Thyroid Disease prediction through Machine learning techniques with enhanced datasets

- Aryan Malhotra (2019ABPS0893P)

Abstract

1- Introduction

The thyroid gland plays an important role in regulating metabolism, heart rate, body temperature and functioning of various other organs in the human body. The thyroid gland produces Levothyroxine (T4) and Triiodothyronine (T3) amongst other hormones to carry out its functions. A reduction in the production of these hormones may lead to hyperthyroidism and hypothyroidism.

Medical diagnoses are often highly specific due to the need for accurate identification of particular conditions. The aim of this study is to augment the number of target classes used from the initial five to seven distinct classes, enriching the diversity of diagnostic possibilities in Thyroid disease classification, while retaining model performance.

The initial phase of this research aimed to replicate the high-performance model proposed by [1]. Their study extensively investigated the Thyroid disease dataset, achieving a remarkable classification score of 0.99 using the Random Forest Classifier algorithm with specific parameters and feature selection methods. Building upon their work, our study seeks to expand the scope by using Synthetic minority over sampling technique (SMOTE) class sampling methodologies.

2- Background

The existing work on this dataset includes use of several machine learning and deep learning approaches to maximise model prediction score. The ML models used include but are not limited to Random Forest, Adaboost, Support Vector method with a tuning range of parameters that have been tested and optimised for performance. The DL approaches include using Convolutional neural networks, Long short term memory neural networks etc. Due to dataset size constraints, the Machine learning models performed much better than DL models. The model with the best performance was a Random Forest Classifier model with parameters n_estimators = 200 (number of estimators) and max_depth = 20 (depth of trees) and features selected by implementing Machine Learning Feature selection method with a threshold value of 0.015 for selecting the features to be considered.

Recreating and validating this model with an expanded target class set holds the promise of achieving comparable or superior diagnostic accuracy in the context of a broader disease classification framework. The potential implications of this research stretch beyond the immediate scope of thyroid disease classification. The strategic utilisation of sampling methods to bolster dataset sizes across medical classification domains has broader implications. These methods, when adeptly employed, could serve as a pivotal strategy to enhance dataset richness, facilitating the incorporation of a greater variety of target classes in model development across diverse medical classification tasks.

In essence, this study endeavours to showcase not only the efficacy of machine learning in thyroid disease classification but also the transformative potential of sampling methodologies in expanding dataset size and target class diversity, thereby paving the way for more nuanced and accurate medical diagnostics.

3- Methodology

The methodology for implementing this project first started with acquiring the Thyroid disease dataset. The dataset consists of 29 separate target class labels as the medical diagnosis. The data was preprocessed to include only the relevant target classes for each model. Preprocessing methods such as one hot encoding have been utilised to deal with boolean values in the data.

Feature selection is carried out based on Feature importance scores and model training is done with a train/test split of 0.2 (except for Borderline SMOTE Sampled Model). The model score is evaluated for benchmarking.

All models (except Borderline SMOTE) have been validated using 10-Fold Cross Validation. F1 scores such as accuracy and recall have been tracked across folds and across models graphically. Confusion matrices have been plotted on heatmaps to evaluate model performance per class and in depth.

3.1 - Dataset & EDA

The dataset was acquired from the UCI Machine Learning Repository. The shape of the dataset acquired is as follows:

Number of Samples	Number of features	Number of target labels
9172	31	29

Five target classes have been considered for training in the original model. These classes have been balanced so that there is no bias in training the dataset. For the normal class diagnosis, 400 random samples have been selected for training and testing. The target classes and their counts are as follows:

Class Name	Target Label	Original Count	Final Count
Normal	-	6771	400
Primary Hypothyroid	F	233	233
Increased binding protein	I	346	346
Compensated Hypothyroid	G	359	359
Concurrent non-thyroidal illness	K	436	436

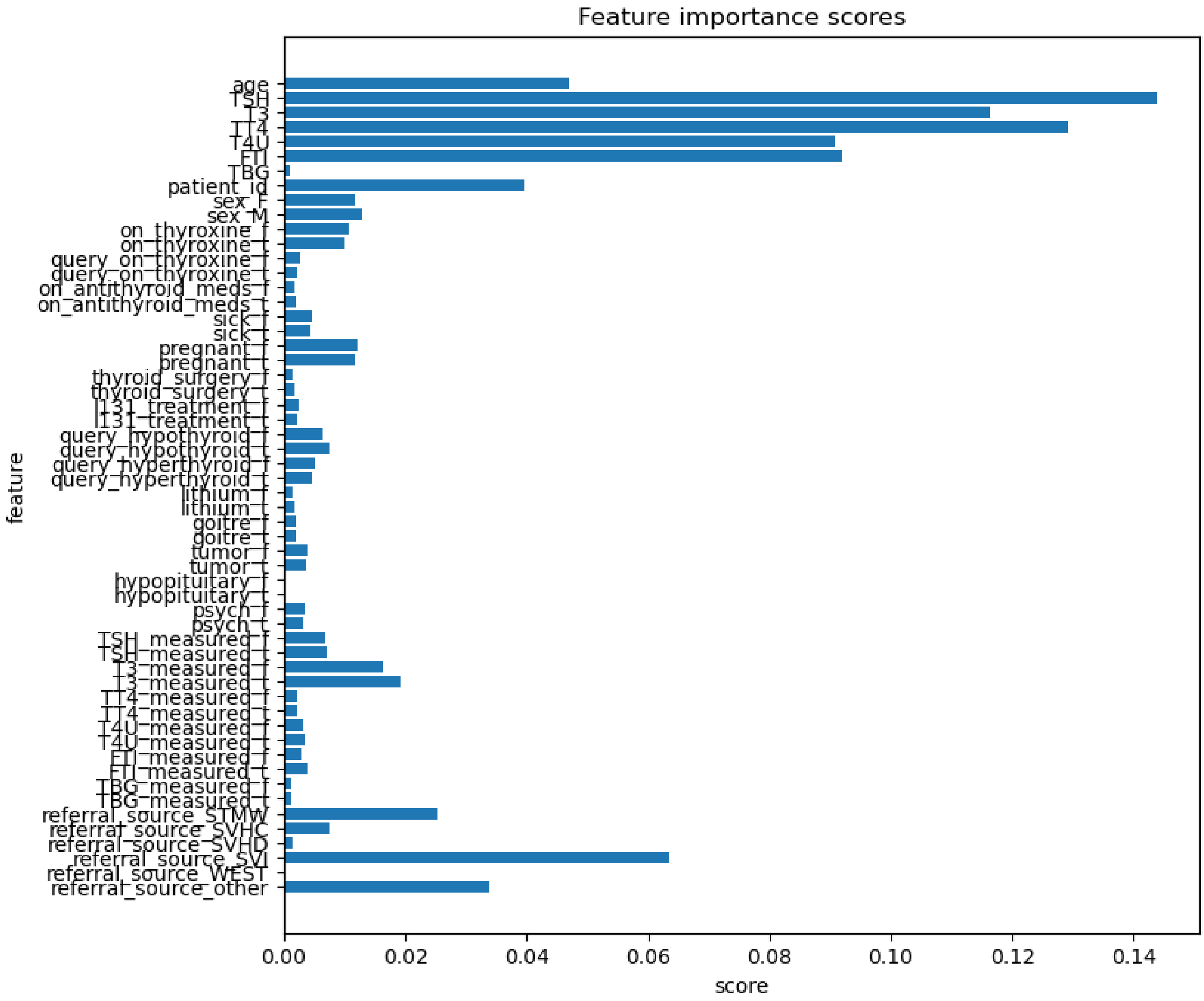
Out of 28 features, 19 features are of boolean type and the remaining 9 features are continuous variable type. In the preprocessing phase, one-hot encoding was employed to transform boolean data into a format compatible with machine learning models.

3.2 - Feature Selection and Model training

In the research, an Extra Tree Classifier with a threshold value of 0.015 for feature selection is used. 11 features were extracted and utilised for model training in all the models. The features selected for consideration include 'age', 'TSH', 'T3', 'TT4', 'T4U', 'FTI', 'patient_id', 'T3_measured_f', 'T3_measured_t', 'referral_source_SVI', and 'referral_source_other'.

For Model training, a Random Forest Classifier with parameters of 200 decision trees (n_estimators) and a maximum tree depth of 20 (max_depth) were the parameters used. This model performed the best in earlier studies [1] and therefore serves as the base model on which all the sampled data is trained and tested.

The feature importance scores for all the features are displayed below :



3.3 - SMOTE sampled dataset model analysis

Synthetic Minority Over-sampling Technique (SMOTE) is a data sampling technique which works by interpolating between existing data points and synthesising new samples. This strategic approach was instrumental in expanding the scope of target classes from an initial count of 5 to a more diverse count of 7. This technique was essential in maintaining class balances while performing model training.

The resultant dataset was preprocessed, wherein features earlier identified through the Machine Learning feature selection process were retained. This curated dataset was then subjected to the Random Forest Classifier, maintaining a consistent configuration of 200 decision trees and a maximum depth parameter set at 20. This uniformity in model configuration ensured consistency and comparability in evaluating the model's performance across diverse datasets.

The target classes and their counts for the SMOTE sampled model are as follows:

Class Name	Target Label	Count
Normal	-	400
Primary Hypothyroid	F	233
Increased binding protein	I	346
Compensated Hypothyroid	G	359
Concurrent non-thyroidal illness	K	436
Discordant assay results	R	231
Hyperthyroid	A	231

3.4 - Borderline SMOTE sampled Model analysis

Borderline SMOTE is a data sampling strategy used in augmenting datasets by synthesising new samples proximal to the decision boundary. The k_neighbours parameter is used to define the number of nearest neighbours for synthetic sample generation, whereas the m_neighbours parameter determines the threshold for determining borderline instances. The application of Borderline SMOTE with k_neighbors set to 5 and m_neighbors set to 2 was implemented.

Consequently, the resultant dataset expanded to a substantial count of 121,992 samples, a figure that surpassed initial expectations significantly. The original samples were removed from the sampled dataset and set aside for testing purposes, intended to validate the performance of the trained model.

Post preprocessing, a selection of 10 balanced target class labels, each encompassing approximately 6,500 samples, was considered for model training. This selection was made to ensure a balanced representation across the spectrum of target classes, reducing chances of bias in the trained model.

The feature selection criteria and the algorithm used was the same as in the earlier two models above.

The target classes and their counts for the Borderline SMOTE sampled model are as follows:

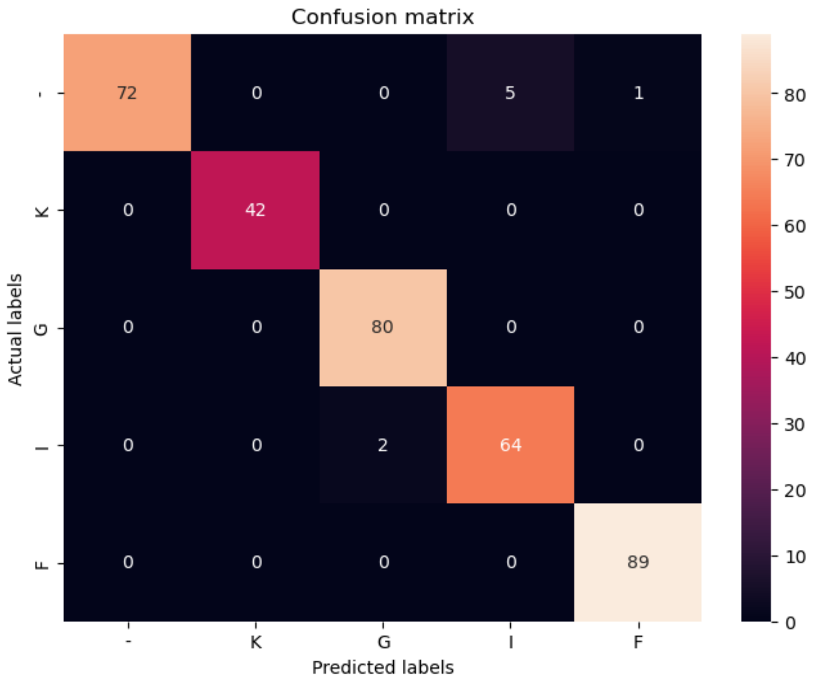
Class Name	Target Label	Count
Primary Hypothyroid	F	6538

Increased binding protein	I	6425
Compensated Hypothyroid	G	6412
Concurrent non-thyroidal illness	K	6335
Discordant assay results	R	6575
Hyperthyroid	A	6624
Elevated TBG	S	6686
Over Replaced	N	6661
Consistent with replacement therapy	L	6656
Under Replaced	M	6660

4 - Results and Analysis

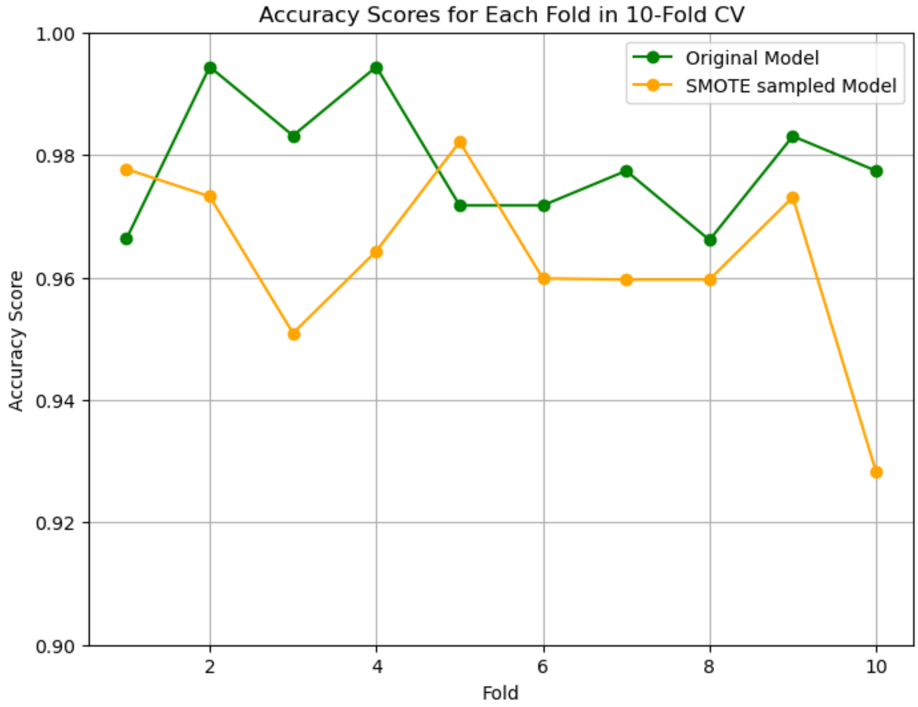
The original model with feature extraction and no data sampling performed with a mean accuracy score of 0.9786 and a mean recall score of 0.9805. The standard deviation across folds was less than 0.01 (0.0097, 0.0087 respectively) in the accuracy and recall score, giving us the sense that the dataset was well balanced.

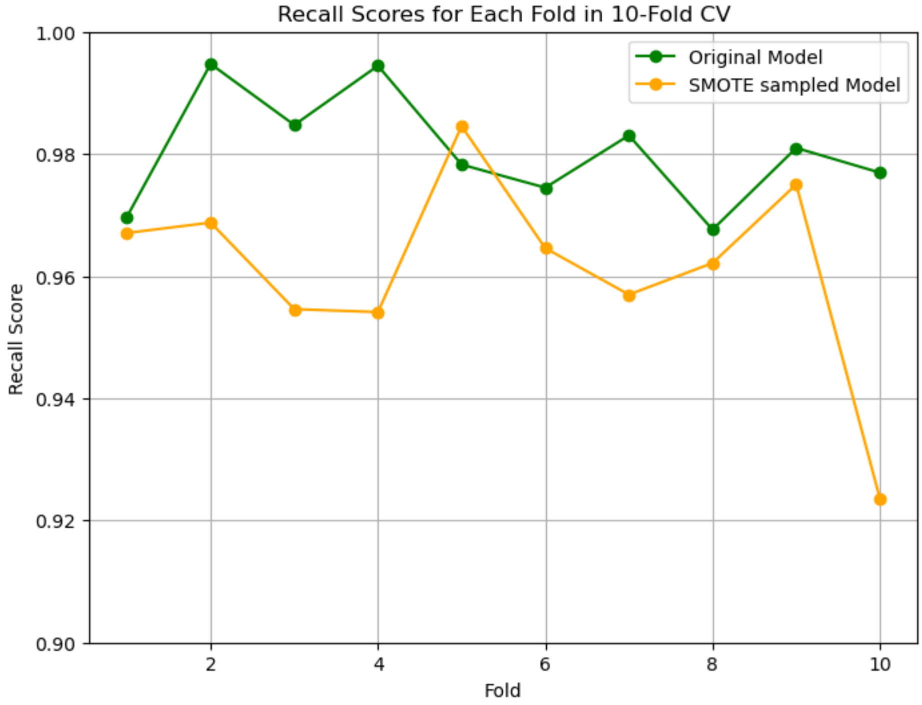
A confusion matrix was plotted with the model predictions and truth labels to ascertain model performance across classes. The confusion matrix heatmap for the original model (mod_v1) is shown below :



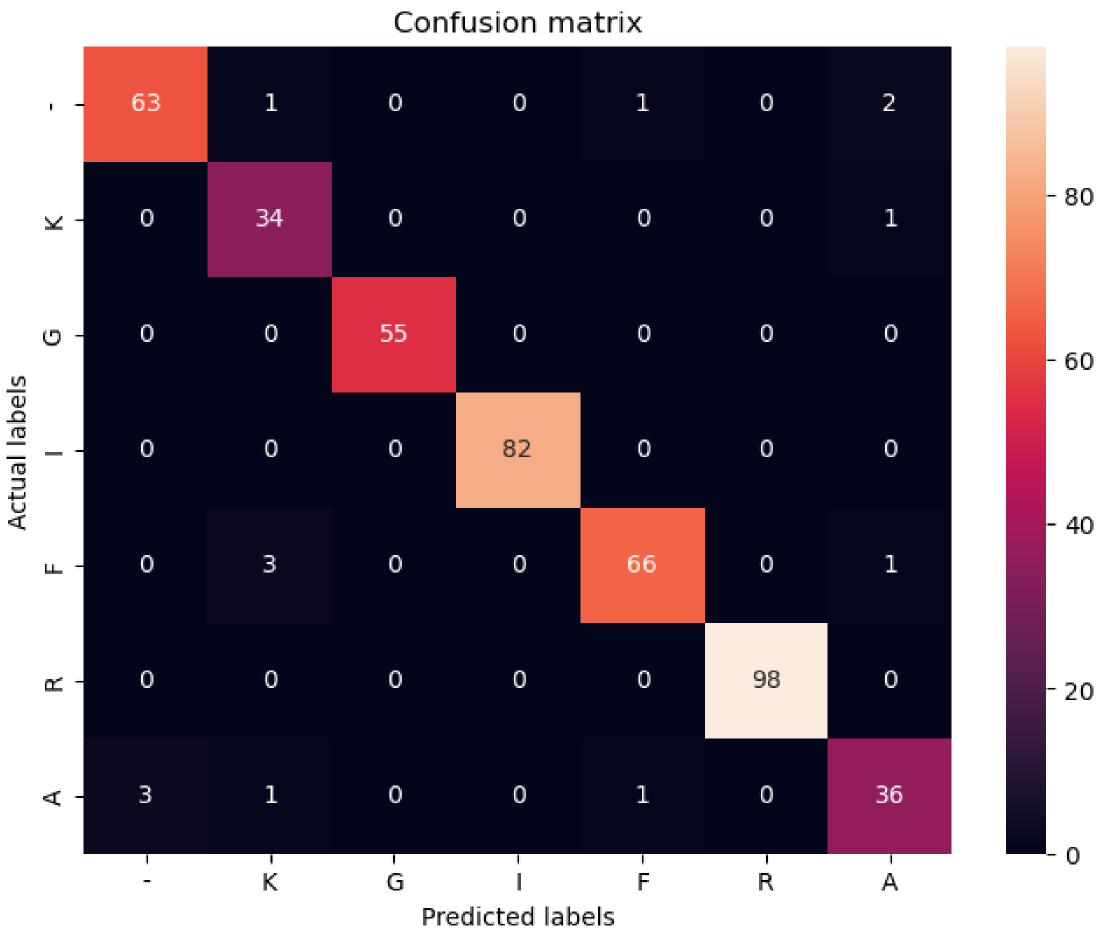
Mod_v2 is the model with SMOTE sampled data added to increase the target classes to 7. 10 Fold Cross validation of the model reveals that the model performed with a mean accuracy score of 0.9629 and mean recall score of 0.9611. The standard deviation measured across the folds for the 2 parameters were 0.0148 and 0.0154 respectively.

Graphs of the accuracy and recall scores for the models are shown below :





The heatmap for the SMOTE sampled model is shown below:



	-	K	G	I	F	R	A
-	63	1	0	0	1	0	2
K	0	34	0	0	0	0	1
G	0	0	55	0	0	0	0
I	0	0	0	82	0	0	0
F	0	3	0	0	66	0	1
R	0	0	0	0	0	98	0
A	3	1	0	0	1	0	36

This tells us that while the model performed better on original data, different models can be utilised to address target class inclusions. While some models may predict a better score, other models can be used to ensure a more diverse class representation during classification.

Borderline SMOTE model (Mod_v3) was tested on the entire preprocessed original dataset. A dataset using a sampling ratio of 13 performing well was an ambitious target. This model performed with a score of 0.8807 while predicting 10 target classes.

The performance of Mod_v3 tells us that maintaining a healthy mix of sampled and original data is important for model prediction. Using a high sampling ratio can compromise model performance.

5 - Conclusion and future research scope

Machine Learning models are used for an innumerable number of tasks in today's time. The requirements of each model can be different and the performance metrics to look at change with the desired results. While one model may prioritise best performance scores, another may prioritise the ability to classify a larger set of data points or unique classes.

Given the nature of medical datasets being high in specificity, there are many use cases for a broader categorisation of classes. While sampling methods are limited in the ways they can enhance a machine learning model, there is plenty of scope for research in identifying how to extract the maximum value from medical datasets.

Future research on this exact study can look at Transfer learning as one of the most promising ways to enhance datasets usability.

References

[1] - Chaganti, R., Rustam, F., De La Torre Díez, I., Vidal Mazón, J. L., Rodríguez, C. L., & Ashraf, I. (2022). "Thyroid Disease Prediction Using Selective Features and Machine Learning Techniques." Cancers(Basel)