**Experimentation done:**

**Performance measures:**

***With all features:***

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Missing values imputed using | Classes balanced using | Algorithms applied | Accuracy | Precision | Recall | F1 score | Specificity |
| Mean | SMOTE | SVM | 0.924 | 0.89 | 0.924 | 0.906 | 0.924 |
| LR | 0.908 | 0.871 | 0.904 | 0.887 | 0.911 |
| RF | 0.944 | 0.914 | 0.949 | 0.931 | 0.941 |
| XGBoost | 0.959 | 0.949 | 0.949 | 0.949 | 0.966 |
| Naïve Bayes | 0.878 | 0.83 | 0.873 | 0.851 | 0.881 |
| Decision Tree | 0.911 | 0.855 | 0.936 | 0.894 | 0.894 |
| ROSE | SVM | 0.928 | 0.939 | 0.924 | 0.932 | 0.933 |
| LR | 0.908 | 0.913 | 0.913 | 0.913 | 0.903 |
| RF | 0.92 | 0.943 | 0.902 | 0.922 | 0.939 |
| XGBoost | 0.934 | 0.945 | 0.929 | 0.937 | 0.939 |
| Naïve Bayes | 0.874 | 0.885 | 0.875 | 0.88 | 0.873 |
| Decision Tree | 0.894 | 0.951 | 0.842 | 0.893 | 0.952 |
| Over & Under Sampling using ovun.sample() | SVM | 0.888 | 0.883 | 0.888 | 0.886 | 0.888 |
| LR | 0.894 | 0.903 | 0.876 | 0.89 | 0.911 |
| RF | 0.974 | 0.982 | 0.965 | 0.973 | 0.983 |
| XGBoost | 0.968 | 0.965 | 0.971 | 0.968 | 0.966 |
| Naïve Bayes | 0.877 | 0.885 | 0.859 | 0.872 | 0.894 |
| Decision Tree | 0.954 | 0.994 | 0.912 | 0.951 | 0.994 |
| KNN | SMOTE | SVM | 0.891 | 0.839 | 0.898 | 0.868 | 0.886 |
| LR | 0.934 | 0.928 | 0.904 | 0.916 | 0.953 |
| RF | 0.969 | 0.95 | 0.975 | 0.962 | 0.966 |
| XGBoost | 0.975 | 0.957 | 0.981 | 0.969 | 0.97 |
| Naïve Bayes | 0.901 | 0.893 | 0.854 | 0.873 | 0.932 |
| Decision Tree | 0.929 | 0.945 | 0.873 | 0.907 | 0.966 |
| ROSE | SVM | 0.923 | 0.934 | 0.918 | 0.926 | 0.927 |
| LR | 0.908 | 0.913 | 0.913 | 0.913 | 0.903 |
| RF | 0.931 | 0.955 | 0.913 | 0.933 | 0.952 |
| XGBoost | 0.928 | 0.954 | 0.908 | 0.93 | 0.952 |
| Naïve Bayes | 0.88 | 0.89 | 0.88 | 0.885 | 0.879 |
| Decision Tree | 0.897 | 0.951 | 0.848 | 0.897 | 0.952 |
| Over & Under Sampling using ovun.sample() | SVM | 0.814 | 0.85 | 0.768 | 0.807 | 0.860 |
| LR | 0.905 | 0.934 | 0.876 | 0.904 | 0.936 |
| RF | 0.966 | 1 | 0.932 | 0.965 | 1 |
| XGBoost | 0.96 | 0.988 | 0.932 | 0.959 | 0.988 |
| Naïve Bayes | 0.854 | 0.894 | 0.808 | 0.849 | 0.901 |
| Decision Tree | 0.943 | 0.97 | 0.915 | 0.942 | 0.971 |
| RF | SMOTE | SVM | 0.95 | 0.955 | 0.919 | 0.937 | 0.971 |
| LR | 0.945 | 0.943 | 0.919 | 0.931 | 0.963 |
| RF | 0.97 | 0.975 | 0.95 | 0.962 | 0.983 |
| XGBoost | 0.968 | 0.951 | 0.969 | 0.96 | 0.967 |
| Naïve Bayes | 0.901 | 0.923 | 0.82 | 0.868 | 0.955 |
| Decision Tree | 0.935 | 0.953 | 0.882 | 0.916 | 0.971 |
| ROSE | SVM | 0.908 | 0.947 | 0.875 | 0.91 | 0.945 |
| LR | 0.848 | 0.874 | 0.832 | 0.852 | 0.867 |
| RF | 0.897 | 0.94 | 0.859 | 0.898 | 0.939 |
| XGBoost | 0.897 | 0.92 | 0.88 | 0.9 | 0.915 |
| Naïve Bayes | 0.817 | 0.87 | 0.766 | 0.815 | 0.873 |
| Decision Tree | 0.874 | 0.898 | 0.859 | 0.878 | 0.891 |
| Over & Under Sampling using ovun.sample() | SVM | 0.951 | 0.965 | 0.938 | 0.952 | 0.965 |
| LR | 0.926 | 0.937 | 0.916 | 0.926 | 0.936 |
| RF | 0.986 | 0.994 | 0.978 | 0.986 | 0.994 |
| XGBoost | 0.986 | 0.994 | 0.978 | 0.986 | 0.994 |
| Naïve Bayes | 0.903 | 0.919 | 0.888 | 0.903 | 0.918 |
| Decision Tree | 0.986 | 0.994 | 0.978 | 0.986 | 0.994 |

The NA values after being imputed by RF and the classes after balancing by Over & under sampling, the highest accuracy is achieved with Random Forest and XGBoost = 98.6%

After applying feature selection techniques:

Mean Decrease in Gini

Column\_Number MeanDecreaseGini

T3 18 233.99682551

T4U 20 51.63309127

referral\_source 22 51.11592465

TT4 19 37.35201536

age 1 35.68528263

FTI 21 32.67990820

TSH 17 20.40800860

query\_hypothyroid 10 3.27433674

on\_thyroxine 3 2.94722776

sex 2 2.25570211

sick 6 1.58941511

query\_hyperthyroid 11 1.33128019

query\_on\_thyroxine 4 0.99591384

hypopituitary 15 0.79384091

thyroid\_surgery 8 0.48770653

I131\_treatment 9 0.41859711

psych 16 0.39607795

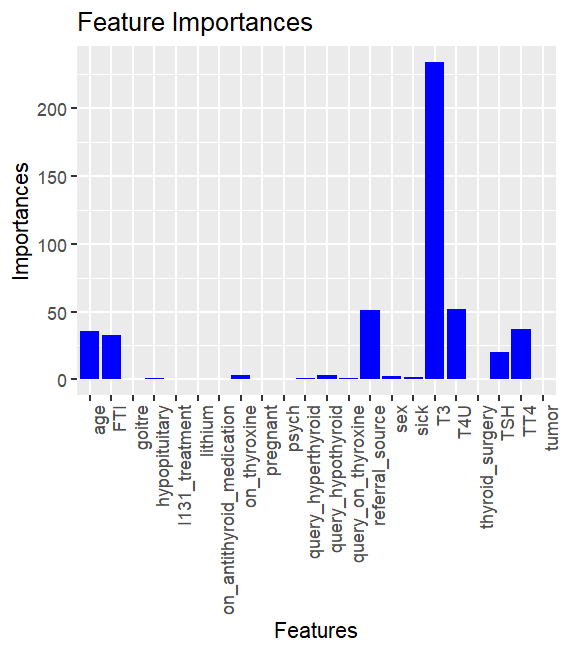
goitre 13 0.38869855

tumor 14 0.34636455

lithium 12 0.20376768

on\_antithyroid\_medication 5 0.12038057

pregnant 7 0.03011414



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.974 | 0.988 | 0.961 | 0.974 | 0.988 |
| LR | 0.923 | 0.936 | 0.91 | 0.923 | 0.936 |
| RF | 0.986 | 0.994 | 0.978 | 0.986 | 0.994 |
| XGBoost | 0.986 | 0.994 | 0.978 | 0.986 | 0.994 |
| Naïve Bayes | 0.897 | 0.908 | 0.888 | 0.898 | 0.906 |
| Decision Tree | 0.986 | 0.994 | 0.978 | 0.986 | 0.994 |

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 0.940 0.8799 0.020485 0.04112

2 0.983 0.9660 0.008215 0.01642

3 0.986 0.9720 0.009602 0.01921

4 0.985 0.9700 0.007048 0.01409

5 0.985 0.9700 0.008479 0.01696

8 0.987 0.9740 0.011612 0.02324

10 0.990 0.9800 0.009428 0.01886 \*

12 0.987 0.9740 0.011612 0.02324

22 0.987 0.9740 0.011612 0.02324

The top 5 variables (out of 10):

T3, FTI, TT4, T4U, TSH

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.974 | 0.988 | 0.961 | 0.974 | 0.988 |
| LR | 0.923 | 0.936 | 0.91 | 0.923 | 0.936 |
| RF | 0.983 | 0.994 | 0.972 | 0.983 | 0.994 |
| XGBoost | 0.986 | 0.989 | 0.983 | 0.986 | 0.988 |
| Naïve Bayes | 0.871 | 0.876 | 0.871 | 0.873 | 0.871 |
| Decision Tree | 0.986 | 0.989 | 0.983 | 0.986 | 0.988 |

Correlation Matrix: T3 TT4, T3 T4U, TT4 FTI T3, FTI, referral\_source, age, TSH

age TSH T3 TT4 T4U FTI class

age 1.00000000 -0.01674288 -0.4042367 -0.1242969 -0.26677254 0.01685703 0.40285546

TSH -0.01674288 1.00000000 -0.1685945 -0.2576854 0.07747266 -0.29540006 0.03101319

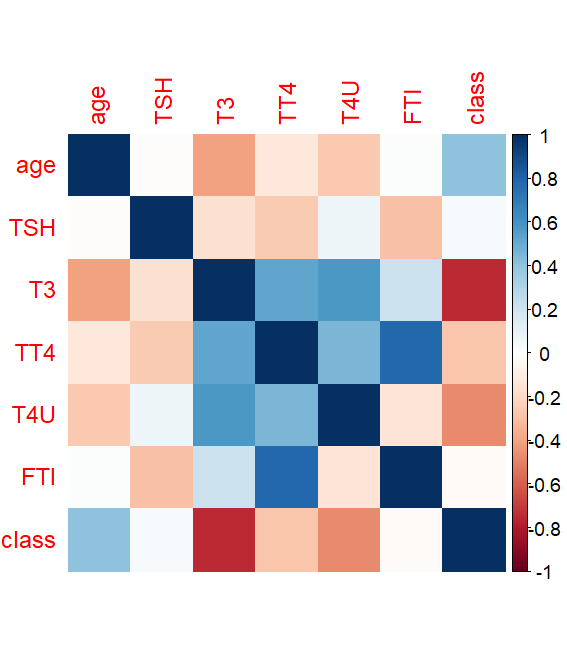
T3 -0.40423673 -0.16859451 1.0000000 0.5211614 0.57887305 0.21201338 -0.74036591

TT4 -0.12429693 -0.25768544 0.5211614 1.0000000 0.45667452 0.78586262 -0.27642204

T4U -0.26677254 0.07747266 0.5788730 0.4566745 1.00000000 -0.14082705 -0.47859514

FTI 0.01685703 -0.29540006 0.2120134 0.7858626 -0.14082705 1.00000000 -0.02536916

class 0.40285546 0.03101319 -0.7403659 -0.2764220 -0.47859514 -0.02536916 1.00000000



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Algorithm | Accuracy | Precision | Recall | F1 score | Specificity |
| SVM | 0.963 | 0.961 | 0.966 | 0.964 | 0.959 |
| LR | 0.923 | 0.936 | 0.91 | 0.923 | 0.936 |
| RF | 0.98 | 0.983 | 0.978 | 0.98 | 0.982 |
| XGBoost | 0.986 | 0.983 | 0.989 | 0.986 | 0.982 |
| Naïve Bayes | 0.914 | 0.92 | 0.91 | 0.915 | 0.918 |
| Decision Tree | 0.986 | 0.989 | 0.983 | 0.986 | 0.988 |

**Results:**

XGBoost and Decision Tree perform the best with 5 features selected using RFE and Correlation Matrix. Hence, we have created the Shiny App with XGBoost model to predict whether the patient has hypothyroid or not.