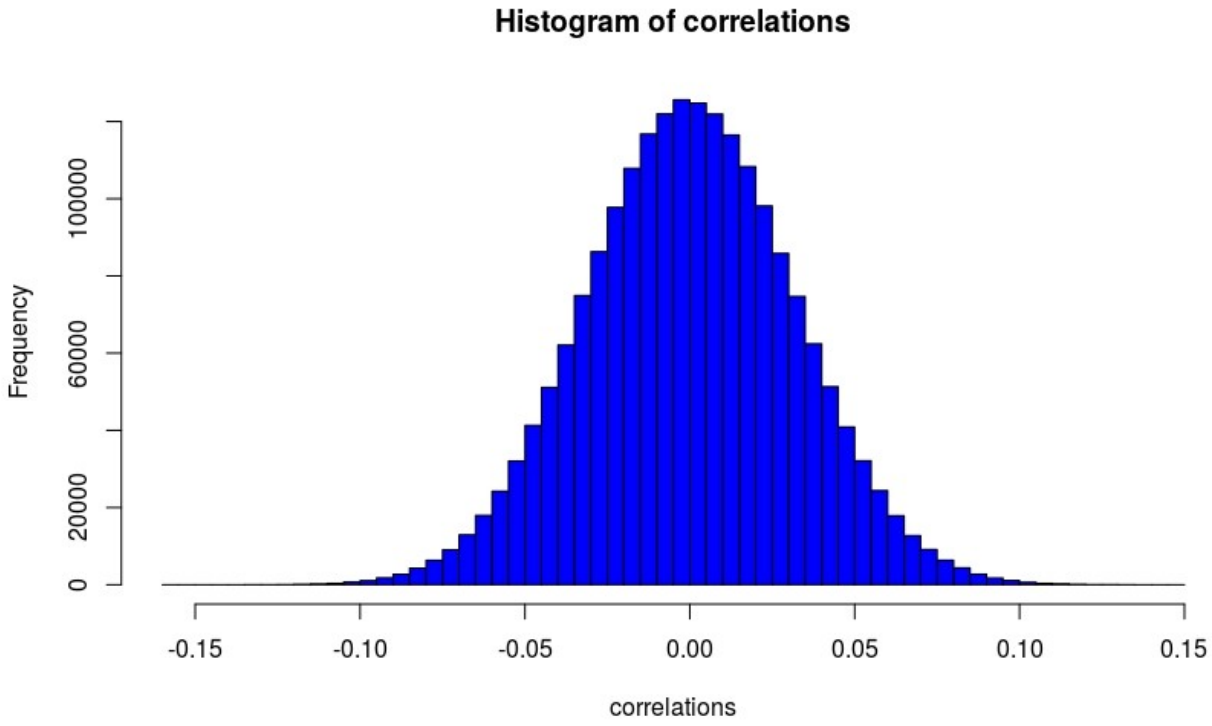
Raport

1. Zbiór Protein.Rdata

W tym zbiorze nie istnieją kolumny o innych nazwach ale identycznej zawartości. Histogram korelacji różnych kolumn zamieściłam poniżej.



Kolumny są słabo skorelowane, najwyższa wartość bezwzględna korelacji między kolumnami wynosi 0.1479086.

Najpierw użyłam regresji liniowej z metodą wyboru podzbiorów forward stepwise (przy użyciu funkcji regsubsets z biblioteki leaps) poszukując najlepszych podzbiorów parametrów posiadających co najwyżej 10 parametrów (czyli z parametrem nvmax=10).
W celu wybrania liczby parametrów prowadzącej do najmniejszego błędu średniokwadratowego użyłam 10-krotnej kroswalidacji. Najmniejszy błąd kroswalidacji (dążący do estymaji błędu testowego) został osiągnięty dla zbiorów 5-elementowych i wynosił 1.081719. Po zastosowaniu metody forward stepwise do całego zbioru treningowego najlepszy podzbiór 5-elementowy zawierał (najważniejsze) predykatory o nazwach x1, x2, x3, x4, x5.

Następnie użyłam regresji liniowej z metodą regularyzacji ridge korzystając z biblioteki glmnet. Użyłam 10-krotnej kroswalidacji na zbiorze w celu wybrania najlepszego parametru lambda metody, który wyniósł 0.6579332 a odpowiadający mu błąd średniokwadratowy z kroswalidacji 412.3325.

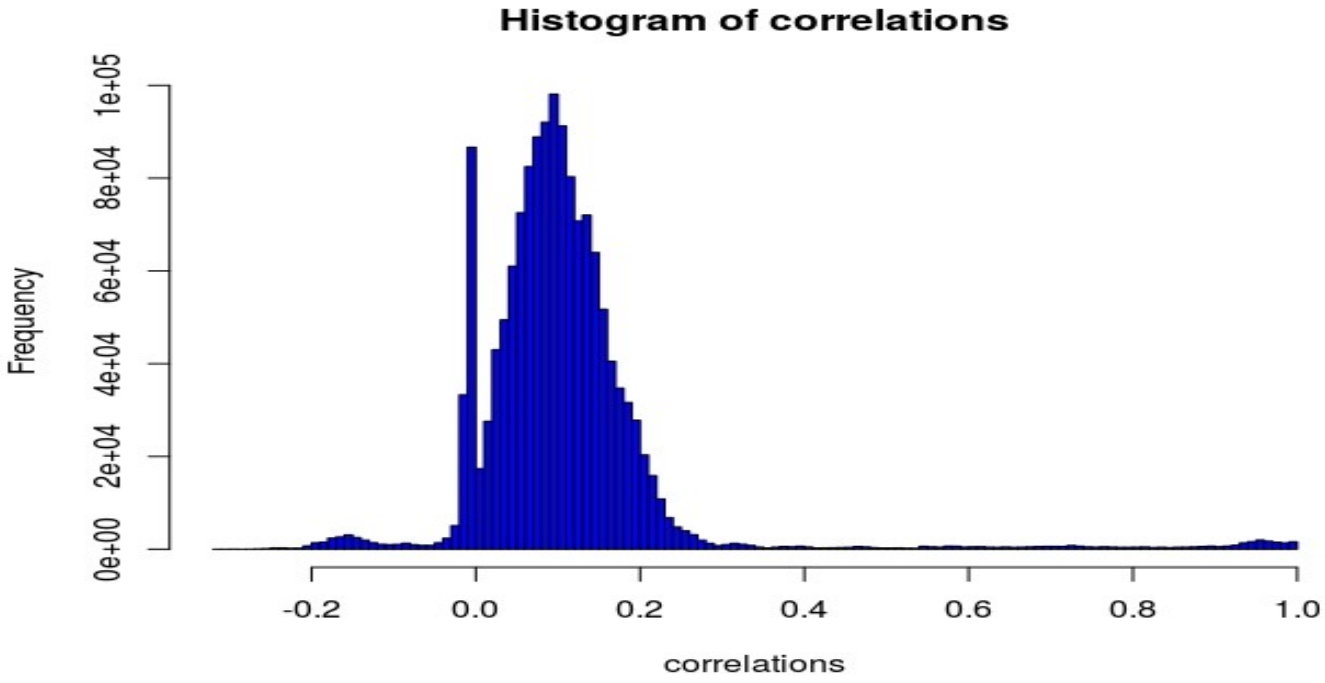
Jako ostatniej użyłam też metody lasso analogicznie jak dla ridge powyżej, dostając z kroswalidacji parametr lambda 0.07054802 i błąd średniokwadratowy 1.128402.

Najlepszy wydaje się być model liniowy używający 5 predykatorów, uzyskany przy użyciu metody forward stepwise, ponieważ prowadzi on do najmniejszego oszacowania błędu z kroswalidacji.

2. Zbiór Cancer.Rdata

Najpierw przekształciłam wszystkie kolumny predyktorów do postaci numerycznej. Były wśród

nich liczne kolumny o innych nazwach ale identycznej zawartości. Po usunięciu powtarzających się kolumn z początkowych 3622 kolumn zostało 1713. Dodatkowo usunęłam jedną zerową kolumnę. Poniżej przedstawiam histogram korelacji różnych pozostałych kolumn.



Około 94,5% wszystkich korelacji zajduje się w przedziale [-0.5, 0.25].
Występują kolumny mocno skorelowane, maksymalna korelacja między kolumnami wyniosła 0.999473. Z tego powodu nie było możliwe skorzystanie z
funkcji anneal z pakietu subselect w celu wybrania najlepszych podzbiorów predyktorów do metody lda (linear discriminant analysis). Żeby zmniejszyć te korelacje zastosowałam funkcję trim.matrix z tego pakietu do macierzy korelaji kolumn, z parametrem tolval=e-3, co pozwoliło zredukować liczbę parametrów do 456. Następnie zastosowałam metodę anneal do wyboru najlepszych 50 i 100 predyktorów dla metody lda. Najlepsze wybrane 50 predyktorów to:

```
 [1] "ACBD4_ac"      "ANKS1B_deac"   "ASB7_deac"     "ATRN_deac"     "C11ORF65_deac"
 [6] "C17ORF97_deac" "CAPZB_ac"      "CASP8_ac"      "CDKN2B_deac"   "CSTF2T_deac"
[11] "DDAH1_deac"    "DDX51_ac"      "DNM2_deac"     "E2F3_ac"       "ESR1_deac"
[16] "FAM122A_ac"    "FBRS_ac"       "FOXP1_deac"    "GCSAML_deac"   "HOXA13_ac"
[21] "IMMP2L_deac"   "IRF2BP1_deac"  "KATNAL1_deac"  "KIAA0368_deac" "MAPK1_deac"
[26] "MDGA2_ac"      "MECOM_deac"    "METTL17_deac"  "MSI2_deac"     "PDE11A_deac"
[31] "PDIA5_ac"      "PHKB_deac"     "PIK3R1_ac"     "PLAGL2_ac"     "PLB1_deac"
[36] "PMAIP1_ac"     "RASGRF2_deac"  "SAMD4B_ac"     "SFN_deac"      "SLC6A12_deac"
[41] "SMAD4_deac"    "SPDYC_ac"      "SPON2_deac"    "TCF7L2_ac"     "TET2_ac"
[46] "TUSC3_deac"    "U2AF1_deac"    "VPS45_ac"      "YWHAZ_ac"      "ZNF703_ac"
```

Najlepsze wybrane 100 predyktorów to:

```
  [1] "ACBD4_deac"    "ACTN2_deac"    "ANK1_deac"     "APOLD1_deac"   "ATM_ac"
  [6] "B2M_deac"      "BRCA2_ac"      "C11ORF65_deac" "C17ORF97_deac" "CACNA1A_deac"
 [11] "CALM3_ac"      "CBFB_ac"       "CCL11_ac"      "CCND1_ac"      "CCSER1_deac"
 [16] "CDH8_deac"     "CDK6_deac"     "CDKN2B_deac"   "COLEC12_ac"    "CREBBP_ac"
 [21] "CRKL_ac"       "CSMD1_deac"    "CSNK2A1_ac"    "CSNK2A1_deac"  "CTNNB1_ac"
 [26] "DDAH1_deac"    "DDX51_ac"      "DPF2_deac"     "E2F3_ac"       "ELP4_deac"
 [31] "EMB_deac"      "ERBB3_ac"      "FAT1_ac"       "GIGYF2_deac"   "GMDS_deac"
```

```
 [36] "GNAQ_deac"     "HCN4_deac"     "HNRNPA3_ac"    "IMMP1L_ac"     "IMMP2L_ac"
 [41] "INSR_deac"     "ITPR1_ac"      "KATNAL1_deac"  "KCNIP4_deac"   "KDM2B_deac"
 [46] "KIF3A_deac"    "KMT2C_deac"    "KRAS_ac"       "LINS1_ac"      "LRP1B_deac"
 [51] "LTBP4_deac"    "LYRM2_ac"      "MAPK1_deac"    "MDGA2_ac"      "MLLT4_deac"
 [56] "MROH1_ac"      "MTOR_deac"     "MYCN_deac"     "NAPSA_deac"    "NFATC1_deac"
 [61] "NKAIN2_deac"   "NKD2_ac"       "NPEPL1_deac"   "NPM1_ac"       "NRG1_deac"
 [66] "NRXN3_deac"    "OR5H6_ac"      "PAK1_deac"     "PARD3_deac"    "PDGFA_ac"
 [71] "PKD2_ac"       "PLGRKT_ac"     "PMAIP1_ac"     "POLR2L_ac"     "PRRX1_ac"
 [76] "PTEN_ac"       "PTPRD_deac"    "RCBTB1_deac"   "REL_ac"        "ROBO1_deac"
 [81] "RTF1_ac"       "SDC3_ac"       "SDK1_deac"     "SMARCA4_ac"    "SMYD3_ac"
 [86] "SOX2_deac"     "TCF7L2_deac"   "TGFBR2_deac"   "TMEM94_deac"   "TNRC6A_deac"
 [91] "TP53_ac"       "TRIM33_ac"     "TUBD1_ac"      "TWF1_deac"     "U2AF1_ac"
 [96] "U2AF1_deac"    "USP22_deac"    "VAV2_ac"       "YWHAZ_deac"    "ZNF703_ac"
```

Zastosowałam trzy różne metody uczenia maszynowego do przewidywania typu nowotworu używając powyższych predyktorów – lda, knn (k-nearest neighbours), i random forest, przewidując dokładnosć tych metod (zdefiniowaną jak w treści zadania) korzystając z 10-krotnej kroswalidacji. Dla metody lda dostałam oszacowanie dokładności 0.533 dla 50 predyktorów i 0.605 dla 100 predyktorów. Dla metody knn dla 50 predyktorów najwyższą dokładność otrzymałam dla k=8 i wyniosła ona 0.482, natomiast dla 100 predyktorów dla k=7 i wyniosła ona 0.492. Dla metody random forest dla 50 predyktorów i z parametrem mtry=7 otrzymałam dokładność 0.536, a dla 100 predyktorów i z parametrem mtry=10 dokładność 0.587.
Dla 50 predyktorów najlepszy wydaje się model random forest, natomiast dla 100 predyktorów – lda, gdyż dla nich otrzymałam największą oszacowaną dokładność z kroswalidacji.