# PH125.9x Data Science Capstone: Differentiated Thyroid Cancer Recurrence

Krithika Ganeshkumar

2023-12-13

## Contents

# 1 Executive Summary

Thyroid is a small butterfly-shaped gland, part of the endocrine system located at the base of the neck. The hormones produced by the gland affects every cell in the body. It is essential to regulate the rate at which the body uses fats and carbohydrates. Thyroid cancer is a type of cancer that occurs in the thyroid gland. This type of cancer mostly cannot be seen or felt. Even though the mortality due to thyroid cancer remains low, the risk of recurrence is high. It warrants frequent follow ups and management.

The primary goal of this project is to train machine learning models such as naive Bayes, random forest, support vector machine, k-nearest neighbors, generalized linear model and gradient boosting machine to predict the likelihood of thyroid cancer recurrence in patients diagnosed with differentiated thyroid cancer. The dataset is available at UCI. This dataset is analyzed by utilizing the various tools learned throughout the courses in the PH125.x data science series. The objective is to determine a model that yields the highest score for recall/sensitivity and AUC (Area Under the Curve) among the trained models. This dataset is part of research in the field of AI and Medicine and does not contain any sensitive data. It does not have any missing (NA) values.

In this report, several key steps are followed to achieve the objective by exploratory analysis of data including checking for NAs, proportion of target variable, distribution of feature variables through data visualization, training and validating the models using the dataset provided. All models are trained on the train set which is 80% of the dataset. The validation set is derived from 20% of the dataset. Repeated cross-validation method is employed to estimate the performance of the model. Confusion matrix results are analyzed at each step and eventually, the model with the best metrics for predicting the recurrence of thyroid cancer from the validation set is selected as the best model for this project. After training and validating various models, gradient boosting machine (GBM) yielded the best results when compared to other models trained for this project.

# 2 Analysis and Methods

## 2.1 Dataset Analysis

The dataset consists of 383 rows and 17 columns. The top 6 rows of the dataset is displayed in table 1.

Table 1: Top 6 rows of thyroid cancer recurrence dataset

| Age | Gender | Smoking | Hx.Smoking | Hx.Radiothreapy | Thyroid.Function | Physical.Examination | Adenopathy | Pathology | Focality | Risk | T | N | M | Stage | Response | Recurred |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 27 | F | No | No | No | Euthyroid | Single nodular goiter-left | No | Micropapillary | Uni-Focal | Low | T1a | N0 | M0 | I | Indeterminate | No |
| 34 | F | No | Yes | No | Euthyroid | Multinodular goiter | No | Micropapillary | Uni-Focal | Low | T1a | N0 | M0 | I | Excellent | No |
| 30 | F | No | No | No | Euthyroid | Single nodular goiter-right | No | Micropapillary | Uni-Focal | Low | T1a | N0 | M0 | I | Excellent | No |
| 62 | F | No | No | No | Euthyroid | Single nodular goiter-right | No | Micropapillary | Uni-Focal | Low | T1a | N0 | M0 | I | Excellent | No |
| 62 | F | No | No | No | Euthyroid | Multinodular goiter | No | Micropapillary | Multi-Focal | Low | T1a | N0 | M0 | I | Excellent | No |
| 52 | M | Yes | No | No | Euthyroid | Multinodular goiter | No | Micropapillary | Multi-Focal | Low | T1a | N0 | M0 | I | Indeterminate | No |

The dimensions of the dataset are,

```
## [1] 383  17
```

From the output of the structure of the dataset, it is evident that most of the variables are categorical. For the purposes of statistical modeling, these features need to be factorized.

```
str(thyroid_dataset)
```

```
## 'data.frame':    383 obs. of  17 variables:
##  $ Age                 : int  27 34 30 62 62 52 41 46 51 40 ...
##  $ Gender              : chr  "F" "F" "F" "F" ...
##  $ Smoking             : chr  "No" "No" "No" "No" ...
##  $ Hx.Smoking          : chr  "No" "Yes" "No" "No" ...
##  $ Hx.Radiothreapy     : chr  "No" "No" "No" "No" ...
##  $ Thyroid.Function    : chr  "Euthyroid" "Euthyroid" "Euthyroid" "Euthyroid" ...
##  $ Physical.Examination: chr  "Single nodular goiter-left" "Multinodular goiter" "Single nodular goi
##  $ Adenopathy          : chr  "No" "No" "No" "No" ...
##  $ Pathology           : chr  "Micropapillary" "Micropapillary" "Micropapillary" "Micropapillary" ..
##  $ Focality            : chr  "Uni-Focal" "Uni-Focal" "Uni-Focal" "Uni-Focal" ...
##  $ Risk                : chr  "Low" "Low" "Low" "Low" ...
##  $ T                   : chr  "T1a" "T1a" "T1a" "T1a" ...
##  $ N                   : chr  "N0" "N0" "N0" "N0" ...
##  $ M                   : chr  "M0" "M0" "M0" "M0" ...
##  $ Stage               : chr  "I" "I" "I" "I" ...
##  $ Response            : chr  "Indeterminate" "Excellent" "Excellent" "Excellent" ...
##  $ Recurred            : chr  "No" "No" "No" "No" ...
```

After factorizing the variables besides the Age feature, the structure looks like this.

```
str(thyroid_dataset)
```

```
## 'data.frame':    383 obs. of  17 variables:
##  $ Age                 : int  27 34 30 62 62 52 41 46 51 40 ...
##  $ Gender              : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 1 1 1 1 ...
##  $ Smoking             : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 1 ...
##  $ Hx.Smoking          : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 2 1 1 1 ...
##  $ Hx.Radiothreapy     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ Thyroid.Function    : Factor w/ 5 levels "Clinical Hyperthyroidism",..: 3 3 3 3 3 3 1 3 3 3 ...
##  $ Physical.Examination: Factor w/ 5 levels "Diffuse goiter",..: 4 2 5 5 2 2 5 5 5 5 ...
##  $ Adenopathy          : Factor w/ 6 levels "Bilateral","Extensive",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Pathology           : Factor w/ 4 levels "Follicular","Hurthel cell",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Focality            : Factor w/ 2 levels "Multi-Focal",..: 2 2 2 2 1 1 2 2 2 2 ...
##  $ Risk                : Factor w/ 3 levels "High","Intermediate",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ T                   : Factor w/ 7 levels "T1a","T1b","T2",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ N                   : Factor w/ 3 levels "N0","N1a","N1b": 1 1 1 1 1 1 1 1 1 1 ...
##  $ M                   : Factor w/ 2 levels "M0","M1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Stage               : Factor w/ 5 levels "I","II","III",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Response            : Factor w/ 4 levels "Biochemical Incomplete",..: 3 2 2 2 2 3 2 2 2 2 ...
##  $ Recurred            : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

Overall summary of the dataset is,

```
summary(thyroid_dataset)
```

```
##       Age        Gender  Smoking   Hx.Smoking Hx.Radiothreapy
##  Min.   :15.00   F:312   No :334   No :355    No :376
##  1st Qu.:29.00   M: 71   Yes: 49   Yes: 28    Yes:  7
##  Median :37.00
##  Mean   :40.87
##  3rd Qu.:51.00
##  Max.   :82.00
##
##                     Thyroid.Function                  Physical.Examination
##  Clinical Hyperthyroidism   : 20     Diffuse goiter             :  7
##  Clinical Hypothyroidism    : 12     Multinodular goiter        :140
##  Euthyroid                  :332     Normal                     :  7
##  Subclinical Hyperthyroidism:  5     Single nodular goiter-left : 89
##  Subclinical Hypothyroidism : 14     Single nodular goiter-right:140
##
##
##      Adenopathy           Pathology          Focality             Risk
##  Bilateral: 32   Follicular    : 28   Multi-Focal:136   High        : 32
##  Extensive:  7   Hurthel cell  : 20   Uni-Focal  :247   Intermediate:102
##  Left     : 17   Micropapillary: 48                     Low         :249
##  No       :277   Papillary     :287
##  Posterior:  2
##  Right    : 48
##
##    T         N        M       Stage                        Response   Recurred
##  T1a: 49   N0 :268   M0:365   I  :333   Biochemical Incomplete: 23   No :275
##  T1b: 43   N1a: 22   M1: 18   II : 32   Excellent             :208   Yes:108
##  T2 :151   N1b: 93            III:  4   Indeterminate         : 61
##  T3a: 96                      IVA:  3   Structural Incomplete : 91
##  T3b: 16                      IVB: 11
##  T4a: 20
##  T4b:  8
```

### 2.2 Dataset Cleaning

Let's explore the dataset for NAs and remove them if present. The output of anyNA yields, FALSE. Overall, the dataset appears to be clean and does not require further pre-processing since we have already factorized the categorical variables. Table 2 displays the variables and their NA count if any.

Table 2: Table of variables and count of NA

| Variable | NA count |
|---|---|
| Age | 0 |
| Gender | 0 |
| Smoking | 0 |
| Hx.Smoking | 0 |
| Hx.Radiothreapy | 0 |
| Thyroid.Function | 0 |
| Physical.Examination | 0 |
| Adenopathy | 0 |
| Pathology | 0 |
| Focality | 0 |
| Risk | 0 |
| T | 0 |
| N | 0 |
| M | 0 |
| Stage | 0 |
| Response | 0 |
| Recurred | 0 |

## 2.3 Dataset Exploration

The proportion of the target variable, Recurred is plotted below. Recurrence appears to be 28.1984334% whereas non-recurrence is 71.8015666%.

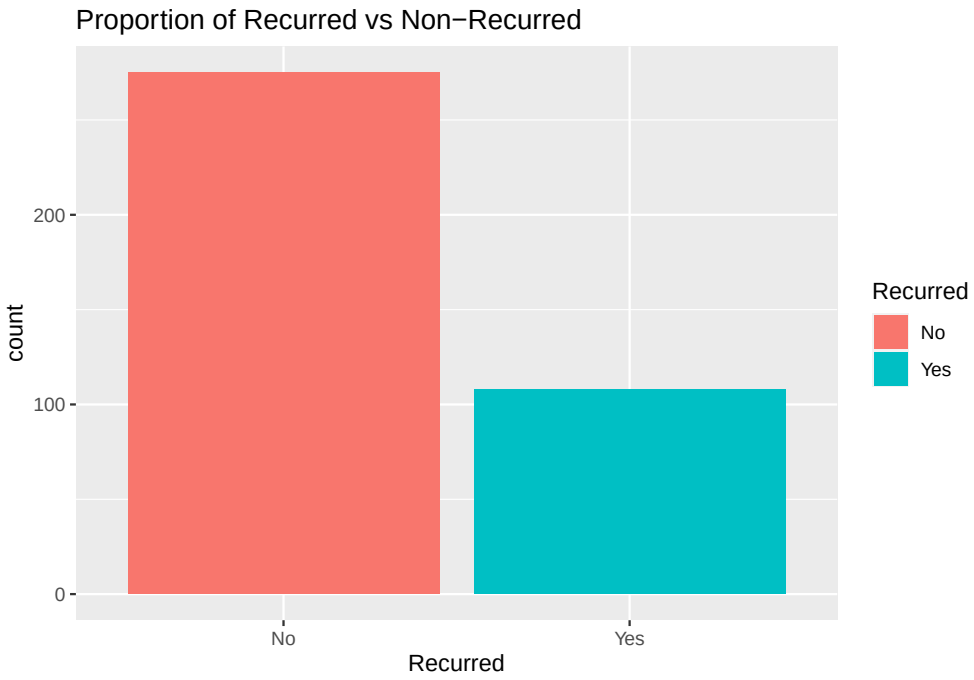
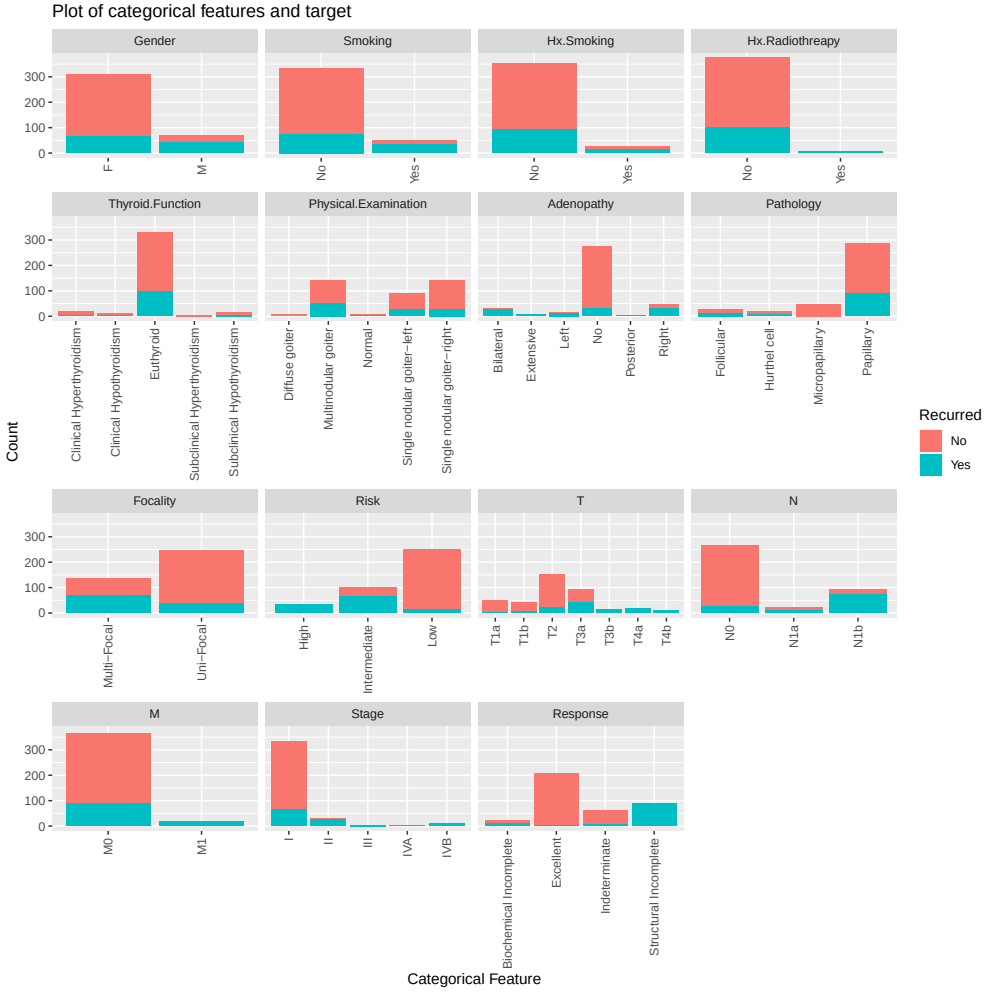



Table 3: Count of Yes and No for the Recurred target variable

| Recurred | NA count |
|---|---|
| No | 275 |
| Yes | 108 |

A plot of the categorical features and target variable is displayed below. The count of recurrence vs non-recurrence appears to be mostly normally distributed across the features.



A plot of the Age feature which is of integer type and the target variable is plotted below. Its distribution is also similar to the previous histogram plot.

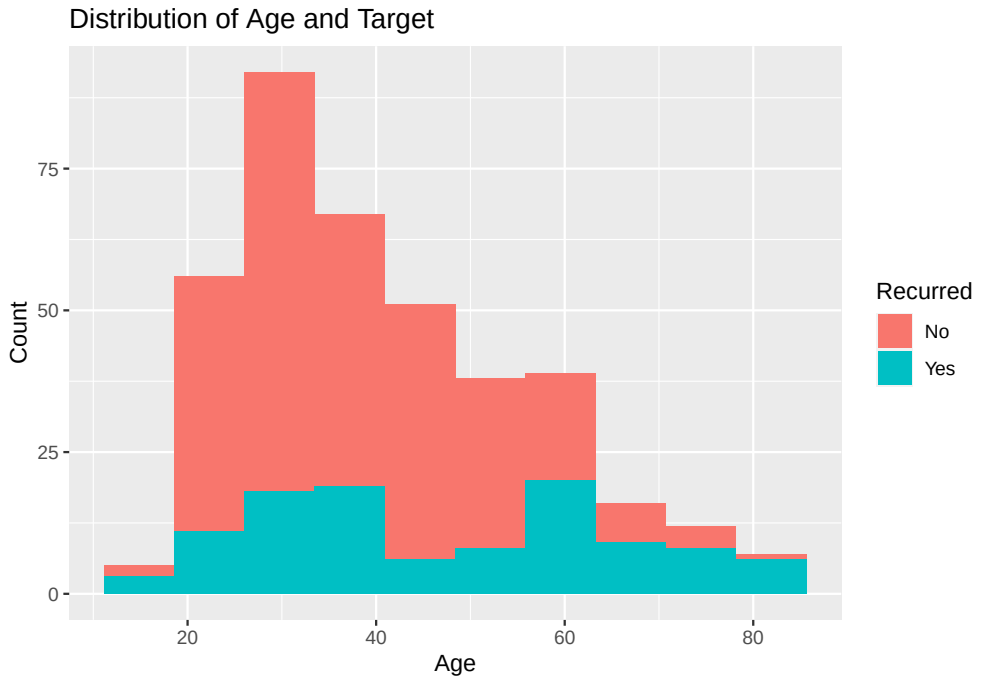
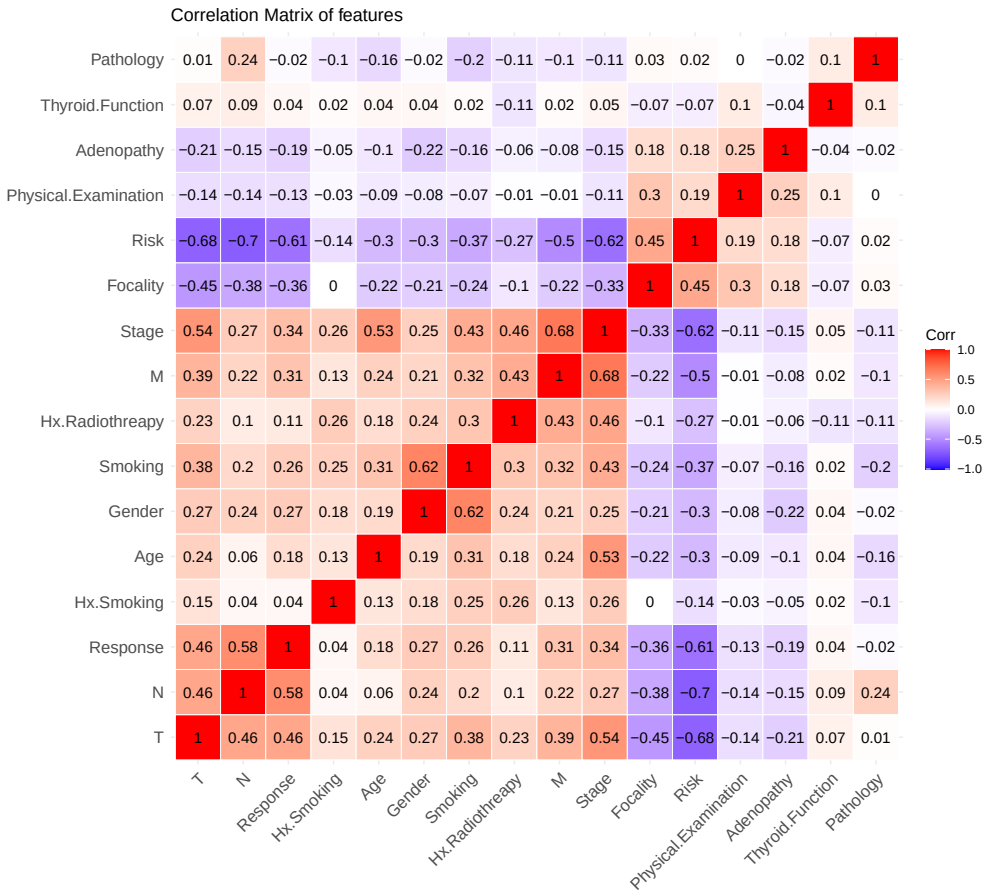



Table 4 displays the top 6 rows of the factorized categorical features to aid with statistical modeling and computations when training and validating the machine learning models.

Table 4: Top 6 rows of factorized categorical features

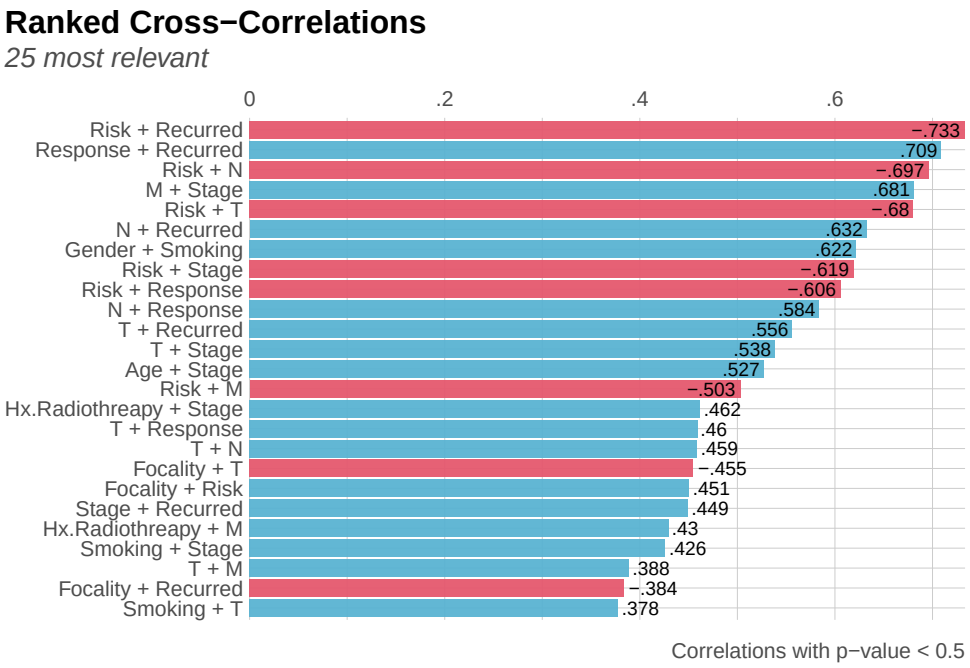
| Age | Gender | Smoking | Hx.Smoking | Hx.Radiothreapy | Thyroid.Function | Physical.Examination | Adenopathy | Pathology | Focality | Risk | T | N | M | Stage | Response | Recurred |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 27 | 1 | 1 | 1 | 1 | 3 | 4 | 4 | 3 | 2 | 3 | 1 | 1 | 1 | 1 | 3 | 1 |
| 34 | 1 | 1 | 2 | 1 | 3 | 2 | 4 | 3 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 |
| 30 | 1 | 1 | 1 | 1 | 3 | 5 | 4 | 3 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 |
| 62 | 1 | 1 | 1 | 1 | 3 | 5 | 4 | 3 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 |
| 62 | 1 | 1 | 1 | 1 | 3 | 2 | 4 | 3 | 1 | 3 | 1 | 1 | 1 | 1 | 2 | 1 |
| 52 | 2 | 2 | 1 | 1 | 3 | 2 | 4 | 3 | 1 | 3 | 1 | 1 | 1 | 1 | 3 | 1 |

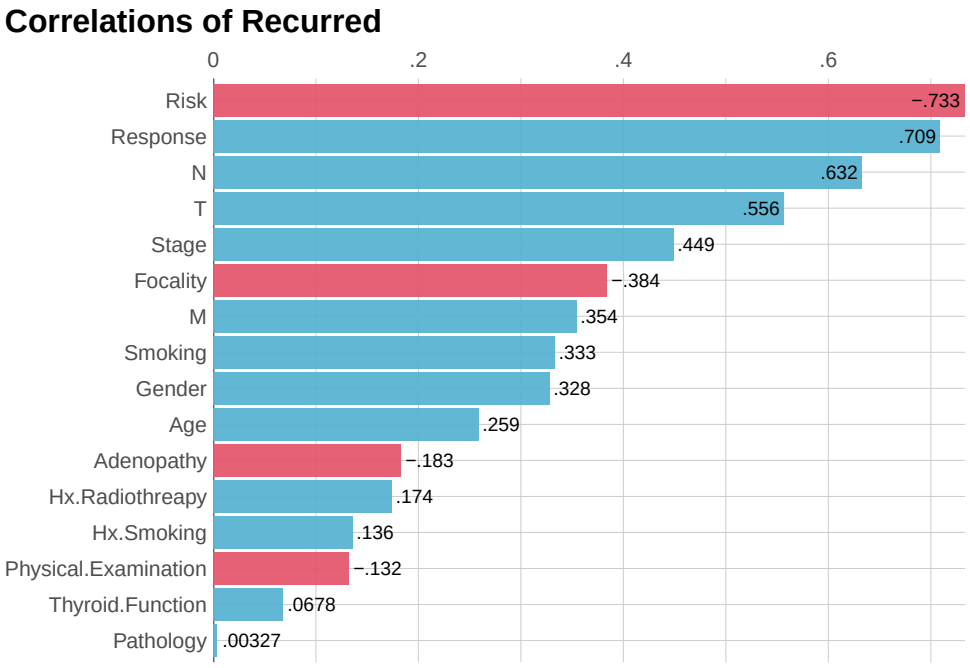
Next step is to identify the correlation matrix between the features.The plot below illustrates the correlation matrix. There does not seem to be a strong correlation between the features. This eliminates the need to remove any features when training the models. All correlations are less than 0.9 which indicates there is no strong correlation.




```
## All correlations <= 0.9
```

The plots below illustrate the ranked cross correlations and correlations of the target variable, Recurred. Once again, we can observe that there are no strong correlations between the feature set.



### 2.4 Machine Learning Modeling Methods

In this section, we will split the thyroid dataset into train and validation sets.The train set is 80% and validation is 20% of the dataset.As stated in this technical report, empirical studies confirm that best results are achieved if we split 70-80% of the data for training and 20-30% for testing. Since the dataset is relatively small, it is split into train and validation rather than train, test and validation. Nevertheless, cross-validation technique is employed which can be leveraged for smaller datasets. Also, techniques such as repeatedcv is used to help with model evaluation on different subsets of the dataset.

Caret package, which is short for Classification and REgression Training contains functions which are used in this project to train and validate the models.The caret package uses a number of R packages and performs hyperparameter tuning by default. Parameters to control training model is configured so it is consistent across the models where applicable. The sampling method used is repeatedcv and the number of resampling iterations is 10. The number of complete sets of folds to compute is set to 10. Summary function used is twoClassSummary, which is used to compute performance metric across resamples. This summary function computes the AUC (Area under the ROC curve) as well as the sensitivity and specificity metrics which will be described later in this report. ROC (Receiver Operating Charateristic Curve) is a graph showing the performance of a model. It plots TPR (True Positive Rate) and FPR (False Positive Rate). TPR is also called as recall.The class probabilities to be computed for classification models in each resample is enabled.Repeatedcv is chosen over cv because repeatedcv performs 10-fold cross-validation on the training data using a different set of folds for each cross-validation and this aids with yielding more robust and accurate results. The caret package contains train function which includes the model method, configures tuning parameters for that corresponding method, and also ROC metric to be measured by training that method. The predict function is used to make predictions on the trained dataset.

Confusion Matrix is used to evaluate performance of a machine learning model including the accuracy of a model. It includes true positives, true negatives, false positives and false negatives. In the table below, the columns represent the actual values of the target variable, whereas the rows represent the predicted values of the target variable. The confusion matrix compares the actual value of the target with those predicted by the models. It is used to calculate accuracy, precision, recall and F1 score.

Table 5: Confusion Matrix

| | Positive | Negative |
|---|---|---|
| Positive | TP | FP |
| Negative | FN | TN |

TP = Number of samples correctly predicted as positive

FP = Number of samples incorrectly predicted as positive

TN = Number of samples correctly predicted as negative

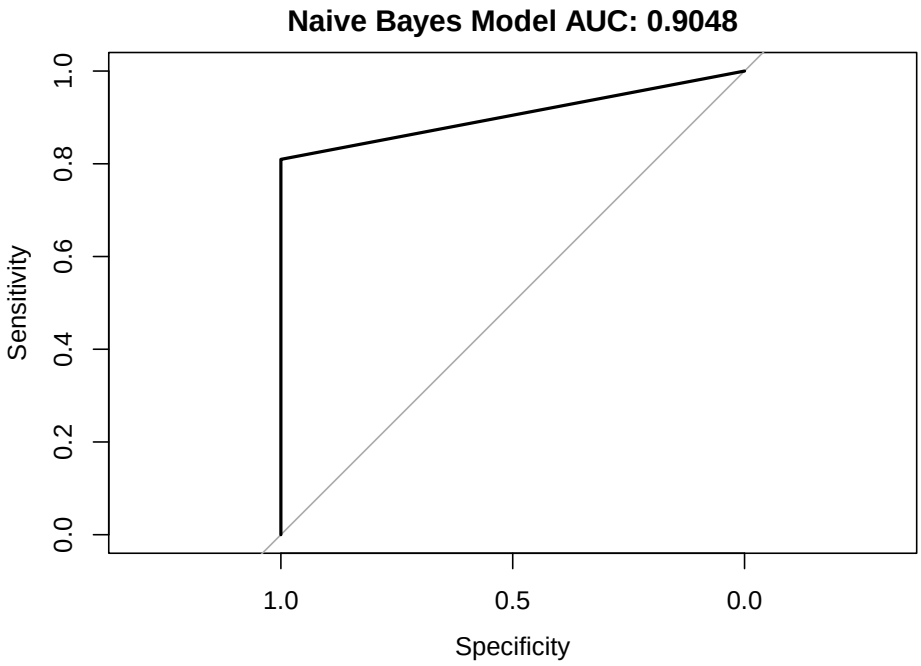
FN = Number of samples incorrectly predicted as negative

Precision is calculated as $\frac{TP}{TP+FP}$, recall is calculated as $\frac{TP}{TP+FN}$ and accuracy is calculated as $\frac{TP+TN}{TP+FP+TN+FN}$. Precision metric is used when false positive(FP) trumps false negatives(FN) and recall is used when false negative(FN) trumps false positive(FP). In this project, it is crucial for note recall metrics since in medical reports we do not want to miss actual positive cases. Recall is also called as sensitivity (TPR) because it is highly sensitive and thereby fewer false negatives. Specificity is also referred as True Negative Rate (TNR). Balanced Accuracy in the confusion matrix output represents the AUC.

### 2.4.1 Naive Bayes (NB)

Naive Bayes is one of the basic classification based machine learning algorithms.It assumes that all features are equally distributed to the outcome and predictors are conditionally independent. Despite being simple, Naive Bayes trained model performs well especially on smaller datasets. The results of confusion matrix and AUC is displayed below.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No  55   4
##        Yes  0  17
##
##                Accuracy : 0.9474
##                  95% CI : (0.8707, 0.9855)
##     No Information Rate : 0.7237
##     P-Value [Acc > NIR] : 6.695e-07
##
##                   Kappa : 0.8602
##
##  Mcnemar's Test P-Value : 0.1336
##
##             Sensitivity : 0.8095
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 0.9322
##              Prevalence : 0.2763
##          Detection Rate : 0.2237
##    Detection Prevalence : 0.2237
##       Balanced Accuracy : 0.9048
##
##        'Positive' Class : Yes
##
```

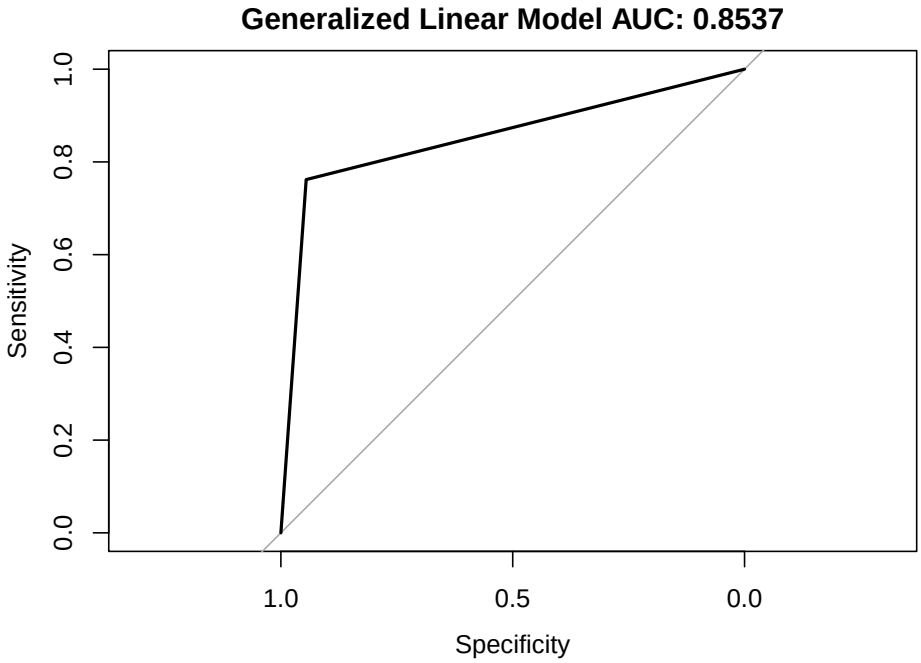



### 2.4.2 Generalized Linear Model (GLM)

Generalized Linear Model is a class of models that allows flexible and non-linear relationships between response and predictor variables. The caret package 'glm' is used to train the train set. ROC is a metric used to measure the model performance. Prediction is performed by leveraging the predict function with the trained model and validation set as inputs. Confusion matrix is used to evaluate the performance of this classification model by comparing the predicted values against the actual target values.The results of confusion matrix and AUC is displayed below.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No  52   5
##        Yes  3  16
##
##                Accuracy : 0.8947
##                  95% CI : (0.8031, 0.9534)
##     No Information Rate : 0.7237
##     P-Value [Acc > NIR] : 0.0002533
##
##                   Kappa : 0.7288
##
##  Mcnemar's Test P-Value : 0.7236736
##
##             Sensitivity : 0.7619
##             Specificity : 0.9455
##          Pos Pred Value : 0.8421
##          Neg Pred Value : 0.9123
##              Prevalence : 0.2763
```

```
##          Detection Rate : 0.2105
##    Detection Prevalence : 0.2500
##       Balanced Accuracy : 0.8537
##
##        'Positive' Class : Yes
##
```

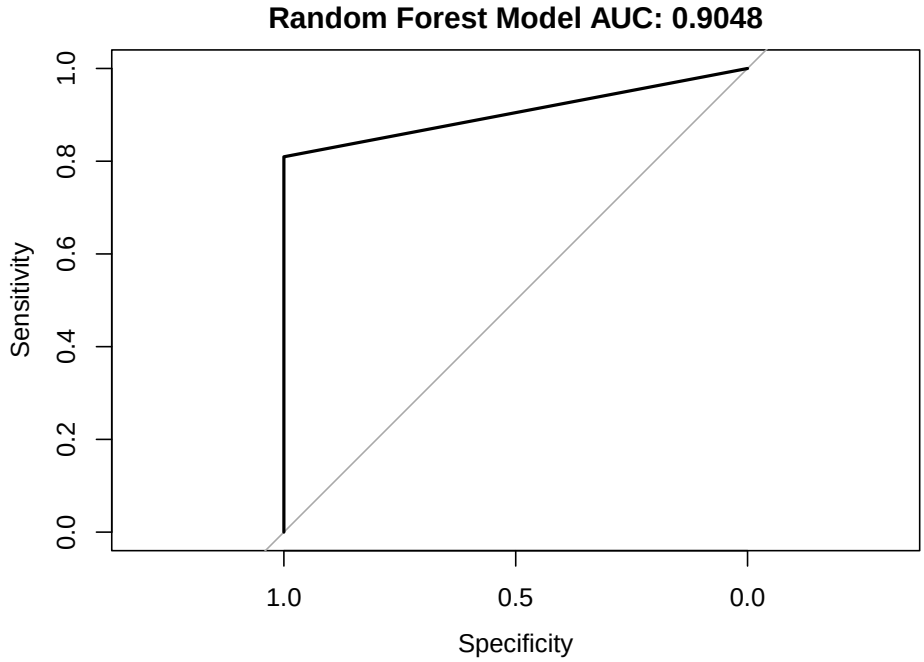



### 2.4.3 Random Forest (RF)

Random forest is a machine learning algorithm that combines the output of multiple decision trees to achieve a single output. It performs well on both classification and regression based problems. In this project, the decision tree is an example of a classification problem, where the labels are 'Yes' or 'No' for 'Recurred' target variable.They are prone to predict accurate results especially when features are uncorrelated with each other.The results of confusion matrix and AUC is displayed below.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No  55   4
##        Yes  0  17
##
##               Accuracy : 0.9474
##                 95% CI : (0.8707, 0.9855)
##    No Information Rate : 0.7237
##    P-Value [Acc > NIR] : 6.695e-07
##
##                  Kappa : 0.8602
##
##  Mcnemar's Test P-Value : 0.1336
```

```
##
##             Sensitivity : 0.8095
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 0.9322
##              Prevalence : 0.2763
##          Detection Rate : 0.2237
##    Detection Prevalence : 0.2237
##       Balanced Accuracy : 0.9048
##
##        'Positive' Class : Yes
##
```

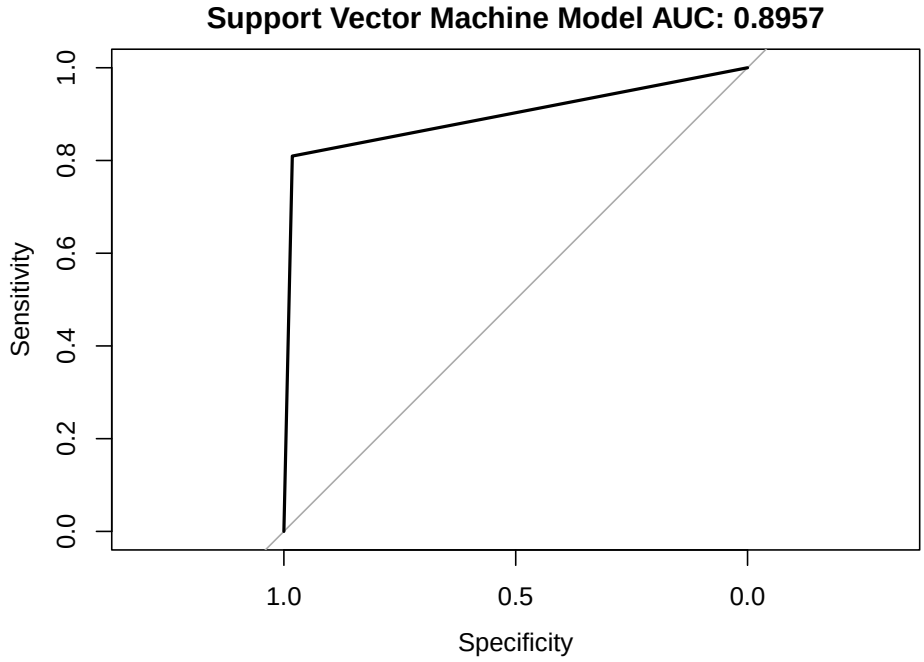



### 2.4.4 Support Vector Machine Model (SVM)

Support Vector Machine (SVM) is used to find a hyperplane (decision boundary) that classifies the data points. To classify the data points, many possible hyperplanes could be chosen. The goal is to find a plane that yields the maximum distance between the data points of both classes and in this project, the classes are 'Yes' and 'No' of the target variable, Recurred.SVMLinear method fits a linear SVM model and supports binary classification method with linear kernel. The results of confusion matrix and AUC is displayed below.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No  54   4
##        Yes  1  17
##
##                Accuracy : 0.9342
##                  95% CI : (0.8531, 0.9783)
```

```
##     No Information Rate : 0.7237
##     P-Value [Acc > NIR] : 3.843e-06
##
##                   Kappa : 0.8279
##
##  Mcnemar's Test P-Value : 0.3711
##
##             Sensitivity : 0.8095
##             Specificity : 0.9818
##          Pos Pred Value : 0.9444
##          Neg Pred Value : 0.9310
##              Prevalence : 0.2763
##          Detection Rate : 0.2237
##    Detection Prevalence : 0.2368
##       Balanced Accuracy : 0.8957
##
##        'Positive' Class : Yes
##
```

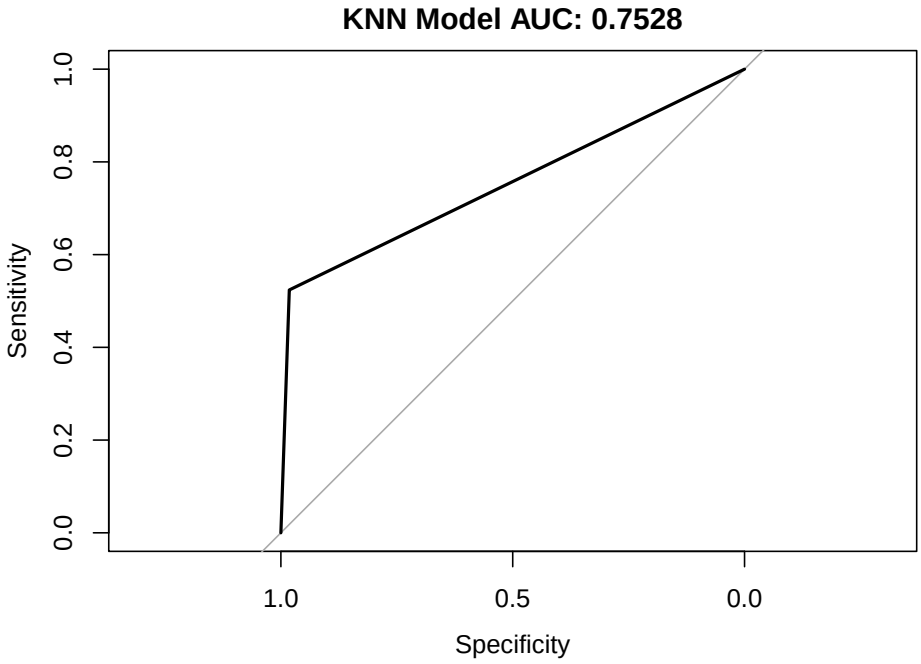



### 2.4.5 K-Nearest Neighbor Model (KNN)

K-Nearest Neighbor (KNN) is a widely used algorithm for classification problems in machine learning. It fairs across calculation and prediction time. The 'K' in KNN is the nearest neighbor(s) to be considered for computation. 'K' is typically an odd number to avoid ties in classification. In this project, a grid of 'k' values 1 through 21 are used. The results of confusion matrix and AUC is displayed below.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
```

```
##        No  54  10
##        Yes  1  11
##
##                Accuracy : 0.8553
##                  95% CI : (0.7558, 0.9255)
##     No Information Rate : 0.7237
##     P-Value [Acc > NIR] : 0.005153
##
##                   Kappa : 0.5828
##
##  Mcnemar's Test P-Value : 0.015861
##
##             Sensitivity : 0.5238
##             Specificity : 0.9818
##          Pos Pred Value : 0.9167
##          Neg Pred Value : 0.8438
##              Prevalence : 0.2763
##          Detection Rate : 0.1447
##    Detection Prevalence : 0.1579
##       Balanced Accuracy : 0.7528
##
##        'Positive' Class : Yes
##
```

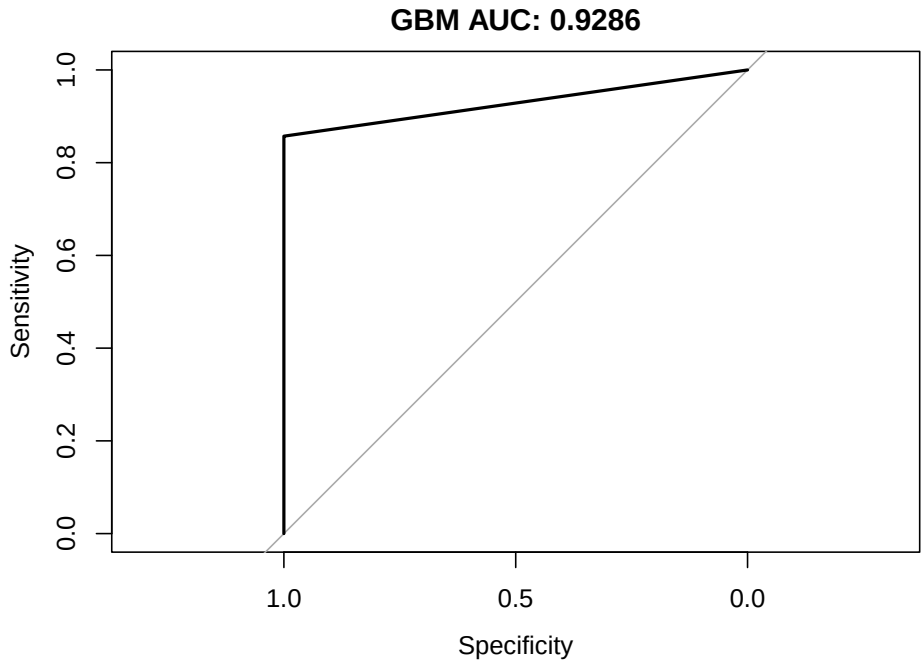



### 2.4.6 Gradient Boosting Machine Model (GBM)

Gradient Boosting Machine (GBM) is a popular machine learning algorithms. Unlike linear regression, naive Bayes and SVM algorithms which are based on a single predictive model, and unlike random forest which is based on building an ensemble model, boosting adds new models to the ensemble sequentially. This algorithm provides several tuning parameters and caret handles the hyper parameters tuning automatically.

In this project, 'gbm' method is used which is the original R implementation of GBM and is stochastic based. The results of confusion matrix and AUC is displayed below.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction No Yes
##        No  55   3
##        Yes  0  18
##
##                Accuracy : 0.9605
##                  95% CI : (0.8889, 0.9918)
##     No Information Rate : 0.7237
##     P-Value [Acc > NIR] : 9.227e-08
##
##                   Kappa : 0.8967
##
##  Mcnemar's Test P-Value : 0.2482
##
##             Sensitivity : 0.8571
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 0.9483
##              Prevalence : 0.2763
##          Detection Rate : 0.2368
##    Detection Prevalence : 0.2368
##       Balanced Accuracy : 0.9286
##
##        'Positive' Class : Yes
##
```



# 3 Results

A tabular structure of machine learning models and their confusion matrix metrics results are displayed in table 5. The model with the highest value for most metrics especially sensitivity and balanced accuracy is selected as the 'best model' among the other models trained and validated for this project. The Gradient Boosting Machine (GBM) model is selected as the best model for this project based on the confusion matrix results.

Table 6: Confusion Matrix Results across ML models

| | **NB** | **GLM** | **RF** | **SVM** | **KNN** | **GBM** | **Best Model** |
|---|---|---|---|---|---|---|---|
| Sensitivity | 0.8095238 | 0.7619048 | 0.8095238 | 0.8095238 | 0.5238095 | 0.8571429 | GBM |
| Specificity | 1.0000000 | 0.9454545 | 1.0000000 | 0.9818182 | 0.9818182 | 1.0000000 | RF |
| Pos Pred Value | 1.0000000 | 0.8421053 | 1.0000000 | 0.9444444 | 0.9166667 | 1.0000000 | RF |
| Neg Pred Value | 0.9322034 | 0.9122807 | 0.9322034 | 0.9310345 | 0.8437500 | 0.9482759 | GBM |
| Precision | 1.0000000 | 0.8421053 | 1.0000000 | 0.9444444 | 0.9166667 | 1.0000000 | RF |
| Recall | 0.8095238 | 0.7619048 | 0.8095238 | 0.8095238 | 0.5238095 | 0.8571429 | GBM |
| F1 | 0.8947368 | 0.8000000 | 0.8947368 | 0.8717949 | 0.6666667 | 0.9230769 | GBM |
| Prevalence | 0.2763158 | 0.2763158 | 0.2763158 | 0.2763158 | 0.2763158 | 0.2763158 | KNN |
| Detection Rate | 0.2236842 | 0.2105263 | 0.2236842 | 0.2236842 | 0.1447368 | 0.2368421 | GBM |
| Detection Prevalence | 0.2236842 | 0.2500000 | 0.2236842 | 0.2368421 | 0.1578947 | 0.2368421 | GLM |
| Balanced Accuracy | 0.9047619 | 0.8536797 | 0.9047619 | 0.8956710 | 0.7528139 | 0.9285714 | GBM |

# 4 Conclusion

In summary, the goal of this project is to train and validate machine learning algorithms to identity the best model to classify the recurrence of differentiated thyroid cancer based on the feature variables provided in the dataset. After training various models such as Linear Model, Naive Bayes, Support Vector Machine, Random Forest, and KNN, Gradient Boosting Machine (GBM) model yielded the best results for sensitivity which is a crucial diagnostic metric in medical field. This model also produced highest value for Area Under the Curve (AUC).

Even though GBM model yielded the best metrics for this project, there are various hyper parameters that could be fine tuned to generate better results. Future work would include additional research on much larger datasets and train advanced algorithms that could involve intensive computing. Also, other packages besides caret could also be utilized to try various techniques and explore advanced metrics besides cross-validation to develop an ideal model.

# 5 References

Irizarry, Rafael A. Introduction to Data Science, rafalab.dfci.harvard.edu/dsbook-part-1/ https://rafalab.dfci.harvard.edu/dsbook-part-1/

Irizarry, Rafael A. Introduction to Data Science, rafalab.dfci.harvard.edu/dsbook-part-2/ https://rafalab.dfci.harvard.edu/dsbook-part-2/

Irizarry, Rafael A. Introduction to Data Science Data Analysis and Prediction Algorithms with R. CRC Press, 2020.

https://archive.ics.uci.edu/dataset/915/differentiated+thyroid+cancer+recurrence

Borzooei,Shiva and Tarokhian,Aidin. (2023). Differentiated Thyroid Cancer Recurrence. UCI Machine Learning Repository. https://doi.org/10.24432/C5632J.