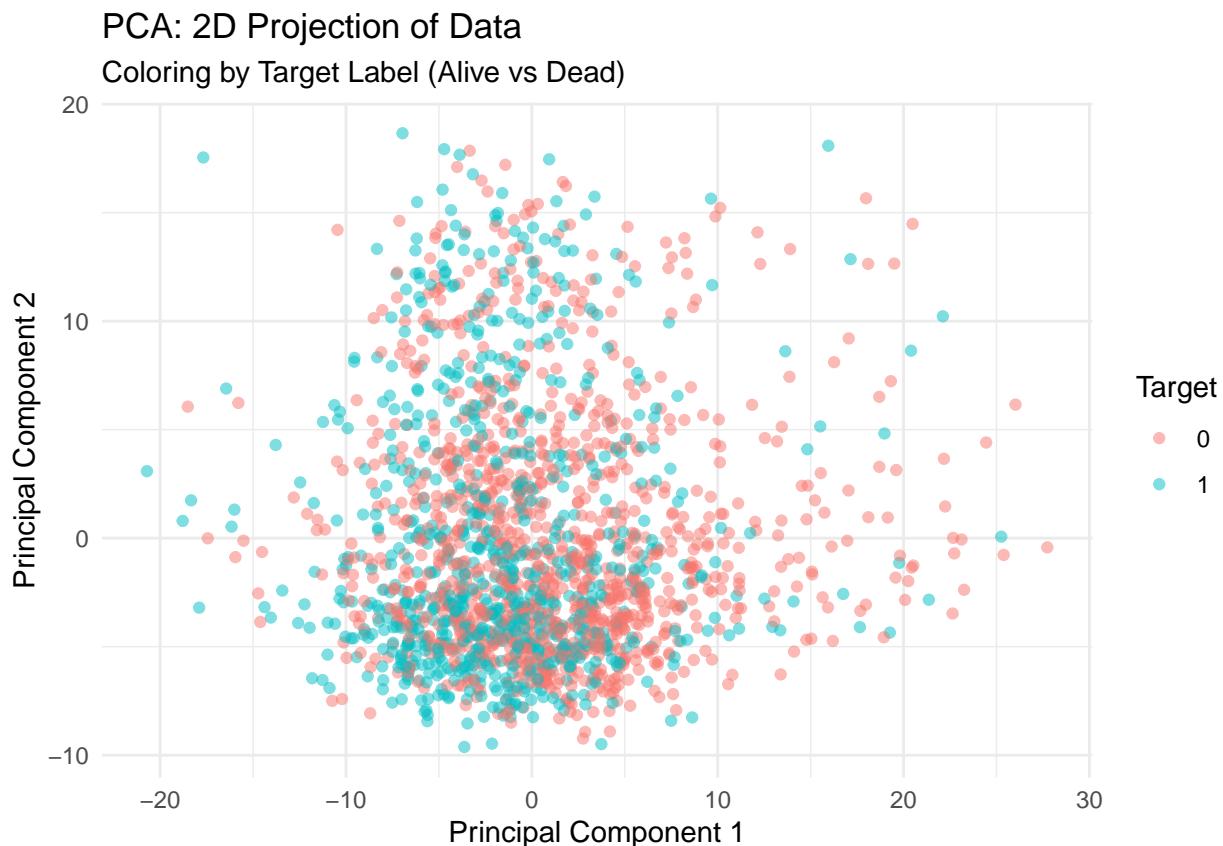
```
#scale Data
```

```
df_scaled <- scale(df_final_numeric)
```

```
#PCA
pca_res <- prcomp(df_scaled, scale. = FALSE)
pca_df <- as.data.frame(pca_res$x[, 1:2]) # 2D
pca_df$Target <- as.factor(df_clean$overall_survival)
```

2 dimension PCA plot colored by overall survival label

```
ggplot(pca_df, aes(x = PC1, y = PC2, color = Target)) +
  geom_point(alpha = 0.5) +
  theme_minimal() +
  labs(title = "PCA: 2D Projection of Data",
       subtitle = "Coloring by Target Label (Alive vs Dead)",
       x = "Principal Component 1", y = "Principal Component 2")
```



According to the above plot, no clear separation can be observed; therefore, to achieve better class prediction, it is preferable to incorporate additional auxiliary variables.

```
#Final data
final_data <- data.frame(df_final_numeric, death_from_cancer = df_clean$death_from_cancer)
```

Part 2: Baseline Classification Models (Lasso & Ridge)

```
#Turn target column to binary type
df_binary <- df_clean %>%
  filter(death_from_cancer %in% c("Died of Disease", "Living")) %>%
  mutate(death_from_cancer = factor(death_from_cancer))
```

```
y <- df_binary$death_from_cancer
```

Remove features with Var=0 or Leakage(A function from target column)

```
leakage_cols <- c("overall_survival_months", "overall_survival", "vital_status",
                  "death_from_cancer", "patient_id")
```

```
X_all <- df_binary %>%
  select(-one_of(intersect(names(.), leakage_cols))) %>%
  select_if(is.numeric) %>%
  select(which(apply(., 2, var, na.rm=TRUE) > 0))
```

Imputation missing value with mean

```
X_all <- as.data.frame(lapply(X_all, function(x) {
  x[is.na(x)] <- mean(x, na.rm = TRUE)
  return(x)
})))
```

Separate features to non-genetic, genetic and both.

```
clinical_features <- names(X_all)[1:min(31, ncol(X_all))]
genetic_features  <- names(X_all)[(min(31, ncol(X_all))+1):ncol(X_all)]

feature_subsets <- list(
  "Clinical_Only" = clinical_features,
  "Genetic_Only"  = genetic_features,
  "Combined"      = names(X_all)
)
```

Stratified Split dataset to train-test

```
train_idx <- createDataPartition(y, p = 0.8, list = FALSE)
X_train <- X_all[train_idx, ]
X_test  <- X_all[-train_idx, ]
y_train <- y[train_idx]
y_test  <- y[-train_idx]
```

Create a function for report model metrics Accuracy,AUC,..

```
evaluate_model <- function(model, X_test, y_test) {
  probs <- predict(model, s = "lambda.min", newx = X_test, type = "response")
  preds <- predict(model, s = "lambda.min", newx = X_test, type = "class")

  cm <- confusionMatrix(as.factor(preds), y_test)

  acc <- cm$overall['Accuracy']
  precision <- cm$byClass['Pos Pred Value'] # Precision
  recall    <- cm$byClass['Sensitivity']     # Recall
  f1 <- 2 * (precision * recall) / (precision + recall)
  roc_auc <- auc(roc(y_test, as.vector(probs), quiet = TRUE))

  return(list(Accuracy = acc, AUC = roc_auc, F1 = f1, Precision = precision, Recall = recall))
}
```

Lasso Ridge for feature subsets

```

results_list <- list()

for(subset_name in names(feature_subsets)) {
  cols <- feature_subsets[[subset_name]]

  X_train_sub <- as.matrix(X_all[train_idx, cols])
  X_test_sub <- as.matrix(X_all[-train_idx, cols])
  y_train_sub <- y[train_idx]
  y_test_sub <- y[-train_idx]

  # 5-fold CV Lasso :
  cv_lasso <- cv.glmnet(X_train_sub, y_train_sub, family = "binomial", alpha = 1, nfolds = 5)
  res_lasso <- evaluate_model(cv_lasso, X_test_sub, y_test_sub)

  # 5-fold CV Ridge :
  cv_ridge <- cv.glmnet(X_train_sub, y_train_sub, family = "binomial", alpha = 0, nfolds = 5)
  res_ridge <- evaluate_model(cv_ridge, X_test_sub, y_test_sub)

  #save results
  results_list[[paste0(subset_name, "_Lasso")]] <- res_lasso
  results_list[[paste0(subset_name, "_Ridge")]] <- res_ridge
}

```

Comparison Table :

```

comparison_df <- do.call(rbind, lapply(results_list, as.data.frame))
print(comparison_df)

```

	Accuracy	AUC	F1	Precision	Recall
## Clinical_Only_Lasso	0.6795775	0.7251008	0.5919283	0.6666667	0.5322581
## Clinical_Only_Ridge	0.6830986	0.7173387	0.6086957	0.6603774	0.5645161
## Genetic_Only_Lasso	0.6232394	0.6791331	0.5069124	0.5913978	0.4435484
## Genetic_Only_Ridge	0.6161972	0.6798891	0.5022831	0.5789474	0.4435484
## Combined_Lasso	0.6936620	0.7454133	0.6133333	0.6831683	0.5564516
## Combined_Ridge	0.6760563	0.7189012	0.6166667	0.6379310	0.5967742

Based on the above results, we conclude that a combination of genetic and non-genetic features leads to a better model. Furthermore, we observe that LASSO performs better than Bridge because LASSO eliminates less important features.

```

# Best Coef in Lasso on combined model
best_lasso_coefs <- coef(cv_lasso, s = "lambda.min")
important_features <- as.matrix(best_lasso_coefs)
important_features <- important_features[important_features != 0, , drop=FALSE]
print(head(important_features, 20))

```

	lambda.min
## (Intercept)	2.492401141
## age_at_diagnosis	-0.017046408
## lymph_nodes_examined_positive	-0.084080436
## nottingham_prognostic_index	-0.174152307
## radio_therapy	0.116326443
## tumor_size	-0.008200294
## tumor_stage	-0.110136750
## chek2	0.042500283


```
## mlh1                0.104816164
## ccnb1               -0.168406007
## jak2                0.050011340
## stat2              -0.075160277
## stat5a             0.214720365
## cul1               0.076953423
## hey1              -0.024339845
## mam11              -0.055875676
## notch2             0.044600036
## rbpj1             -0.020001908
## hes2               0.019627764
## acvr1c             0.021550084
```

Part 3: Ensemble and Nonlinear Models - FULLY COMPLIANT

```
# 5-fold CV :
ctrl <- trainControl(method = "cv", number = 5,
                     classProbs = TRUE,
                     summaryFunction = twoClassSummary,
                     savePredictions = "final")
#Use caret library for c-v

subsets <- list(
  Clinical = clinical_features,
  Genetic = genetic_features,
  Combined = names(X_all)
)
# Feature_subsets
```

Make syntactically valid names out of character vectors for train/test.

```
y_train_f <- make.names(y_train)
y_test_f  <- make.names(y_test)

final_results_p3 <- data.frame()
```

Train models with tuned hyperparameters for all models

```
for(s_name in names(subsets)){
  cols <- subsets[[s_name]]

  #Random Forest :
  print(paste("Tuning RF for:", s_name))
  rf_tuned <- train(x = X_train[, cols], y = y_train_f,
                   method = "rf",
                   ntree = 100,
                   trControl = ctrl,
                   metric = "ROC",
                   tuneGrid = expand.grid(mtry = c(2, 5, 10)))

  # SVM
  print(paste("Tuning SVM for:", s_name))
  svm_tuned <- train(x = X_train[, cols], y = y_train_f,
                   method = "svmRadial",
```

```

        trControl = ctrl,
        metric = "ROC",
        tuneLength = 5)

# KNN
print(paste("Tuning KNN for:", s_name))
knn_tuned <- train(x = X_train[, cols], y = y_train_f,
                  method = "knn",
                  trControl = ctrl,
                  metric = "ROC",
                  tuneGrid = expand.grid(k = c(3, 5, 7, 9)))

# Decision Tree
print(paste("Tuning Decision Tree for:", s_name))
dt_tuned <- train(x = X_train[, cols], y = y_train_f,
                 method = "rpart", trControl = ctrl,
                 metric = "ROC", tuneLength = 10)

# AdaBoost
print(paste("Tuning AdaBoost for:", s_name))
ada_tuned <- train(x = X_train[, cols], y = y_train_f,
                  method = "gbm", trControl = ctrl,
                  metric = "ROC", verbose = FALSE)
temp_models <- list(RF=rf_tuned, SVM=svm_tuned, KNN=knn_tuned, DT=dt_tuned, AdaBoost=ada_tuned)

for(m_name in names(temp_models)){
  m <- temp_models[[m_name]]
  # best result in CV
  best_row <- m$results[which.max(m$results$ROC), ]

  # Test on hold-out dataset (test)
  preds <- predict(m, newdata = X_test[, cols])
  cm <- confusionMatrix(preds, as.factor(y_test_f))

  # report :
  new_row <- data.frame(
    Subset = s_name,
    Model = m_name,
    ROC_Mean = best_row$ROC,
    ROC_SD = best_row$ROCSd,
    Accuracy = cm$overall['Accuracy']
  )
  final_results_p3 <- rbind(final_results_p3, new_row)
}
}

```

```

## [1] "Tuning RF for: Clinical"
## [1] "Tuning SVM for: Clinical"
## [1] "Tuning KNN for: Clinical"
## [1] "Tuning Decision Tree for: Clinical"
## [1] "Tuning AdaBoost for: Clinical"
## [1] "Tuning RF for: Genetic"
## [1] "Tuning SVM for: Genetic"
## [1] "Tuning KNN for: Genetic"
## [1] "Tuning Decision Tree for: Genetic"
## [1] "Tuning AdaBoost for: Genetic"

```

```
## [1] "Tuning RF for: Combined"
## [1] "Tuning SVM for: Combined"
## [1] "Tuning KNN for: Combined"
## [1] "Tuning Decision Tree for: Combined"
## [1] "Tuning AdaBoost for: Combined"
```

```
# Comparison Table
```

```
print(final_results_p3)
```

	Subset	Model	ROC_Mean	ROC_SD	Accuracy
## Accuracy	Clinical	RF	0.7319492	0.03779529	0.6760563
## Accuracy1	Clinical	SVM	0.7263503	0.04305771	0.6830986
## Accuracy2	Clinical	KNN	0.6656729	0.04441370	0.6549296
## Accuracy3	Clinical	DT	0.6547478	0.04090628	0.6690141
## Accuracy4	Clinical	AdaBoost	0.7230557	0.03738038	0.6901408
## Accuracy5	Genetic	RF	0.6904900	0.05780538	0.6267606
## Accuracy6	Genetic	SVM	0.6857091	0.03280443	0.6161972
## Accuracy7	Genetic	KNN	0.6102183	0.05310781	0.6267606
## Accuracy8	Genetic	DT	0.5969233	0.04491897	0.6161972
## Accuracy9	Genetic	AdaBoost	0.6758850	0.02390542	0.6161972
## Accuracy10	Combined	RF	0.7254311	0.03755847	0.6373239
## Accuracy11	Combined	SVM	0.7527862	0.02734056	0.6619718
## Accuracy12	Combined	KNN	0.7014608	0.04153411	0.6866197
## Accuracy13	Combined	DT	0.6482264	0.02852804	0.6514085
## Accuracy14	Combined	AdaBoost	0.7460395	0.01119258	0.6760563

Based on the above results, the AdaBoost on Combined feature subset model performs better than the other models(High Roc mean and less ROC std) all models work better in combined feature subset.

Best

1 : combin adaboost 2 : combin svm 3 : clinical rf

Part 4: Model Evaluation, Comparison, and Interpretation (SHAP)

```
#Global comparision
```

```
final_summary_table <- final_results_p3 %>%
  select(Subset, Model, ROC_Mean, ROC_SD, Accuracy) %>%
  arrange(desc(ROC_Mean))
```

```
print(final_summary_table)
```

	Subset	Model	ROC_Mean	ROC_SD	Accuracy
## Accuracy11	Combined	SVM	0.7527862	0.02734056	0.6619718
## Accuracy14	Combined	AdaBoost	0.7460395	0.01119258	0.6760563
## Accuracy	Clinical	RF	0.7319492	0.03779529	0.6760563
## Accuracy1	Clinical	SVM	0.7263503	0.04305771	0.6830986
## Accuracy10	Combined	RF	0.7254311	0.03755847	0.6373239
## Accuracy4	Clinical	AdaBoost	0.7230557	0.03738038	0.6901408
## Accuracy12	Combined	KNN	0.7014608	0.04153411	0.6866197
## Accuracy5	Genetic	RF	0.6904900	0.05780538	0.6267606
## Accuracy6	Genetic	SVM	0.6857091	0.03280443	0.6161972
## Accuracy9	Genetic	AdaBoost	0.6758850	0.02390542	0.6161972
## Accuracy2	Clinical	KNN	0.6656729	0.04441370	0.6549296
## Accuracy3	Clinical	DT	0.6547478	0.04090628	0.6690141

```
## Accuracy13 Combined      DT 0.6482264 0.02852804 0.6514085
## Accuracy7  Genetic      KNN 0.6102183 0.05310781 0.6267606
## Accuracy8  Genetic      DT 0.5969233 0.04491897 0.6161972
```

Choose best model base on Accuracy

```
# best model ada_tunes
best_model_fit <- ada_tuned
```

```
#sample = 5 -> lack of resource
X_sample <- X_train[sample(1:nrow(X_train), 5), subsets$Combined]
```

```
p_function_fixed <- function(object, newdata) {

  results <- predict(object, newdata = newdata, type = "prob")
  return(results[, grep("Living", colnames(results))])
}
```

```
#predict function
```

```
library(fastshap)
```

```
## Warning: package 'fastshap' was built under R version 4.5.2
```

```
##
```

```
## Attaching package: 'fastshap'
```

```
## The following object is masked from 'package:dplyr':
```

```
##
```

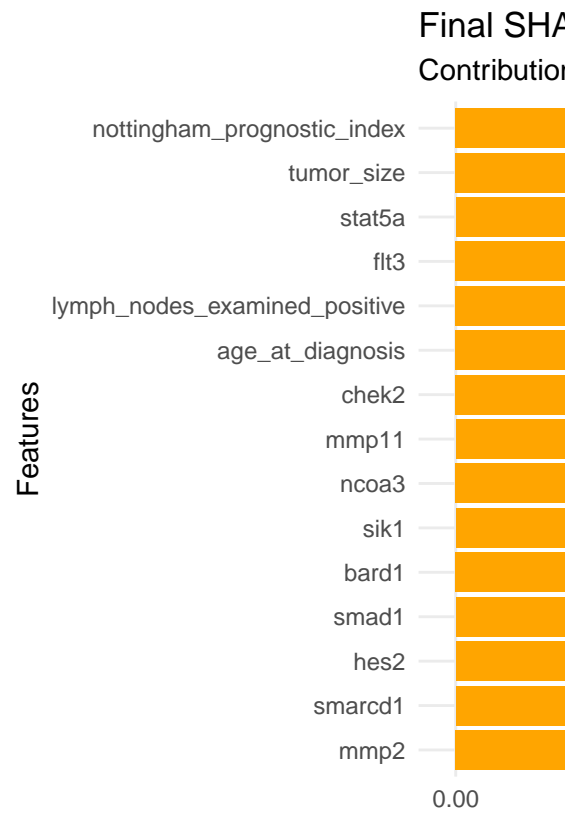
```
## explain
```

```
shap_values <- explain(
  object = best_model_fit,
  X = X_sample,
  pred_wrapper = p_function_fixed,
  nsim = 2
)
```

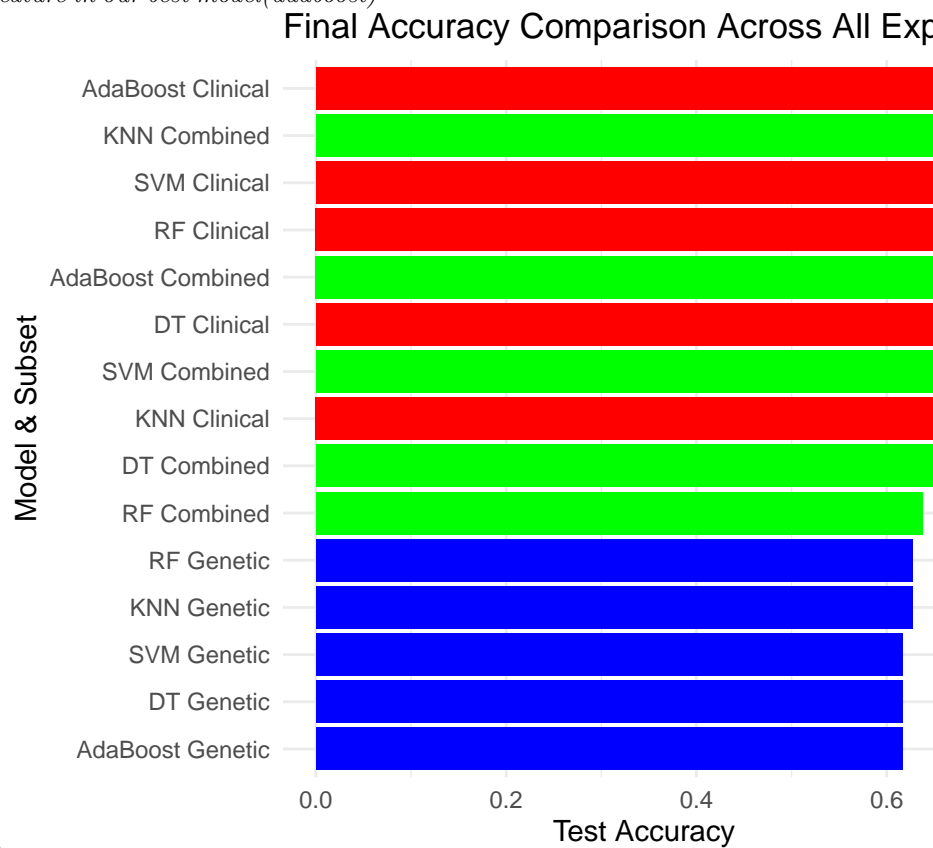
```
#calculating shap
```

Create a data from show importance(SHAP value) per feature in our best model(adaboost)

```
shap_imp_data <- data.frame(
  Feature = colnames(X_sample),
  Importance = colMeans(abs(shap_values))
) %>% arrange(desc(Importance))
```



Plot histogram importance(SHAP value) per feature in our best model(adaboost)



Plot test accuracy of each model in all subsets.