Dataset: Hypothyroid

The dataset originally has 30 columns (29 features and 1 target variable). There are 3772 observations (rows).

**DATA PREPROCESSING**

The dataset has many missing values. Hence, while reading the data into a data frame, the values “?” have been replaced with “NA”

There are a few unnecessary columns like TSH\_measured, T3\_measured, TT4\_measured, T4U\_measured, FTI\_measured and TBG\_measured. Hence, these have been removed. There is a feature called TBG which only NA values. Hence, that column is also removed. The new subset of data has 23 columns (22 features and 1 target variable).

The “age” column has only one NA value and “sex” has 150 NA values. So, these observations are deleted. Now the subset has 23 columns and 3621 observations which has been saved as “hypothyroid\_subset.csv”. The other NA values were handled as follows:

1. Mean imputation

The missing values in the numeric columns “TSH”, “T3”, “TT4”, “T4U”, “FTI” have been imputed by Mean. All numeric values are then normalized and character values converted to factors.

1. KNN imputation

The missing values in the numeric columns have also been imputed by KNN. All numeric values are then normalized and character values converted to factors.

1. Rf impute

The missing values in the numeric columns have also been imputed by RF impute function. All numeric values are then normalized and character values converted to factors.

Out of the 3621 observations, there 3396 with “negative” and 225 with “sick” class. The classes were hence balanced using the following techniques:

1. SMOTE applied on MEAN, KNN and rf imputed dataset:

After applying SMOTE, the observations with negative class are 1800 and those with the sick class are 2025

1. ROSE applied on MEAN, KNN and rf imputed dataset:

After applying ROSE, the observations with negative class are 1820 and those with the sick class are 1801

1. Ovun.sample applied on MEAN, KNN and rf imputed dataset:

After applying ovun.sample, the observations with negative class are 1797 and those with the sick class are 1824

**FEATURE SELECTION**

1. Mean decrease in Gini

Mean-Smote

Column\_Number MeanDecreaseGini

T3 18 814.0031273

referral\_source 22 260.0825868

T4U 20 195.3419623

age 1 189.7599080

FTI 21 121.9174560

TT4 19 117.3072355

TSH 17 67.1540053

on\_thyroxine 3 24.2057139

sex 2 9.1422951

psych 16 6.1007155

query\_hyperthyroid 11 4.3422058

query\_hypothyroid 10 4.0673323

sick 6 2.2124128

goitre 13 1.8047176

tumor 14 1.5819254

thyroid\_surgery 8 1.4998265

query\_on\_thyroxine 4 1.3848575

I131\_treatment 9 1.2464466

on\_antithyroid\_medication 5 0.7543691

hypopituitary 15 0.6122859

pregnant 7 0.4680570

lithium 12 0.3766368

Mean-Rose

Column\_Number MeanDecreaseGini

T3 18 756.0465432

referral\_source 22 254.4308970

T4U 20 165.1007985

TT4 19 123.5725242

FTI 21 117.2531055

age 1 101.1977932

TSH 17 79.0230606

query\_hypothyroid 10 16.4757541

on\_thyroxine 3 16.2550393

sick 6 13.9982320

sex 2 13.5728542

query\_hyperthyroid 11 5.7695398

thyroid\_surgery 8 4.3075105

I131\_treatment 9 3.3045501

goitre 13 2.4436923

query\_on\_thyroxine 4 2.2342846

psych 16 1.9759156

hypopituitary 15 1.7217605

tumor 14 1.5483947

on\_antithyroid\_medication 5 1.1229724

pregnant 7 1.1083610

lithium 12 0.4129565

Mean-OvunSample

Column\_Number MeanDecreaseGini

T3 18 847.9770743

referral\_source 22 234.9708071

T4U 20 162.3476681

TT4 19 146.5859798

FTI 21 128.9706866

age 1 102.5918279

TSH 17 73.3008539

on\_thyroxine 3 11.8394582

query\_hypothyroid 10 7.8111351

sex 2 7.6983726

sick 6 5.8029024

query\_hyperthyroid 11 4.4663561

hypopituitary 15 2.3581717

query\_on\_thyroxine 4 2.0836308

psych 16 1.9869865

tumor 14 1.6431211

pregnant 7 1.2705593

I131\_treatment 9 1.2536472

goitre 13 1.1138598

thyroid\_surgery 8 0.9068260

on\_antithyroid\_medication 5 0.3962473

lithium 12 0.3152360

KNN-Smote

Column\_Number MeanDecreaseGini

T3 18 868.3299607

referral\_source 22 260.5540144

age 1 203.7628214

T4U 20 143.5690396

FTI 21 127.5480698

TT4 19 113.7547297

TSH 17 63.1645332

on\_thyroxine 3 20.3560276

sex 2 9.0260199

psych 16 5.5466235

query\_hyperthyroid 11 4.2402162

query\_hypothyroid 10 3.2469108

sick 6 2.0066240

query\_on\_thyroxine 4 1.5456242

tumor 14 1.3867105

thyroid\_surgery 8 1.2084197

I131\_treatment 9 0.9581929

goitre 13 0.8041837

hypopituitary 15 0.7066149

on\_antithyroid\_medication 5 0.5726835

lithium 12 0.4433625

pregnant 7 0.3510414

KNN-Rose

Column\_Number MeanDecreaseGini

T3 18 777.7647477

referral\_source 22 261.2711229

T4U 20 152.5600484

TT4 19 116.4453861

FTI 21 115.0454687

age 1 98.9911917

TSH 17 75.9224132

on\_thyroxine 3 16.7490396

query\_hypothyroid 10 16.0612153

sick 6 14.5898628

sex 2 13.2409407

query\_hyperthyroid 11 6.0303830

thyroid\_surgery 8 4.2463443

I131\_treatment 9 2.7890932

query\_on\_thyroxine 4 2.5770995

psych 16 1.9983583

hypopituitary 15 1.6771179

tumor 14 1.5004410

on\_antithyroid\_medication 5 1.1438714

pregnant 7 1.0719226

goitre 13 0.7123117

lithium 12 0.3444328

KNN-OvunSample

Column\_Number MeanDecreaseGini

T3 18 894.0042762

referral\_source 22 243.6451251

T4U 20 142.3236388

TT4 19 138.6491326

FTI 21 118.4816182

age 1 96.2658245

TSH 17 71.5051022

on\_thyroxine 3 12.6077571

query\_hypothyroid 10 7.5361426

sex 2 6.6615998

sick 6 4.7475929

query\_hyperthyroid 11 4.1049642

psych 16 2.0572992

hypopituitary 15 1.9169068

query\_on\_thyroxine 4 1.8909790

tumor 14 1.4047267

I131\_treatment 9 1.1349861

pregnant 7 1.0406158

thyroid\_surgery 8 0.7205958

goitre 13 0.4364175

lithium 12 0.3696585

on\_antithyroid\_medication 5 0.3506334

RF-Smote

Column\_Number MeanDecreaseGini

T3 18 783.2797205

referral\_source 22 258.2431578

T4U 20 224.9154181

age 1 195.8883496

FTI 21 128.1915768

TT4 19 115.6168844

TSH 17 64.4939083

on\_thyroxine 3 21.3182959

sex 2 9.6222070

psych 16 5.8732111

query\_hyperthyroid 11 4.0672397

query\_hypothyroid 10 3.9399560

sick 6 2.0878503

goitre 13 1.7325752

tumor 14 1.6225462

query\_on\_thyroxine 4 1.4180853

thyroid\_surgery 8 1.2652155

I131\_treatment 9 0.9644089

hypopituitary 15 0.8411157

lithium 12 0.4421000

pregnant 7 0.3385132

on\_antithyroid\_medication 5 0.3052474

RF-Rose

Column\_Number MeanDecreaseGini

T3 18 751.1361038

referral\_source 22 250.7496155

T4U 20 171.7860330

TT4 19 122.7469453

FTI 21 119.6372904

age 1 102.8471800

TSH 17 76.0282339

query\_hypothyroid 10 16.1714956

on\_thyroxine 3 16.0828655

sick 6 14.8963371

sex 2 14.5176402

query\_hyperthyroid 11 6.1399009

thyroid\_surgery 8 3.8608939

I131\_treatment 9 3.2149157

goitre 13 2.5080287

query\_on\_thyroxine 4 2.0355543

psych 16 2.0022629

hypopituitary 15 1.7814860

tumor 14 1.5232264

pregnant 7 1.1650415

on\_antithyroid\_medication 5 0.6506811

lithium 12 0.3777808

RF-OvunSample

Column\_Number MeanDecreaseGini

T3 18 847.8687688

referral\_source 22 243.6090265

T4U 20 172.4483958

TT4 19 148.2397870

FTI 21 127.2116090

age 1 96.4355368

TSH 17 72.1947754

on\_thyroxine 3 11.4817452

query\_hypothyroid 10 7.8585481

sex 2 7.3553540

sick 6 4.8550565

query\_hyperthyroid 11 3.6608811

hypopituitary 15 1.9190537

psych 16 1.7715659

query\_on\_thyroxine 4 1.7486690

I131\_treatment 9 1.3499048

tumor 14 1.3099414

pregnant 7 1.0241863

goitre 13 0.8885276

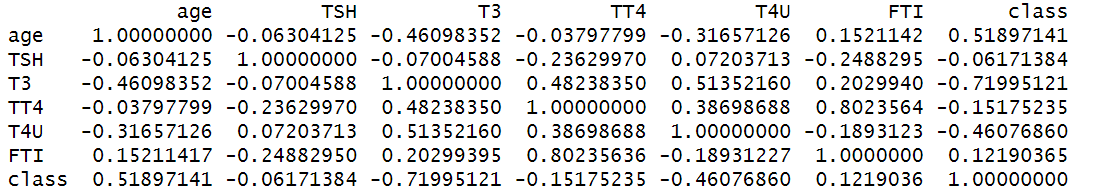
thyroid\_surgery 8 0.7404489

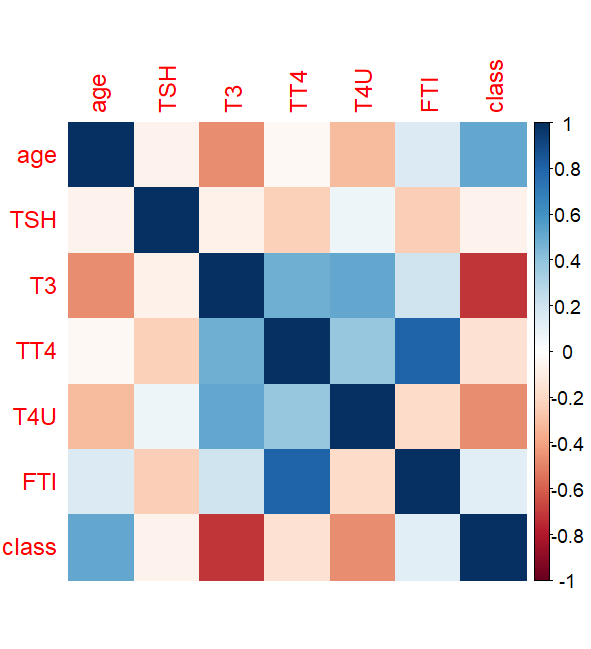
lithium 12 0.4819335

on\_antithyroid\_medication 5 0.2973734

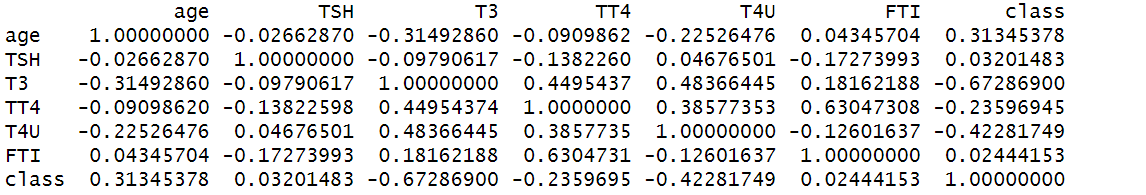
1. Correlation matrix

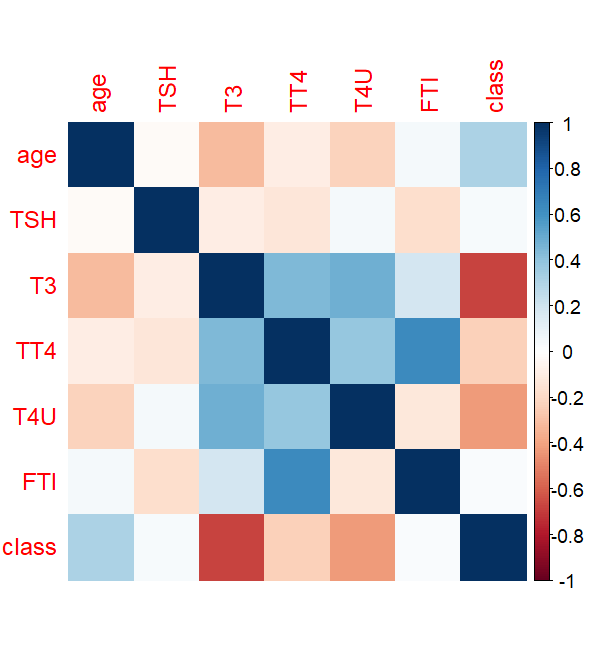
Mean-Smote



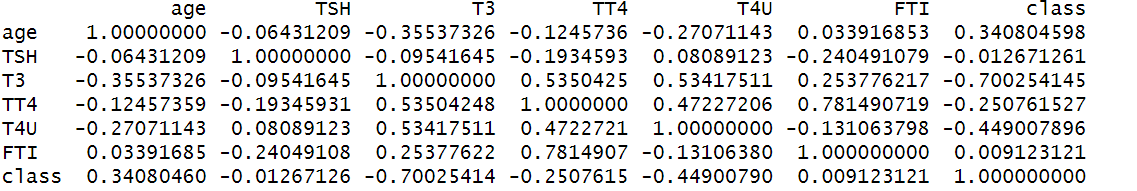


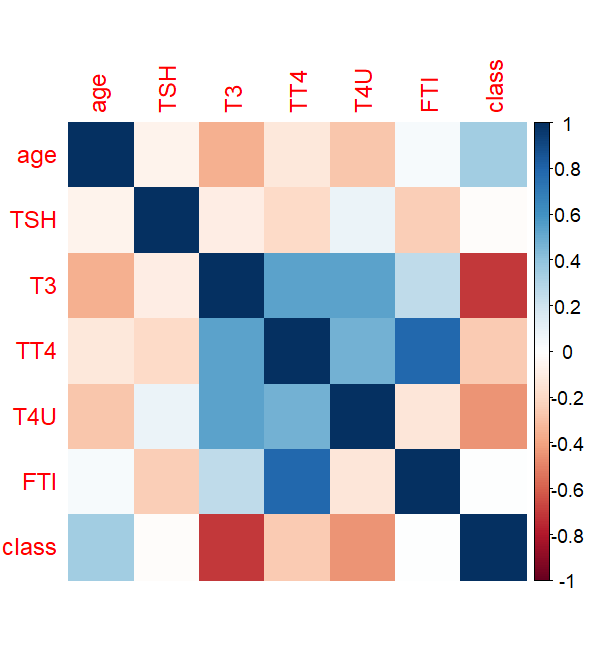
Mean-Rose



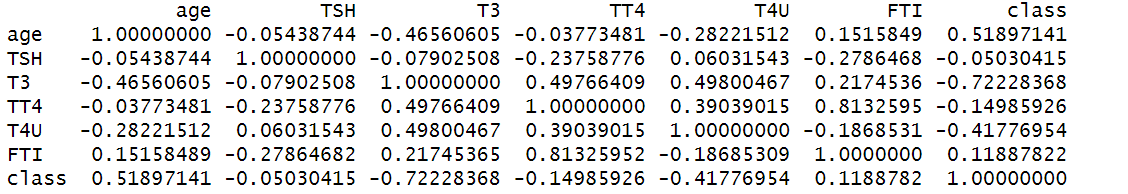


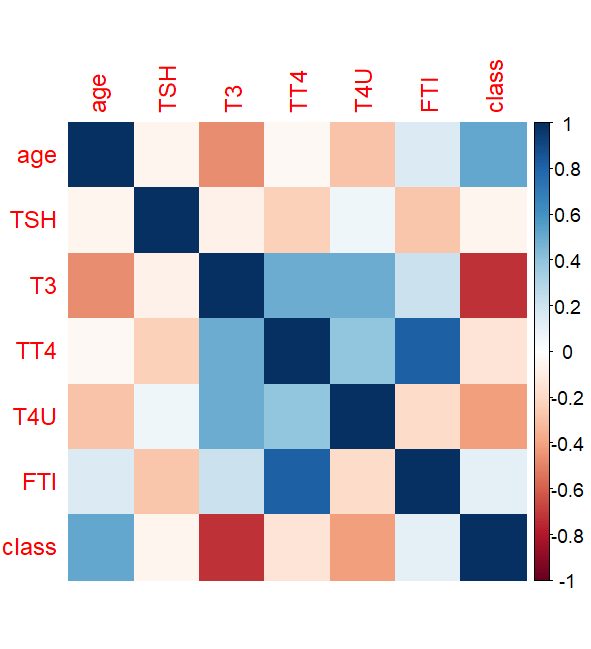
Mean-OvunSample



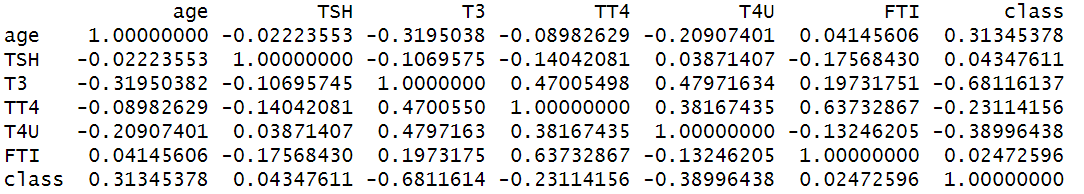


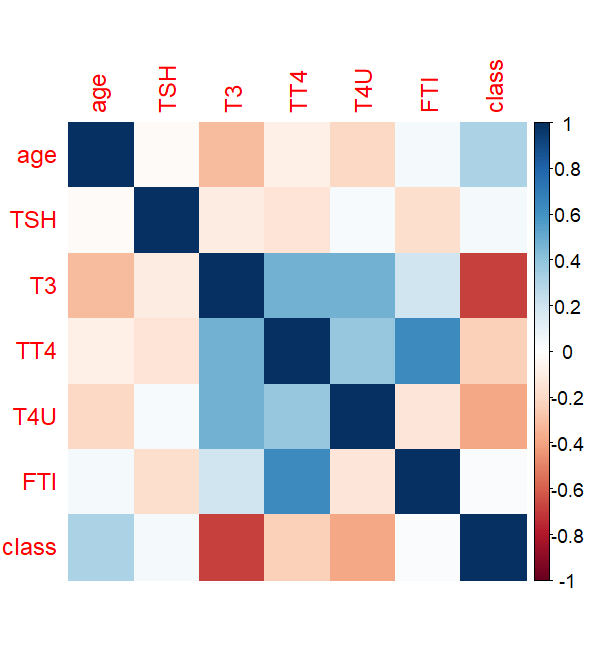
KNN-Smote



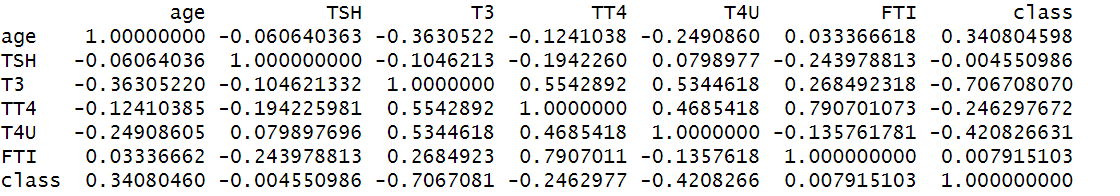


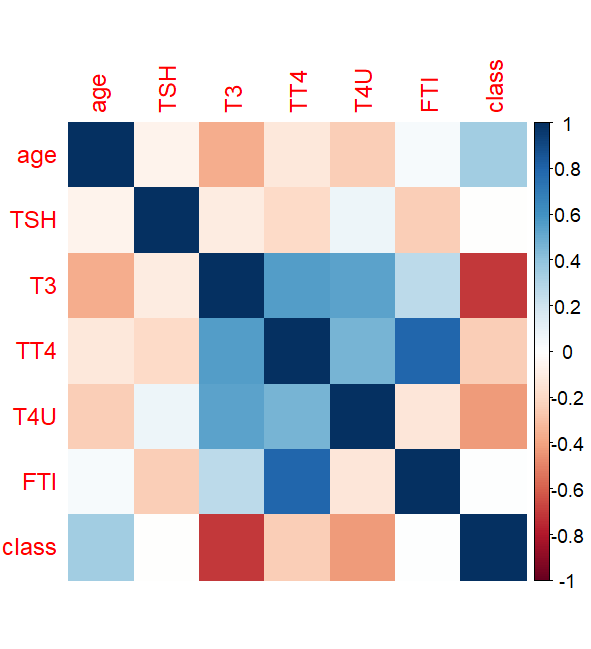
KNN-Rose



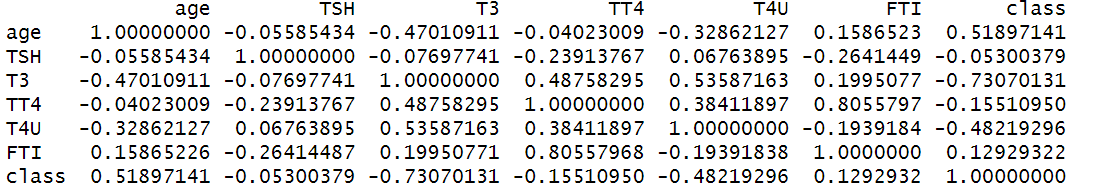


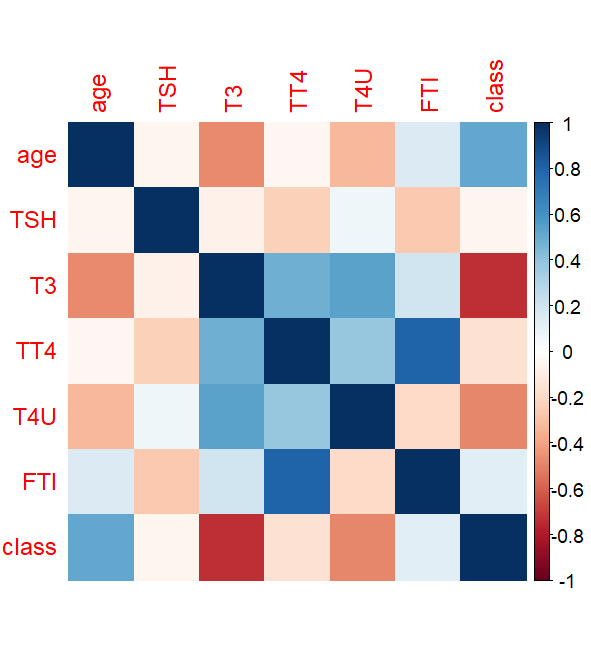
KNN-OvunSample



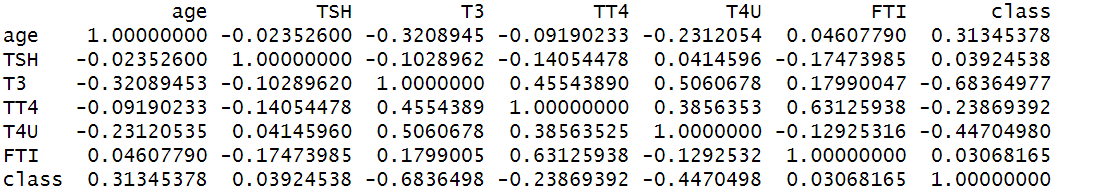


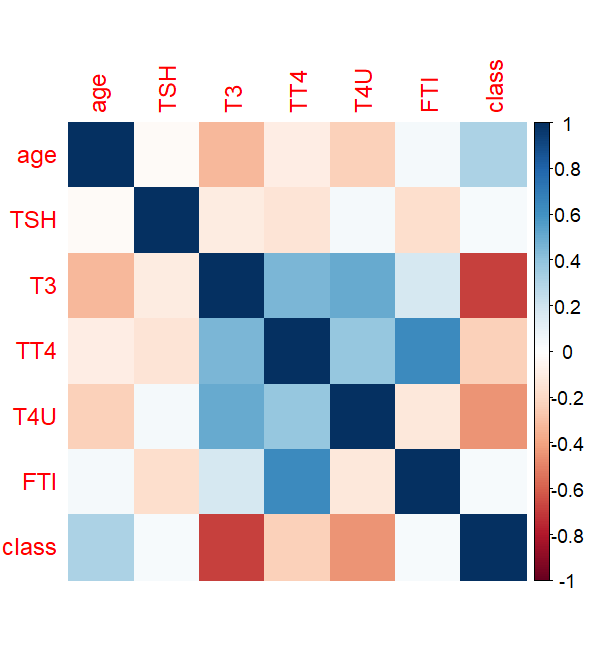
RF-Smote



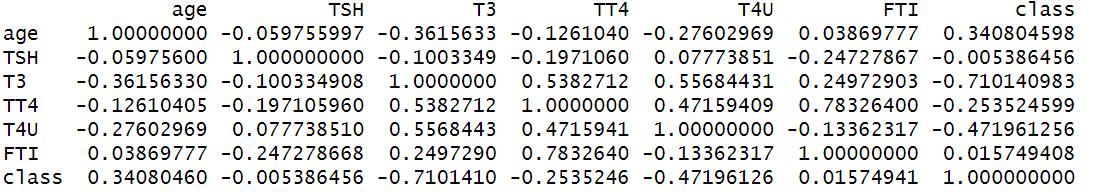


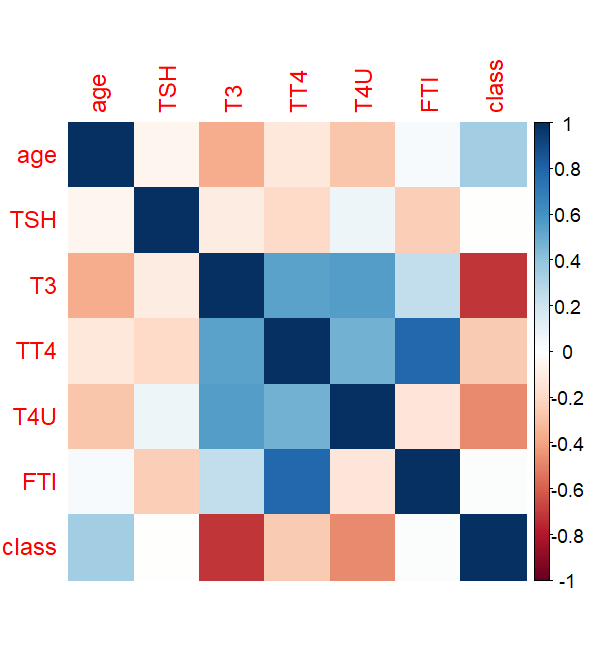
RF-Rose





RF-OvunSample





1. Recursive Feature Elimination

Mean-Smote

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables Accuracy Kappa AccuracySD KappaSD Selected

1 0.9705 0.9407 0.006052 0.012129

2 0.9569 0.9133 0.028366 0.057185

3 0.9686 0.9370 0.005778 0.011603

4 0.9757 0.9512 0.007723 0.015464

5 0.9765 0.9528 0.006179 0.012390

8 0.9793 0.9586 0.006080 0.012187

10 0.9809 0.9617 0.004784 0.009578 \*

12 0.9799 0.9596 0.004616 0.009247

22 0.9780 0.9559 0.006538 0.013125

The top 5 variables (out of 10):

T3, FTI, referral\_source, TT4, T4U

> predictors(result)

[1] "T3" "FTI" "referral\_source" "TT4"

[5] "T4U" "age" "TSH" "on\_thyroxine"

[9] "sex" "query\_hypothyroid"

Mean-Rose

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables Accuracy Kappa AccuracySD KappaSD Selected

1 0.8056 0.6111 0.024287 0.04857

2 0.8904 0.7807 0.019193 0.03840

3 0.9130 0.8260 0.008391 0.01677

4 0.9152 0.8304 0.011064 0.02213

5 0.9191 0.8382 0.011294 0.02259

8 0.9334 0.8669 0.012533 0.02507

10 0.9357 0.8713 0.011283 0.02256

12 0.9390 0.8779 0.011539 0.02308

22 0.9390 0.8779 0.011312 0.02263 \*

The top 5 variables (out of 22):

T3, referral\_source, FTI, TT4, T4U

Mean-OvunSample

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables Accuracy Kappa AccuracySD KappaSD Selected

1 0.9329 0.8658 0.014178 0.028304

2 0.9873 0.9746 0.005846 0.011699

3 0.9928 0.9856 0.003726 0.007453

4 0.9934 0.9867 0.006655 0.013314

5 0.9945 0.9890 0.004504 0.009011

8 0.9945 0.9890 0.005810 0.011624

10 0.9948 0.9895 0.005107 0.010218 \*

12 0.9945 0.9890 0.005810 0.011624

22 0.9934 0.9867 0.006261 0.012526

The top 5 variables (out of 10):

T3, FTI, TT4, TSH, referral\_source

KNN-Smote

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables Accuracy Kappa AccuracySD KappaSD Selected

1 0.9673 0.9344 0.005554 0.011114

2 0.9725 0.9449 0.008579 0.017193

3 0.9692 0.9381 0.003828 0.007732

4 0.9754 0.9507 0.005537 0.011079

5 0.9778 0.9554 0.002557 0.005102

8 0.9817 0.9633 0.005782 0.011590

10 0.9820 0.9638 0.004176 0.008365

12 0.9833 0.9664 0.005391 0.010810 \*

22 0.9804 0.9607 0.006874 0.013800

The top 5 variables (out of 12):

T3, FTI, referral\_source, TT4, T4U

KNN-Rose

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables Accuracy Kappa AccuracySD KappaSD Selected

1 0.8141 0.6282 0.027885 0.05575

2 0.8973 0.7945 0.019538 0.03908

3 0.9183 0.8365 0.013048 0.02608

4 0.9166 0.8332 0.011407 0.02280

5 0.9229 0.8459 0.010785 0.02156

8 0.9379 0.8757 0.011382 0.02275

10 0.9409 0.8818 0.011804 0.02360

12 0.9412 0.8824 0.010500 0.02099 \*

22 0.9403 0.8807 0.009765 0.01953

The top 5 variables (out of 12):

T3, referral\_source, FTI, TT4, T4U

KNN-OvunSample

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables Accuracy Kappa AccuracySD KappaSD Selected

1 0.9392 0.8785 0.014042 0.028042

2 0.9848 0.9696 0.007835 0.015677

3 0.9909 0.9818 0.004513 0.009029

4 0.9939 0.9878 0.005334 0.010670

5 0.9945 0.9889 0.004318 0.008639

8 0.9939 0.9878 0.005328 0.010659

10 0.9950 0.9901 0.004465 0.008933 \*

12 0.9948 0.9895 0.005107 0.010218

22 0.9942 0.9884 0.005585 0.011173

The top 5 variables (out of 10):

T3, FTI, TSH, TT4, referral\_source

RF-Smote

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables Accuracy Kappa AccuracySD KappaSD Selected

1 0.9608 0.9213 0.004941 0.009875

2 0.9564 0.9124 0.023788 0.047955

3 0.9712 0.9423 0.006049 0.012139

4 0.9762 0.9523 0.006916 0.013839

5 0.9754 0.9507 0.007413 0.014863

8 0.9796 0.9591 0.004577 0.009173

10 0.9801 0.9601 0.004118 0.008252

12 0.9807 0.9612 0.003726 0.007468 \*

22 0.9793 0.9586 0.005429 0.010902

The top 5 variables (out of 12):

T3, FTI, TT4, referral\_source, T4U

RF-Rose

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables Accuracy Kappa AccuracySD KappaSD Selected

1 0.8064 0.6128 0.02508 0.05016

2 0.8895 0.7790 0.02027 0.04055

3 0.9125 0.8249 0.01153 0.02304

4 0.9127 0.8255 0.01155 0.02309

5 0.9188 0.8376 0.01190 0.02380

8 0.9323 0.8647 0.01418 0.02835

10 0.9381 0.8763 0.01279 0.02557

12 0.9390 0.8779 0.01117 0.02233

22 0.9401 0.8802 0.01018 0.02036 \*

The top 5 variables (out of 22):

T3, referral\_source, FTI, T4U, TT4

RF-OvunSample

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables Accuracy Kappa AccuracySD KappaSD Selected

1 0.9384 0.8769 0.014589 0.029135

2 0.9879 0.9757 0.005843 0.011692

3 0.9917 0.9834 0.003678 0.007358

4 0.9931 0.9862 0.006272 0.012548

5 0.9942 0.9884 0.004404 0.008811

8 0.9950 0.9901 0.004465 0.008933 \*

10 0.9950 0.9901 0.004465 0.008933

12 0.9945 0.9890 0.005199 0.010402

22 0.9931 0.9862 0.006131 0.012266

The top 5 variables (out of 8):

T3, FTI, TT4, TSH, referral\_source

1. Chi-square test

statistic.X-squared p\_value

sex 28.2298195 1.077320e-07

on\_thyroxine 63.8237247 1.360651e-15

query\_on\_thyroxine 0.7591791 3.835855e-01

on\_antithyroid\_medication 18.4373141 1.755858e-05

sick 57.9794140 2.648752e-14

pregnant 32.8335359 1.003978e-08

thyroid\_surgery 20.4854458 6.008641e-06

I131\_treatment 7.7903356 5.252644e-03

query\_hypothyroid 25.6928160 4.003145e-07

query\_hyperthyroid 14.6170638 1.317165e-04

lithium 1.0451110 3.066358e-01

goitre 0.9765844 3.230435e-01

tumor 18.3822084 1.807377e-05

hypopituitary 5.0638385 2.443007e-02

psych 8.9959738 2.705750e-03

referral\_source 1257.9203762 4.419510e-271

Class 3617.0008824 0.000000e+00

**Performance measures:**

***With all features:***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Missing values imputed using | Classes balanced using | Algorithms applied | Accuracy | Precision | Recall | F1 score | Specificity |
| Mean | SMOTE | SVM | 0.958 | 0.966 | 0.944 | 0.955 | 0.97 |
| LR | 0.929 | 0.939 | 0.909 | 0.924 | 0.947 |
| RF | 0.973 | 0.967 | 0.976 | 0.971 | 0.97 |
| XGBoost | 0.984 | 0.983 | 0.983 | 0.983 | 0.985 |
| Naïve Bayes | 0.895 | 0.913 | 0.857 | 0.884 | 0.928 |
| Decision Tree | 0.934 | 0.939 | 0.919 | 0.929 | 0.947 |
| ROSE | SVM | 0.953 | 0.975 | 0.93 | 0.952 | 0.976 |
| LR | 0.903 | 0.918 | 0.886 | 0.902 | 0.92 |
| RF | 0.943 | 0.969 | 0.916 | 0.942 | 0.97 |
| XGBoost | 0.938 | 0.958 | 0.918 | 0.937 | 0.959 |
| Naïve Bayes | 0.891 | 0.912 | 0.868 | 0.889 | 0.915 |
| Decision Tree | 0.89 | 0.916 | 0.859 | 0.887 | 0.92 |
| Over & Under Sampling using ovun.sample() | SVM | 0.967 | 0.985 | 0.948 | 0.966 | 0.985 |
| LR | 0.906 | 0.924 | 0.883 | 0.903 | 0.929 |
| RF | 0.995 | 1 | 0.991 | 0.995 | 1 |
| XGBoost | 0.995 | 1 | 0.991 | 0.995 | 1 |
| Naïve Bayes | 0.881 | 0.905 | 0.85 | 0.877 | 0.912 |
| Decision Tree | 0.948 | 0.962 | 0.933 | 0.947 | 0.963 |
| KNN | SMOTE | SVM | 0.964 | 0.973 | 0.95 | 0.962 | 0.977 |
| LR | 0.935 | 0.953 | 0.907 | 0.93 | 0.96 |
| RF | 0.976 | 0.971 | 0.98 | 0.975 | 0.974 |
| XGBoost | 0.983 | 0.983 | 0.98 | 0.981 | 0.985 |
| Naïve Bayes | 0.896 | 0.92 | 0.854 | 0.886 | 0.934 |
| Decision Tree | 0.922 | 0.971 | 0.861 | 0.913 | 0.977 |
| ROSE | SVM | 0.953 | 0.975 | 0.93 | 0.952 | 0.976 |
| LR | 0.901 | 0.915 | 0.886 | 0.9 | 0.917 |
| RF | 0.941 | 0.969 | 0.912 | 0.94 | 0.97 |
| XGBoost | 0.94 | 0.965 | 0.914 | 0.939 | 0.967 |
| Naïve Bayes | 0.893 | 0.915 | 0.868 | 0.891 | 0.919 |
| Decision Tree | 0.898 | 0.97 | 0.822 | 0.89 | 0.974 |
| Over & Under Sampling using ovun.sample() | SVM | 0.973 | 0.985 | 0.961 | 0.973 | 0.985 |
| LR | 0.912 | 0.933 | 0.885 | 0.909 | 0.938 |
| RF | 0.995 | 1 | 0.991 | 0.995 | 1 |
| XGBoost | 0.993 | 1 | 0.985 | 0.993 | 1 |
| Naïve Bayes | 0.885 | 0.914 | 0.848 | 0.88 | 0.921 |
| Decision Tree | 0.952 | 0.971 | 0.931 | 0.951 | 0.973 |
| RF | SMOTE | SVM | 0.963 | 0.972 | 0.948 | 0.96 | 0.975 |
| LR | 0.934 | 0.944 | 0.913 | 0.928 | 0.952 |
| RF | 0.979 | 0.973 | 0.983 | 0.978 | 0.975 |
| XGBoost | 0.984 | 0.982 | 0.985 | 0.983 | 0.984 |
| Naïve Bayes | 0.897 | 0.914 | 0.863 | 0.888 | 0.928 |
| Decision Tree | 0.934 | 0.939 | 0.919 | 0.929 | 0.947 |
| ROSE | SVM | 0.953 | 0.975 | 0.93 | 0.952 | 0.976 |
| LR | 0.901 | 0.916 | 0.883 | 0.899 | 0.919 |
| RF | 0.941 | 0.969 | 0.912 | 0.94 | 0.97 |
| XGBoost | 0.939 | 0.963 | 0.914 | 0.938 | 0.965 |
| Naïve Bayes | 0.889 | 0.908 | 0.866 | 0.887 | 0.911 |
| Decision Tree | 0.89 | 0.916 | 0.859 | 0.887 | 0.92 |
| Over & Under Sampling using ovun.sample() | SVM | 0.969 | 0.985 | 0.952 | 0.968 | 0.985 |
| LR | 0.907 | 0.928 | 0.881 | 0.904 | 0.932 |
| RF | 0.995 | 1 | 0.991 | 0.995 | 1 |
| XGBoost | 0.994 | 1 | 0.987 | 0.993 | 1 |
| Naïve Bayes | 0.88 | 0.905 | 0.848 | 0.875 | 0.912 |
| Decision Tree | 0.948 | 0.962 | 0.933 | 0.947 | 0.963 |

1. ***With Mean Decrease in Gini:*** When this feature selection method is applied, it is observed that the 7 most important features “T3, referral source, T4U, TT4, age, FTI, TSH” are common to all types of imputed and balanced sets. Hence, the models are trained on these features and the performance is measured.

Mean-Smote

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.963 | 0.97 | 0.952 | 0.961 | 0.974 |
| LR | 0.929 | 0.935 | 0.913 | 0.924 | 0.944 |
| RF | 0.975 | 0.965 | 0.981 | 0.973 | 0.969 |
| XGBoost | 0.98 | 0.976 | 0.981 | 0.979 | 0.979 |
| Naïve Bayes | 0.895 | 0.909 | 0.865 | 0.886 | 0.923 |
| Decision Tree | 0.934 | 0.939 | 0.919 | 0.929 | 0.947 |

Mean-Rose

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.928 | 0.945 | 0.91 | 0.927 | 0.946 |
| LR | 0.893 | 0.906 | 0.879 | 0.892 | 0.907 |
| RF | 0.924 | 0.951 | 0.896 | 0.923 | 0.954 |
| XGBoost | 0.928 | 0.952 | 0.903 | 0.927 | 0.954 |
| Naïve Bayes | 0.892 | 0.907 | 0.875 | 0.891 | 0.909 |
| Decision Tree | 0.89 | 0.916 | 0.859 | 0.887 | 0.92 |

Mean- OvunSample

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.978 | 0.996 | 0.959 | 0.977 | 0.996 |
| LR | 0.907 | 0.924 | 0.885 | 0.904 | 0.929 |
| RF | 0.993 | 1 | 0.985 | 0.993 | 1 |
| XGBoost | 0.994 | 1 | 0.987 | 0.993 | 1 |
| Naïve Bayes | 0.881 | 0.897 | 0.859 | 0.878 | 0.903 |
| Decision Tree | 0.948 | 0.962 | 0.933 | 0.947 | 0.963 |

KNN-Smote

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.963 | 0.97 | 0.952 | 0.961 | 0.974 |
| LR | 0.937 | 0.945 | 0.92 | 0.932 | 0.952 |
| RF | 0.976 | 0.969 | 0.981 | 0.975 | 0.972 |
| XGBoost | 0.983 | 0.981 | 0.981 | 0.981 | 0.984 |
| Naïve Bayes | 0.9 | 0.916 | 0.867 | 0.891 | 0.929 |
| Decision Tree | 0.922 | 0.971 | 0.861 | 0.913 | 0.977 |

KNN-Rose

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.926 | 0.941 | 0.91 | 0.926 | 0.943 |
| LR | 0.895 | 0.915 | 0.872 | 0.893 | 0.919 |
| RF | 0.928 | 0.957 | 0.897 | 0.926 | 0.959 |
| XGBoost | 0.92 | 0.94 | 0.897 | 0.918 | 0.943 |
| Naïve Bayes | 0.893 | 0.909 | 0.875 | 0.892 | 0.911 |
| Decision Tree | 0.898 | 0.97 | 0.822 | 0.89 | 0.974 |

KNN-OvunSample

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.978 | 0.996 | 0.959 | 0.977 | 0.996 |
| LR | 0.907 | 0.931 | 0.878 | 0.904 | 0.936 |
| RF | 0.994 | 1 | 0.989 | 0.994 | 1 |
| XGBoost | 0.994 | 1 | 0.987 | 0.993 | 1 |
| Naïve Bayes | 0.89 | 0.912 | 0.861 | 0.885 | 0.918 |
| Decision Tree | 0.952 | 0.971 | 0.931 | 0.951 | 0.973 |

RF-Smote

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.962 | 0.97 | 0.948 | 0.959 | 0.974 |
| LR | 0.929 | 0.935 | 0.913 | 0.924 | 0.944 |
| RF | 0.973 | 0.962 | 0.981 | 0.972 | 0.965 |
| XGBoost | 0.983 | 0.98 | 0.983 | 0.982 | 0.982 |
| Naïve Bayes | 0.894 | 0.905 | 0.865 | 0.884 | 0.919 |
| Decision Tree | 0.934 | 0.939 | 0.919 | 0.929 | 0.947 |

RF-Rose

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.929 | 0.947 | 0.91 | 0.928 | 0.948 |
| LR | 0.893 | 0.906 | 0.879 | 0.892 | 0.907 |
| RF | 0.928 | 0.94 | 0.916 | 0.928 | 0.941 |
| XGBoost | 0.926 | 0.948 | 0.903 | 0.925 | 0.95 |
| Naïve Bayes | 0.893 | 0.909 | 0.875 | 0.892 | 0.911 |
| Decision Tree | 0.89 | 0.916 | 0.859 | 0.887 | 0.92 |

RF-OvunSample

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.978 | 0.996 | 0.959 | 0.977 | 0.996 |
| LR | 0.904 | 0.924 | 0.879 | 0.901 | 0.929 |
| RF | 0.993 | 1 | 0.985 | 0.993 | 1 |
| XGBoost | 0.993 | 1 | 0.985 | 0.993 | 1 |
| Naïve Bayes | 0.882 | 0.902 | 0.855 | 0.878 | 0.909 |
| Decision Tree | 0.948 | 0.962 | 0.933 | 0.947 | 0.963 |

The highest accuracy with this feature selection method is 99.4 of Mean-ovunsample of XGBoost

1. ***With Correlation matrix:***The features that are correlated were identified and then the models were trained removing one of the correlated features and features from the above 7 most important.

Mean-Smote

T3 T4U, TT4 FTI : Based on the Mean decrease gini scores, T3 & FTI selected

Feature set: "T3", "referral\_source", "age", "FTI", "TSH"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.956 | 0.959 | 0.948 | 0.953 | 0.964 |
| LR | 0.929 | 0.932 | 0.915 | 0.923 | 0.941 |
| RF | 0.969 | 0.96 | 0.976 | 0.968 | 0.964 |
| XGBoost | 0.976 | 0.972 | 0.978 | 0.975 | 0.975 |
| Naïve Bayes | 0.893 | 0.905 | 0.863 | 0.883 | 0.919 |
| Decision Tree | 0.934 | 0.939 | 0.919 | 0.929 | 0.947 |

Mean-Rose

TT4, FTI : Based on the Mean decrease gini scores, TT4 selected

Feature set: "T3", "referral\_source", "T4U", "TT4", "age", "TSH", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.919 | 0.927 | 0.91 | 0.919 | 0.928 |
| LR | 0.893 | 0.9 | 0.886 | 0.893 | 0.9 |
| RF | 0.927 | 0.941 | 0.912 | 0.927 | 0.943 |
| XGBoost | 0.923 | 0.934 | 0.91 | 0.922 | 0.935 |
| Naïve Bayes | 0.89 | 0.908 | 0.868 | 0.888 | 0.911 |
| Decision Tree | 0.893 | 0.906 | 0.879 | 0.892 | 0.907 |

Mean-OvunSample

T3, TT4 T3, T4U TT4,FTI: Based on the Mean decrease gini scores, T3 selected

Feature set: "T3", "referral\_source", "age", "TSH", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.939 | 0.922 | 0.959 | 0.94 | 0.92 |
| LR | 0.9 | 0.913 | 0.881 | 0.897 | 0.918 |
| RF | 0.991 | 0.996 | 0.985 | 0.991 | 0.996 |
| XGBoost | 0.99 | 0.996 | 0.983 | 0.99 | 0.996 |
| Naïve Bayes | 0.885 | 0.901 | 0.863 | 0.882 | 0.907 |
| Decision Tree | 0.946 | 0.914 | 0.983 | 0.947 | 0.909 |

KNN-Smote

TT4,FTI: Based on the Mean decrease gini scores, FTI selected

Feature set: "T3", "referral\_source", "T4U", "age", "FTI", "TSH", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.964 | 0.973 | 0.95 | 0.962 | 0.977 |
| LR | 0.939 | 0.947 | 0.922 | 0.934 | 0.954 |
| RF | 0.974 | 0.967 | 0.978 | 0.972 | 0.97 |
| XGBoost | 0.978 | 0.978 | 0.976 | 0.977 | 0.98 |
| Naïve Bayes | 0.902 | 0.923 | 0.865 | 0.893 | 0.936 |
| Decision Tree | 0.922 | 0.971 | 0.861 | 0.913 | 0.977 |

KNN-Rose

TT4,FTI: Based on the Mean decrease gini scores, TT4 selected

Feature set: "T3", "referral\_source", "T4U", "TT4", "age", "TSH", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.924 | 0.933 | 0.916 | 0.924 | 0.933 |
| LR | 0.895 | 0.915 | 0.872 | 0.893 | 0.919 |
| RF | 0.921 | 0.941 | 0.899 | 0.919 | 0.943 |
| XGBoost | 0.921 | 0.941 | 0.899 | 0.919 | 0.943 |
| Naïve Bayes | 0.893 | 0.913 | 0.87 | 0.891 | 0.917 |
| Decision Tree | 0.901 | 0.901 | 0.901 | 0.901 | 0.9 |

KNN-OvunSample

T3 TT4, T3, T4U TT4, FTI: Based on the Mean decrease gini scores, T3 selected

Feature set: "T3", "referral\_source", "age", "TSH", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.944 | 0.93 | 0.959 | 0.944 | 0.929 |
| LR | 0.915 | 0.932 | 0.894 | 0.913 | 0.936 |
| RF | 0.994 | 0.996 | 0.991 | 0.993 | 0.996 |
| XGBoost | 0.993 | 0.996 | 0.989 | 0.993 | 0.996 |
| Naïve Bayes | 0.884 | 0.904 | 0.857 | 0.88 | 0.91 |
| Decision Tree | 0.95 | 0.922 | 0.983 | 0.952 | 0.918 |

RF-Smote

T3, T4U TT4, FTI: Based on the Mean decrease gini scores, T3, FTI selected

Feature set: "T3", "referral\_source", "age", "FTI", "TSH", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.956 | 0.962 | 0.944 | 0.953 | 0.967 |
| LR | 0.931 | 0.936 | 0.917 | 0.926 | 0.944 |
| RF | 0.973 | 0.965 | 0.978 | 0.971 | 0.969 |
| XGBoost | 0.977 | 0.972 | 0.98 | 0.976 | 0.975 |
| Naïve Bayes | 0.891 | 0.904 | 0.859 | 0.881 | 0.919 |
| Decision Tree | 0.934 | 0.939 | 0.919 | 0.929 | 0.947 |

RF-Rose

T3, T4U TT4, FTI: Based on the Mean decrease gini scores, T3, TT4 selected

Feature set: "T3", "referral\_source", "TT4", "age", "TSH", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.912 | 0.923 | 0.899 | 0.911 | 0.924 |
| LR | 0.89 | 0.907 | 0.872 | 0.889 | 0.909 |
| RF | 0.913 | 0.926 | 0.899 | 0.913 | 0.928 |
| XGBoost | 0.914 | 0.928 | 0.899 | 0.913 | 0.93 |
| Naïve Bayes | 0.879 | 0.9 | 0.855 | 0.877 | 0.904 |
| Decision Tree | 0.893 | 0.906 | 0.879 | 0.892 | 0.907 |

RF-OvunSample

T3, TT4 T3, T4U TT4, FTI: Based on the Mean decrease gini scores, T3 selected

Feature set: "T3", "referral\_source", "age", "TSH", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.939 | 0.922 | 0.959 | 0.94 | 0.92 |
| LR | 0.903 | 0.919 | 0.883 | 0.901 | 0.923 |
| RF | 0.992 | 0.996 | 0.987 | 0.992 | 0.996 |
| XGBoost | 0.988 | 0.996 | 0.98 | 0.988 | 0.996 |
| Naïve Bayes | 0.883 | 0.901 | 0.859 | 0.879 | 0.907 |
| Decision Tree | 0.954 | 0.957 | 0.95 | 0.953 | 0.958 |

The highest accuracy with this feature selection method is 99.4 of KNN-ovunsample of RF model.

1. ***With Recursive Feature Elimination:***

Mean-Smote: "T3", "FTI", "referral\_source", "TT4", "T4U", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.962 | 0.961 | 0.957 | 0.959 | 0.965 |
| LR | 0.936 | 0.941 | 0.922 | 0.932 | 0.949 |
| RF | 0.974 | 0.962 | 0.983 | 0.973 | 0.965 |
| XGBoost | 0.974 | 0.96 | 0.985 | 0.973 | 0.964 |
| Naïve Bayes | 0.903 | 0.91 | 0.881 | 0.896 | 0.923 |
| Decision Tree | 0.934 | 0.939 | 0.919 | 0.929 | 0.947 |

Mean-Rose: "T3", "referral\_source", "FTI", "TT4", "T4U", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.92 | 0.946 | 0.892 | 0.918 | 0.948 |
| LR | 0.892 | 0.902 | 0.881 | 0.892 | 0.904 |
| RF | 0.922 | 0.948 | 0.894 | 0.92 | 0.95 |
| XGBoost | 0.922 | 0.946 | 0.896 | 0.92 | 0.948 |
| Naïve Bayes | 0.885 | 0.898 | 0.87 | 0.884 | 0.9 |
| Decision Tree | 0.89 | 0.916 | 0.859 | 0.887 | 0.92 |

Mean-OvunSample: "T3", "FTI", "TT4", "TSH", "referral\_source", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.982 | 0.996 | 0.967 | 0.981 | 0.996 |
| LR | 0.905 | 0.921 | 0.885 | 0.903 | 0.925 |
| RF | 0.992 | 1 | 0.983 | 0.992 | 1 |
| XGBoost | 0.993 | 1 | 0.985 | 0.993 | 1 |
| Naïve Bayes | 0.899 | 0.921 | 0.87 | 0.895 | 0.927 |
| Decision Tree | 0.948 | 0.962 | 0.933 | 0.947 | 0.963 |

KNN-Smote: "T3", "FTI", "referral\_source", "TT4", "T4U", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.966 | 0.966 | 0.961 | 0.964 | 0.97 |
| LR | 0.941 | 0.95 | 0.922 | 0.936 | 0.957 |
| RF | 0.968 | 0.955 | 0.978 | 0.966 | 0.959 |
| XGBoost | 0.977 | 0.972 | 0.98 | 0.976 | 0.975 |
| Naïve Bayes | 0.908 | 0.914 | 0.887 | 0.9 | 0.926 |
| Decision Tree | 0.922 | 0.971 | 0.861 | 0.913 | 0.977 |

KNN-Rose: "T3", "referral\_source", "FTI", "TT4", "T4U", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.925 | 0.953 | 0.896 | 0.924 | 0.956 |
| LR | 0.895 | 0.915 | 0.872 | 0.893 | 0.919 |
| RF | 0.922 | 0.951 | 0.89 | 0.92 | 0.954 |
| XGBoost | 0.924 | 0.955 | 0.89 | 0.921 | 0.957 |
| Naïve Bayes | 0.89 | 0.905 | 0.874 | 0.889 | 0.907 |
| Decision Tree | 0.898 | 0.97 | 0.822 | 0.89 | 0.974 |

KNN-OvunSample: "T3", "FTI", "TSH", "TT4", "referral\_source", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.976 | 0.99 | 0.961 | 0.976 | 0.991 |
| LR | 0.908 | 0.928 | 0.883 | 0.905 | 0.932 |
| RF | 0.994 | 1 | 0.989 | 0.994 | 1 |
| XGBoost | 0.992 | 1 | 0.983 | 0.992 | 1 |
| Naïve Bayes | 0.899 | 0.923 | 0.868 | 0.895 | 0.929 |
| Decision Tree | 0.952 | 0.971 | 0.931 | 0.951 | 0.973 |

RF-Smote: "T3", "FTI", "TT4", "referral\_source", "T4U", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.961 | 0.963 | 0.954 | 0.958 | 0.967 |
| LR | 0.938 | 0.945 | 0.922 | 0.933 | 0.952 |
| RF | 0.969 | 0.955 | 0.98 | 0.967 | 0.959 |
| XGBoost | 0.979 | 0.971 | 0.985 | 0.978 | 0.974 |
| Naïve Bayes | 0.905 | 0.918 | 0.876 | 0.897 | 0.931 |
| Decision Tree | 0.934 | 0.939 | 0.919 | 0.929 | 0.947 |

RF-Rose: "T3", "referral\_source", "FTI", "T4U", "TT4", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.922 | 0.941 | 0.901 | 0.92 | 0.943 |
| LR | 0.894 | 0.906 | 0.881 | 0.893 | 0.907 |
| RF | 0.92 | 0.942 | 0.896 | 0.918 | 0.944 |
| XGBoost | 0.922 | 0.939 | 0.903 | 0.921 | 0.941 |
| Naïve Bayes | 0.887 | 0.901 | 0.87 | 0.885 | 0.904 |
| Decision Tree | 0.89 | 0.916 | 0.859 | 0.887 | 0.92 |

RF-OvunSample: "T3", "FTI", "TT4", "TSH", "referral\_source", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.983 | 0.996 | 0.968 | 0.982 | 0.996 |
| LR | 0.904 | 0.921 | 0.883 | 0.902 | 0.925 |
| RF | 0.991 | 1 | 0.981 | 0.991 | 1 |
| XGBoost | 0.993 | 1 | 0.985 | 0.993 | 1 |
| Naïve Bayes | 0.895 | 0.917 | 0.866 | 0.891 | 0.923 |
| Decision Tree | 0.948 | 0.962 | 0.933 | 0.947 | 0.963 |

The highest accuracy with this feature selection method is 99.4 of KNN-ovunsample of RF model.

For all feature selection methods, the highest accuracy is 99.4%.

For achieving this highest accuracy,

***Mean Decrease in Gini*** uses 7 features.

***Correlation matrix*** uses 4 features.

***Recursive Feature Elimination*** uses 5 features.

The model which gives the highest accuracy with minimum no. of features is chosen.

Hence, this is chosen for our Shiny app.

***Correlation matrix:*** KNN-OvunSample

T3 TT4, T3, T4U TT4, FTI: Based on the Mean decrease gini scores, T3 selected

Feature set: "T3", "referral\_source", "age", "TSH", "Class"

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.944 | 0.93 | 0.959 | 0.944 | 0.929 |
| LR | 0.915 | 0.932 | 0.894 | 0.913 | 0.936 |
| RF | 0.994 | 0.996 | 0.991 | 0.993 | 0.996 |
| XGBoost | 0.993 | 0.996 | 0.989 | 0.993 | 0.996 |
| Naïve Bayes | 0.884 | 0.904 | 0.857 | 0.88 | 0.91 |
| Decision Tree | 0.95 | 0.922 | 0.983 | 0.952 | 0.918 |

