1. BUSINESS UNDERSTANDING

Business Overview

Hypothyroidism is a common condition where the thyroid does not create or produce enough thyroid hormone into a person's bloodstream. This makes the person's metabolism slow down. Also called underactive thyroid, hypothyroidism can make a person feel tired, gain weight and be unable to tolerate cold temperatures. It is important to note that hypothyroidism is more rampant in females than males.

Classifying whether or not a patient has hypothyroidism can help in reversing the trajectory of the thyroid disease, improve the symptoms and avoid lifelong consequences for the positive patients.

In this regard, Nairobi Hospital conducted a clinical camp to test for hypothyroidism. 18 different tests were conducted in all the patients during the camp. The data collected focused on thyroid patients.

Business Objectives

The main objective of this study was to correctly classify the patients who had hypothyroidism and those who were negative from the test results.

Business Success Criteria

The project will be considered a success if the patients are correctly classified as having hypothyroidism or not based on the symptoms.

Assessing the Situation

Resource inventory:

- Dataset: Can be found here.
- Software: Google collaboratory, Github.

Assumptions:

Provided data is sufficient, correct, and up-to-date.

Constraints:

There were no constraints.

Data Mining Goals

The data mining goals for this project were as follows:

- Identify the key test that can be used to detect hypothyroidism in patients given symptoms.
- Building a model using the available patient data to classify whether a patient has hypothyroidism or not.

Data Mining Success Criteria

The success criteria will be measured by targeting the most prevalent symptoms of hypothyroidism and the key test that can be conducted on patients to ascertain the assumption.

2. DATA UNDERSTANDING

Overview

For the project, the dataset that was used is the Nairobi Hospital data that was collected during the clinical camp that focuses on thyroid patients.

Data Description

The dataset has the following columns:
Status - indicates whether a patient has hypothyroidism or is negative
Age - represents the age of the patient
Sex - represents the gender of the patient
On_thyroxine - whether the patient is on thyroxine or not
Query_on_thyroxine - whether there was question on thyroxine or not
On_antithyroid_medication - whether a patient is on antithyroid medication or not
Thyroid_surgery - whether a patient has had a thyroid surgery or not
Query_hypothyroid
Query_hyperthyroid
Pregnant - whether a patient is pregnant or not
Sick - whether a patient is sick or not at the time of data collection
Tumor - whether a patient has a tumor or not
Lithium
Goitre - whether a patient has goitre or not
Tsh_measured - if TSH test has been conducted or not
Tsh - the TSH test result
T3_measured - if T3 test has been conducted or not
T3 - the T3 test result
Tt4_measured - if TT4 test has been conducted or not
Tt4 - the TT3 test result
T4u_measured - if T4U test has been conducted or not
T4u - the T4U test result
Fti_measured - if FTI test has been conducted or not

Fti - the FTI test result
Tbg_measured - if TBG test has been conducted or not
Tbg - the TBG test result

Verifying Data Quality
Missing data - these were coded as a non-response using the ? symbol.
There were no data errors in the dataset.
There were no measurement errors.
Standard units of measurements and value consistencies were used.
All the metadata had the apparent meaning as stated field name.

3. DATA PREPARATION

Step 1: Load Data
Loaded the dataset from a Google Drive CSV file and then created a data frame from it.

Step 2: Cleaning Data
The following steps were taken to clean the dataset:
i. Validity
No data was excluded from the dataset as they all proved useful to the data mining goals. Outliers were also spotted in the dataset. These, however, were not dropped as they proved useful during the analysis stage.

ii. Accuracy
I believe the data we were dealing with is correct therefore no action was taken in this step.

iii. Completeness
I checked and counted missing values and found that there were no missing values. However, upon close scrutiny, I realized that the missing data were coded as a non-response using the ? symbol. These were appropriately dealt with for each column.

iv. Consistency
In this step, I checked for any duplicates in the dataset and dropped all of them as they would seriously affect the analysis results.

v. Uniformity

For uniformity purposes, I changed the data types of integer/float columns to their appropriate data types and the rest remained as string. The integer/float columns are age, TSH, T3, TT4, T4U, FTI and TBG.

Since the ? symbol in the age column was replaced with 0 before converting to integer, the 0s were replaced by the median value of the different ages. The 0s in the other float columns that replaced the ? symbol were left that way.

Step 3: Data Transformation

The modeling techniques require that the features being used be in numerical format. This was achieved by creating dummy variables of the categorical columns. Another transformation done was standard scaling each attribute in the dataset, without which the classifier would depend much more on attributes which scale is larger than others.

4. ANALYSIS

Performing an analysis on the data using the linear discriminant analysis technique gave an accuracy score of 95.73% with one component and correctly classifying 16 hypothyroid patients and 590 negative patients. The whole analysis can be found in the github repository which I will attach the link to in the modeling section.

5. MODELING

Selecting the Modeling Techniques

This project focuses on a classification problem. In this regard, the following techniques were considered:

1. Decision trees

In this part, I did not use a single tree, rather, I used all the advanced models, that is, random forests, AdaBoost and gradient boost techniques.

2. Support Vector Machines

In this part, I applied polynomial, linear and rbf kernel functions to build the SVM model and then evaluated their performance and picked the kernel that performs the best.

Test Design

The data was divided into train and test sets and the test set used to test the criteria of the goodness of the model. In our case, the model's goodness of fit will be estimated using the mean squared error as I will be using supervised learning models. The hyperparameters were adjusted several times and the models re-ran to give optimal accuracy levels.

Building the Models

Parameter Settings

Decision trees - after testing different parameters on the models, a maximum depth of 5 and a minimum sample split of 20 with 100 trees gave an optimal accuracy score.
SVM - linear gave an optimal accuracy using gamma as auto and degree as 3, polynomial had the gamma set to 5 and a degree of 2 while RBF had its gamma parameter set to 5 as well.

Running the Models

This was pretty straight forward after plugging in the parameters for the different models. However, the text editor kept crashing while trying to view the graphs. The graphs were therefore omitted.

Model Description

The following are the results from the models used:

1. Decision trees

- Random forest - using all the test features, an accuracy score of 99.52% is achieved and 0 and 629 records are correctly classified under hypothyroid and negative classes respectively. After determining the most important features, an accuracy score of 99.05% is achieved, and 23 and 604 records are correctly classified under hypothyroid and negative classes respectively.
- AdaBoost - using the most important features, an accuracy score of 99.05% is achieved and 23 and 604 records are correctly classified under hypothyroid and negative classes respectively.
- Gradient boosting - using the most important features, an accuracy score of 99.05% is achieved and 23 and 604 records are correctly classified under hypothyroid and negative classes respectively.

2. SVM

The most important features were also used in this modeling technique. The results were as follows:

- Linear - an accuracy score of 97.47% is achieved and 16 and 601 records are correctly classified under hypothyroid and negative classes respectively.
- Polynomial - an accuracy score of 97.63% is achieved and 17 and 601 records are correctly classified under hypothyroid and negative classes respectively.

- RBF - an accuracy score of 95.89% is achieved and 1 and 606 records are correctly classified under hypothyroid and negative classes respectively.

6. RECOMMENDATION AND CONCLUSION

Given a set of symptoms that indicate hypothyroidism in patients, a physician would start by asking the patient the age and continue to conduct TSH, T3, TT4, T4U and FTI tests as these were some of the most important features in our data.

For classification using the SVM technique, the polynomial kernel function is highly recommended as it gives us the highest accuracy score of the three kernel functions.