Nairobi hospital hypothyroidism

Introduction

Today we will be analysing and predicting the levels of hypothyroidism in Nairobi Hospital. We will be using one dataset that is the hypothyroidism.csv dataset.

Business Understanding

Business Overview

Predict hypothyroidism in Nairobi Hospital.

Business objective

We would like to know if we can predict whether someone has hypothyroidism given their symptoms to see how we can best prevent it.

Business success criteria

Our business success criteria would be to find out if we are able to accurately predict if a client has hypothyroidism.

Data Handling description

Data Understanding

Our dataset is a csv file with 3163 rows and 26 columns. We used pandas to read this dataset, and we found the following columns namely:

- Age - This gives us the age of the clients
- Sex - This gives us the gender of the clients
- On_thyroxine -This indicates whether or not they are on thyroxine
- Query_on_thyroxine- This indicates whether or not they are on thyroxine
- on_antithyroid_medicationthyroid_surgery
- query_hypothyroid
- query_hyperthyroid
- Pregnant- This indicates whether they are pregnant or not
- Sick- This indicates whether they are sick or not
- Tumor- This indicates whether they have a tumor or not
- Lithium- Are they on medications containing lithium
- Goitre- This indicates whether they have goitre or not
- TSH_measured- This indicates whether they had TSH tested or not
- TSH - This is the amount of TSH found in their blood
- T3_measured- This indicates whether they had T3 measured or not
- T3- This is the amount of T3 found in their blood
- TT4_measured- This indicates whether they had TT4 measured or not
- TT4- This is the amount of T3 found in their blood

Data Cleaning Success Criteria

Eliminating all null values

Removing all duplicates

Data Cleaning procedure

Calculated the sum of the missing values. When looking at them they were too many to drop and this would affect our dataset. So I chose to fill the age column with 0,fill the sex column with unknowns , and fill the rest(since they were columns containing numbers) with the mean.

I then looked at the data types and noticed all the columns were objects but some had numbers inside them, so I decided to change them to their numeric form.

I proceeded to look at the number of duplicated rows, and found that there were 77 duplicated rows. I then proceeded to check them out and realised they had different values so I decided not to drop them as these were different records.

After this I changed the target variable from an object to numeric , this is because we would like to do analysis on this column as it is the target variable.
Next I dropped the columns that were not being used in the analysis which was 'query_hyperthyroid' since we are looking for hypothyroid we dropped this column.
I then proceeded to look for outliers and I noticed that age had no outliers and was slightly negatively skewed. I also noticed that TSH had outliers on the right hand side meaning most of them were larger than normal. However I noticed that T3, TT4,T4U, and FTI had outliers on both the left hand side and the right hand side. This means that there were values that were extremely low and some that were extremely high.
I then went to check out the anomalies and I found that all of the numeric columns had anomalies and when I summed them up 53% of the data. So I decided not to drop them as this would mean losing half of our data, and this can change stuff in analysis.

The analysis

In the analysis part we wanted to check the distribution of the numeric columns, and we found that the data is positively skewed. This was largely because of the TSH ,TBG , FTI ,and T3 columns. They had a large positive skew and hence influenced the negatively skewed which are the Status and Age columns.
We also discovered that the data has a heavy tail to the right hand side and this could be because of TSH,TBG,Status and the FTI columns that had the most kurtosis meaning they had the heaviest tails.
We also got to see that 95.2% of the people in this dataset have hypothyroidism. On further analysis we saw that of the people with hypothyroidism, 68.8% were female,28.9% were male and the rest were unknown. We also saw that 1.4% of the people with hypothyroidism are on medication which is alarming. We saw that 3.2% of the people with hypothyroidism are sick, and that 3.1% have goitre. We also saw that 1.3% of them have tumors and 2.1% are pregnant.
We also observed that TT4 and FTI are the most correlated to the status and we also saw that age, T4U and TBG have no correlation with the status.

Modelling procedure and results

We conducted the modeling in two parts:

Part 1

In part one I used decision trees random forest classifier to build my first model since this was a classification challenge . This model without any parameters specified gave me an root mean squared error of 0.13182400062431848.
I then decided to tune my parameters to get a better score, so to do this I used the random search CV because the grid search CV was taking too long. From the random search CV i got n_estimators =100, criterion='entropy', bootstrap=False. Using these parameters I got an RMSE of 0.12568925295997527 which is lower by about 0.01 that is slightly significant.
I then decided to boost my model using ada boosting method, and this gave me an RMSE of 0.13182400062431848 which is exactly the same as the initial one, meaning the weak learners were not contributing to the model.
For further investigation we wanted to see which columns are contributing the most so we did that and found FTI, TSH, T4U,T3,age,on_antithyroidmedicine,TT4 were the only ones contributing and FTI was the most with 0.66 and the rest were below 0.1.

Part 2

In part two we decided to select two features which are FTI and TSH since these are the ones which had the greatest impact in the first part. We used the Support Vector Machine to build our model. In the first model we built on the kernel linear, and we got an RMSE of 0.1589.
Then we moved to the second model where we built on the kernel poly which is short for polynomial , we chose the degree to be 3 because as our x increased the y decreased so we felt like degree 3 best fitted the model. We got an RMSE of 0.1589 which was the same as the linear .

Then we moved to the third model where we used the kernel of radial basis function. This gave us an RMSE of 0.14 which was the best compared to the rest.
After building the three models I decided to tune my parameters however this did not go into completion due to the RAM on my computer. However I am aware that the optimal kernel is rbf.
When asked to use all the features I used PCA to combine them and give two components. I then used these components in my analysis as well as the rbf kernel, and this gave me an RMSE of 0.195 which was the worst of them all.