Applying rfimpute() on dataset with all features.

RF-Smote

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.963 | 0.976 | 0.946 | 0.961 | 0.979 |
| LR | 0.949 | 0.966 | 0.924 | 0.945 | 0.971 |
| RF | 0.973 | 0.971 | 0.971 | 0.971 | 0.974 |
| XGBoost | 0.982 | 0.985 | 0.977 | 0.981 | 0.987 |
| Naïve Bayes | 0.896 | 0.932 | 0.841 | 0.884 | 0.945 |
| Decision Tree | 0.912 | 0.939 | 0.868 | 0.902 | 0.95 |

RF-Rose

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.944 | 0.956 | 0.931 | 0.943 | 0.958 |
| LR | 0.899 | 0.905 | 0.892 | 0.898 | 0.906 |
| RF | 0.935 | 0.947 | 0.922 | 0.935 | 0.949 |
| XGBoost | 0.926 | 0.933 | 0.917 | 0.925 | 0.935 |
| Naïve Bayes | 0.876 | 0.905 | 0.841 | 0.872 | 0.912 |
| Decision Tree | 0.899 | 0.909 | 0.887 | 0.898 | 0.912 |

RF-ovunsample

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.978 | 0.987 | 0.968 | 0.977 | 0.988 |
| LR | 0.904 | 0.911 | 0.895 | 0.903 | 0.914 |
| RF | 0.997 | 1 | 0.995 | 0.997 | 1 |
| XGBoost | 0.996 | 1 | 0.993 | 0.996 | 1 |
| Naïve Bayes | 0.867 | 0.902 | 0.821 | 0.86 | 0.912 |
| Decision Tree | 0.951 | 0.958 | 0.943 | 0.95 | 0.96 |

After applying feature selection methods:

1. Mean Decrease in Gini: selecting top 7

Column\_Number MeanDecreaseGini

T3 18 905.3788255

referral\_source 22 229.7125397

T4U 20 172.5096236

TT4 19 151.2513202

FTI 21 140.2729948

age 1 87.0798418

TSH 17 75.3694556

on\_thyroxine 3 15.0951106

query\_hypothyroid 10 9.8894296

sex 2 9.4844553

sick 6 7.1091928

hypopituitary 15 3.3726799

query\_hyperthyroid 11 2.7832005

query\_on\_thyroxine 4 2.2251817

psych 16 2.0134446

goitre 13 1.5627746

thyroid\_surgery 8 1.1344663

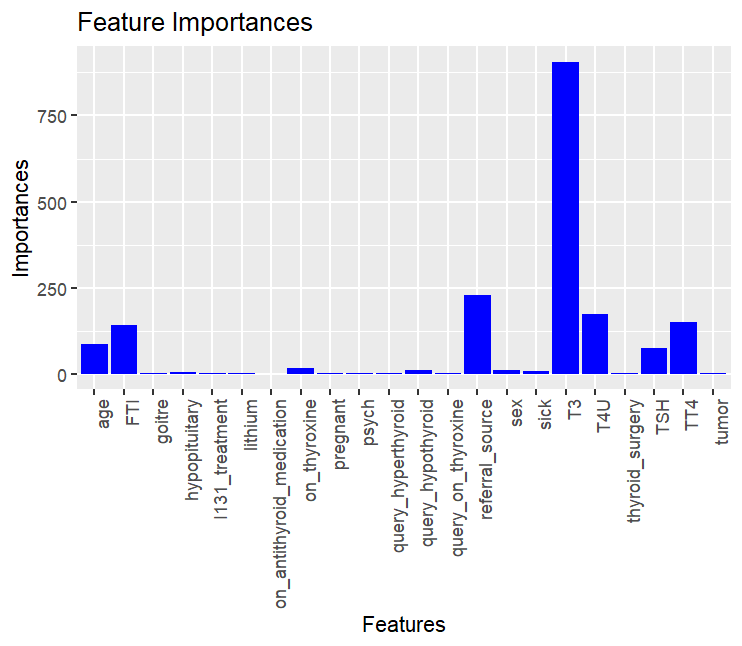
tumor 14 1.1259908

I131\_treatment 9 0.9305384

pregnant 7 0.5334638

lithium 12 0.5268633

on\_antithyroid\_medication 5 0.3303528



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.98 | 0.987 | 0.971 | 0.979 | 0.988 |
| LR | 0.898 | 0.902 | 0.891 | 0.897 | 0.905 |
| RF | 0.996 | 1 | 0.991 | 0.996 | 1 |
| XGBoost | 0.995 | 1 | 0.989 | 0.995 | 1 |
| Naïve Bayes | 0.87 | 0.898 | 0.832 | 0.864 | 0.907 |
| Decision Tree | 0.951 | 0.958 | 0.943 | 0.95 | 0.96 |

1. Recursive feature elimination:

Recursive feature selection

Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

Variables Accuracy Kappa AccuracySD KappaSD Selected

1 0.9327 0.8654 0.012409 0.024796

2 0.9897 0.9793 0.004903 0.009812

3 0.9910 0.9820 0.004186 0.008375

4 0.9931 0.9862 0.004369 0.008740

5 0.9942 0.9883 0.004811 0.009625

8 0.9955 0.9910 0.003965 0.007931 \*

10 0.9952 0.9905 0.003915 0.007831

12 0.9955 0.9910 0.003965 0.007931

22 0.9947 0.9894 0.004679 0.009361

The top 5 variables (out of 8):

T3, FTI, TT4, TSH, referral\_source

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.982 | 0.987 | 0.977 | 0.982 | 0.988 |
| LR | 0.901 | 0.91 | 0.887 | 0.899 | 0.914 |
| RF | 0.994 | 1 | 0.988 | 0.994 | 1 |
| XGBoost | 0.994 | 1 | 0.988 | 0.994 | 1 |
| Naïve Bayes | 0.876 | 0.904 | 0.839 | 0.87 | 0.912 |
| Decision Tree | 0.951 | 0.958 | 0.943 | 0.95 | 0.96 |

1. Correlation matrix: T3 TT4 T3 T4U TT4 FTI **T3, referral\_source, age, TSH**

age TSH T3 TT4 T4U FTI class

age 1.00000000 -0.06242313 -0.34292739 -0.1143843 -0.25235259 0.02398143 0.31253969

TSH -0.06242313 1.00000000 -0.06760509 -0.1867541 0.09502519 -0.22973475 -0.01785488

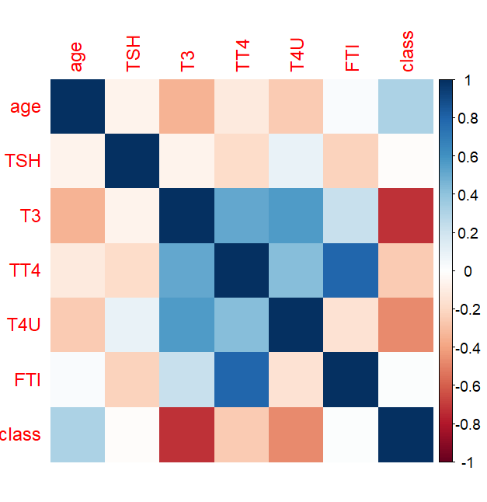
T3 -0.34292739 -0.06760509 1.00000000 0.5116293 0.56923181 0.22292357 -0.72097397

TT4 -0.11438434 -0.18675406 0.51162925 1.0000000 0.42420110 0.79747121 -0.25001253

T4U -0.25235259 0.09502519 0.56923181 0.4242011 1.00000000 -0.15863493 -0.47867318

FTI 0.02398143 -0.22973475 0.22292357 0.7974712 -0.15863493 1.00000000 0.01189464

class 0.31253969 -0.01785488 -0.72097397 -0.2500125 -0.47867318 0.01189464 1.00000000



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.94 | 0.921 | 0.961 | 0.941 | 0.919 |
| LR | 0.898 | 0.904 | 0.889 | 0.896 | 0.907 |
| RF | 0.995 | 1 | 0.989 | 0.995 | 1 |
| XGBoost | 0.994 | 1 | 0.988 | 0.994 | 1 |
| Naïve Bayes | 0.871 | 0.9 | 0.832 | 0.865 | 0.909 |
| Decision Tree | 0.939 | 0.928 | 0.95 | 0.939 | 0.928 |