4 užduotis

Duomenys

Požymis: Smoking\Non-smoking

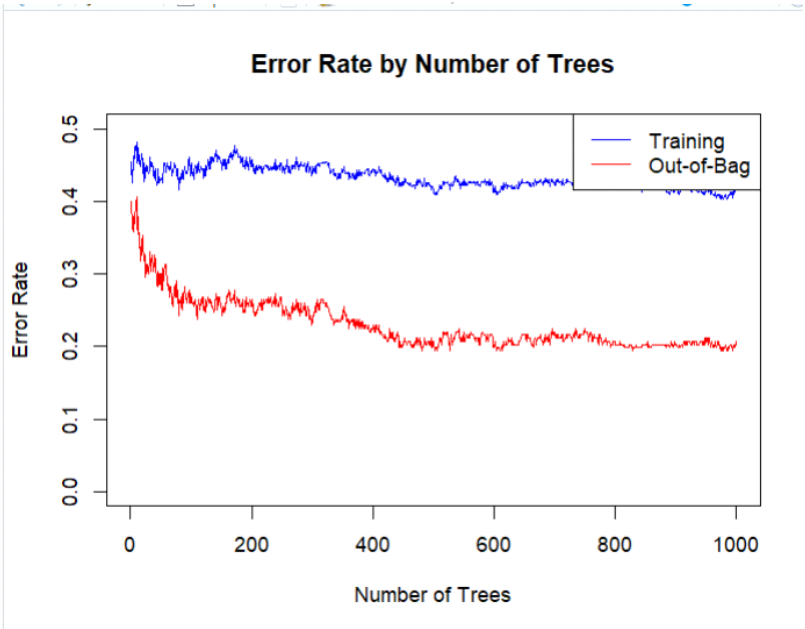
Taip pat, atkiriame 1000 variabiliausių citozino modifikacijos pozicijų

Random Forest:

Duomenų klasifikavimui parinkome „Random Forest“ algoritmą.

```
          Reference
Prediction   non-smoker smoker
  non-smoker         46     19
  smoker             11     17

               Accuracy : 0.6774
                 95% CI : (0.5725, 0.7707)
    No Information Rate : 0.6129
    P-Value [Acc > NIR] : 0.1201

                  Kappa : 0.2912

 Mcnemar's Test P-Value : 0.2012

            Sensitivity : 0.8070
            Specificity : 0.4722
         Pos Pred Value : 0.7077
         Neg Pred Value : 0.6071
             Prevalence : 0.6129
         Detection Rate : 0.4946
   Detection Prevalence : 0.6989
      Balanced Accuracy : 0.6396

       'Positive' Class : non-smoker
```

```
        OOB estimate of  error rate: 41.75%
Confusion matrix:
           non-smoker smoker class.error
non-smoker        181     46   0.2026432
smoker            121     52   0.6994220
```



Matome, kad mūsų modelis gana gerai gali aptažinti „non-smoker“ žmones, bet su rūkančiais žmonėmis mūsų modelis daro nemažai klaidų. Todėl vidutiniška paklaida yra gana didelė, t.y. ~42%.

Validacija:

```
Confusion Matrix and Statistics

            Reference
Prediction   non-smoker smoker
  non-smoker         41     16
  smoker             16     20

               Accuracy : 0.6559
                 95% CI : (0.5502, 0.7514)
    No Information Rate : 0.6129
    P-Value [Acc > NIR] : 0.2293

                  Kappa : 0.2749

 Mcnemar's Test P-Value : 1.0000

            Sensitivity : 0.7193
            Specificity : 0.5556
         Pos Pred Value : 0.7193
         Neg Pred Value : 0.5556
             Prevalence : 0.6129
         Detection Rate : 0.4409
   Detection Prevalence : 0.6129
      Balanced Accuracy : 0.6374

       'Positive' Class : non-smoker
```

Padalinome modelio mokymosi duomenis ir atlikome kryžminę validaciją.
Matome, kad situacija išliko beveik tokią pati. Klasifikatoriaus paklaida yra ~40%.

Išvados apie klasifikavimą:

Manome, kad nors rezultatai ir nėra tobuli, tačiau, pagal kappa rodiklį matome, kad tas atspėjimas nėra atsitiktinis, nors paklaida yra gana didelė. Manau, kad tokios didelės paklaidos priežastis ta, kad mes naudojame ganėtinai mažus duomenis, tikriausiai 1000 pozicijų buvo per mažai, dėl to modelė galėjo „prisitaikyti“ prie duomenų ir dėl to klasių paklaida atrodo tokia „nesubalansuota“

Papildomi klasifikatoriai:

```
                Model  Accuracy       Kappa
1       Random Forest 0.6774194  0.31314623
2                 SVM 0.6559140  0.16357504
3                 KNN 0.5913978  0.12998523
4                 GBM 0.7204301  0.39850746
5 Logistic Regression 0.4946237 -0.02822865
6                 LDA 0.7311828  0.43636364
```

Taip pat pabandėm sudaryti atskirą modelį su kiekvienu iš užduotyje išvardintų klasifikatorių. Tai gavome tokių rezultatų, matome, kad didžiausią „Accuracy“ ir „Kappa“ rodiklį turi GBM ir LDA klasifikatoriai. Taip pat, įdomus rezultatas yra su regresija. Mažiausias „Accuracy“ ir neigiamas „Kappa“, t.y. kad modelis „spėja“ blogiau už atsitiktinį pasirinkimą, kas yra ganėtinai keista.

Principinė komponentė:

Taip pat pabandėme pasinaudoti principine komponente. Pagal mūsų rezultatus matome, kad visgi dominuoja nerūkančiųjų klasė, tačiau su rūkančiųjų klase situacija irgi „pagerėjo“. Bent daugumą rūkančiųjų modelis su principine komponente sugebejo atpažinti, kas yra gan nebelogas rezultatas.

```
Prediction   non-smoker smoker
  non-smoker         40     12
  smoker             17     24

               Accuracy : 0.6882
                 95% CI : (0.5837, 0.7802)
    No Information Rate : 0.6129
    P-Value [Acc > NIR] : 0.08186

                  Kappa : 0.3592

 Mcnemar's Test P-Value : 0.45761

            Sensitivity : 0.7018
            Specificity : 0.6667
         Pos Pred Value : 0.7692
         Neg Pred Value : 0.5854
             Prevalence : 0.6129
         Detection Rate : 0.4301
   Detection Prevalence : 0.5591
      Balanced Accuracy : 0.6842

       'Positive' Class : non-smoker
```