Grupo 45

António Estêvão – Nº 58203 – XX hours.

Jacky Xu – Nº 58218 – YY hours.

Maria Rocha – Nº 58208 – ZZ hours.

Introduction and Goals

Brief Description of the Task

This project aims to analyze a version of the "thyroid0387" dataset (proj-data.csv) to develop effective classification models. The primary task is to classify subjects into one of eight target variable classes related to thyroid diagnoses, including six specific conditions, a class indicating no condition ('-'), and other diagnoses.

Data

The dataset includes various attributes related to thyroid conditions and other demographic information. The data.names file provides detailed descriptions of each attribute.

Goals

The objectives of this project are threefold:

1. **Classification (O1)**: Develop and compare the best possible classification models using methods such as Decision Trees, Naive Bayes, and Logistic Regression, with a preference for simpler configurations.
2. **Predictive Analysis (O2)**: Assess the ability to predict the age and sex of the subjects based on the available attributes.
3. **Feature Importance (O3)**: Identify the most significant features in the best-performing models from objectives O1 and O2.

Each of the following points will be divided by the analyses and experimentation done on each model.

Data Processing

To start the analysis of the data we first ran “from ydata_profiling import ProfileReport” on the data given, in this we saw that multiple columns had missing values, a lot of columns were skewed and age was highly skewed and record identification had only unique values.
Before data processing the classes available in the data were mapped according to the project guidelines (“Classification should be according to 8 classes of the target variable "diagnoses"). After this, a simple DTC model was created to keep checking how the different data processing steps affected it.
The “?” values were replaced by NAN in order to make it easier to detect them. By seeing age was highly skewed we realized there was a need to delete the rows in which the age was illogical, by this I mean ages <0 and >130.
We detected with the profile the need to drop record identification since these values were only for identification of the columns.
Since the models we worked with don’t accept nan values, we needed to imput those. Different imputers were tested but they mostly gave the same values, hence by this line in the project guidelines: “within each model, everything else being similar, the simplest configurations should be preferred.”, it was decided to use simpleImputer with the mode most frequent.

Examples in Naïve Bayes (Note: mean and average cannot be used since F Or M isn’t a value):

imputer : imputer = SimpleImputer(strategy='most_frequent')

The Accuracy is: 0.8004
The Precision is: 0.7605
The Recall is: 0.8004
The F1 score is: 0.7448
The Matthews correlation coefficient is: 0.4171

	precision	recall	f1-score	support
binding protein	0.83	0.19	0.30	54
discordant results	0.00	0.00	0.00	28
general health	0.56	0.07	0.12	72
healthy class	0.81	0.99	0.89	1102
hyperthyroid conditions	0.89	0.26	0.40	31
hypothyroid conditions	0.75	0.55	0.64	92
other class	0.62	0.24	0.34	34
replacement therapy	0.43	0.05	0.10	55
accuracy			0.80	1468
macro avg	0.61	0.29	0.35	1468
weighted avg	0.76	0.80	0.74	1468

imputer : imputer = KNNimputer()

The Accuracy is: 0.8004
The Precision is: 0.7645
The Recall is: 0.8004
The F1 score is: 0.7466
The Matthews correlation coefficient is: 0.4177

	precision	recall	f1-score	support
binding protein	0.77	0.19	0.30	54
discordant results	0.00	0.00	0.00	28
general health	0.56	0.07	0.12	72
healthy class	0.81	0.99	0.89	1102
hyperthyroid conditions	0.88	0.23	0.36	31
hypothyroid conditions	0.75	0.55	0.64	92
other class	0.62	0.24	0.34	34
replacement therapy	0.60	0.11	0.18	55
accuracy			0.80	1468
macro avg	0.62	0.30	0.35	1468
weighted avg	0.76	0.80	0.75	1468

It was detected that the columns with numerical values needed to be encoded since they were being interpreted as strings instead of numbers, to do so an ordinal encoder was used.
Since some of the models (DTC for example) only accept numerical values, it was necessary to encode the categorical values, first it was tested doing this with one hot encoder but in columns with t and f this created a

column for each and gave a value of true or false, since these created a lot of unnecessary columns it was decided to manually encode the true and false columns by making 1 == true and 0 false. The same was done for sex, making 1 being female and 0 male.
The nan values in numerical columns except age mean the test was not realized hence wasn’t measured, since DTC doesn’t accept nan, and naïve bayes doesn’t accept negative numbers, it was used a fillna with a ridiculous value that the decision tree could interpret as a different class.
Values from 999 to 999999 were tested, but since most did not show any significant difference, it was decided that we would use the value 999.

Examples in Logistical Regression:

data.fillna(valor, inplace=True) valor: 999

```
The bias is:  0.5785134159001989
The Accuracy is:  0.2732
The Precision is:  0.7523
The Recall is:  0.2732
The F1 score is:  0.2672
The Matthews correlation coefficient is:  0.2529
The Macro F1 Score is:  0.2826
```

	precision	recall	f1-score	support
binding protein	0.29	0.57	0.39	54
discordant results	0.03	0.29	0.06	28
general health	0.21	0.74	0.33	72
healthy class	0.93	0.14	0.25	1102
hyperthyroid conditions	0.16	0.87	0.26	31
hypothyroid conditions	0.26	0.65	0.38	92
other class	0.11	0.32	0.17	34
replacement therapy	0.28	0.96	0.43	55
accuracy			0.27	1468
macro avg	0.28	0.57	0.28	1468
weighted avg	0.75	0.27	0.27	1468

data.fillna(valor, inplace=True) valor: 999999

```
The bias is:  0.4070066003979032
The Accuracy is:  0.1921
The Precision is:  0.6908
The Recall is:  0.1921
The F1 score is:  0.1598
The Matthews correlation coefficient is:  0.1787
The Macro F1 Score is:  0.2180
```

	precision	recall	f1-score	support
binding protein	0.15	0.44	0.23	54
discordant results	0.05	0.39	0.09	28
general health	0.17	0.68	0.27	72
healthy class	0.87	0.07	0.13	1102
hyperthyroid	0.14	0.68	0.24	31

conditions				
hypothyroid conditions	0.11	0.35	0.17	92
other class	0.13	0.35	0.19	34
replacement therapy	0.27	0.96	0.42	55
accuracy			0.19	1468
macro avg	0.24	0.49	0.22	1468
weighted avg	0.69	0.19	0.16	1468

Notes:
It was decided not to use scaling in Naive Bayes as Naive Bayes relies on the assumption that all features are conditionally independent given the class label. Scaling the features doesn't typically improve or align with this independence assumption. The independence assumption is more critical to the performance of Naive Bayes than the scale of the features.
Undersampling was not used in the models because it was decided that the possible consequences (loss of information, reduced model performance, increased variability, imbalanced subsampling, etc.) did not outweight the possible benefits.

Variable Selection

For feature selection it was first used Spearman correlation to visually detect the variables which had most correlation either negative or positive, and after that we used the sequential feature selection since this chooses the features that influence positively the model. For each model (target=diagnoses, age, sex) we tested them with all columns selected, 30, 20 and 10 (and without variable selection) and the model which gave the best value was selected.

Model Results

DecisionTreeClassifier

Diagnoses

N-Fold Cross Validation Results:
Precision: 0.8921734990287876
Recall: 0.8931492842535788
F1 Score: 0.8923484389769444
Matthews Correlation Coefficient: 0.7599140687987704
Balanced Accuracy: 0.7315090911427939
MSM: 0.7315090911427939
GMTR: 0.7184169366004695
Macro F1 Score: 0.7343603875476511

	precision	recall	f1-score	support
binding protein	0.63	0.65	0.64	230
discordant results	0.64	0.68	0.66	135
general health	0.82	0.89	0.85	274
healthy class	0.95	0.95	0.95	4324
hyperthyroid conditions	0.71	0.60	0.65	111
hypothyroid conditions	0.78	0.78	0.78	384
other class	0.61	0.51	0.56	189
replacement therapy	0.77	0.78	0.77	221

accuracy			0.89	5868
macro avg	0.74	0.73	0.73	5868
weighted avg	0.89	0.89	0.89	5868

Sex
N-Fold Cross Validation Results:
Precision: 0.6516213679583482
Recall: 0.6876278118609407
F1 Score: 0.6442442707197173
Matthews Correlation Coefficient: 0.1697323614241098
Balanced Accuracy: 0.5628939773230264
Macro F1 Score: 0.5548252814334137

	precision	recall	f1-score	support
0	0.44	0.16	0.24	393
1	0.75	0.93	0.83	1075
accuracy			0.72	1468
macro avg	0.60	0.54	0.53	1468
weighted avg	0.67	0.72	0.67	1468

Classification Report for N-Fold Cross Validation:

	precision	recall	f1-score	support
0	0.51	0.22	0.31	1855
1	0.72	0.90	0.80	4013
accuracy			0.69	5868
macro avg	0.61	0.56	0.55	5868
weighted avg	0.65	0.69	0.64	5868

Age
Best Hyperparameters: {'max_depth': 5, 'min_samples_leaf': 4, 'min_samples_split': 2}
Average Evaluation Metrics across Folds:
Average Mean Squared Error: 311.5569131360579
Average Mean Absolute Error: 14.582066759124094
Average R-squared Score: 0.12313342931671956

Hyperparameter Tuning

For hyperparameter tunning it was used GridSearchCV, defining different parameters to test.

Discussion and Conclusions

Objective 1: Classification Models

The Decision Tree Classifier achieved an overall accuracy of 89%, indicating that the Decision Tree Classifier is highly effective. The model's high values suggest that it is both accurate and reliable in its predictions, minimizing both false positives and false negatives.
In comparison, the Naive Bayes model and Logistic Regression model performed less effectively. The Naive Bayes model achieved an accuracy of 80%, which, while reasonable, fell short of the Decision Tree's performance. Logistic Regression, on the other hand, struggled significantly with an accuracy of only 27%, indicating that it is not well-suited for this multi-class classification problem.
The robustness of the Decision Tree Classifier can be attributed to its ability to handle the complexity of the dataset, including interactions between different features and the non-linear relationships inherent in the data.

Objective 2: Predictive Analysis

Age Prediction

Predicting the age of subjects based on the available attributes proved to be a challenging task. The Decision Tree Regressor used for this task resulted in a low R-squared score of 0.123, indicating that the model could not explain much of the variance in the age data. This low score suggests that the features in the dataset do not have a strong relationship with age, making it difficult to predict age accurately.

This difficulty may be due to the nature of thyroid-related attributes, which can manifest at any age, leading to a weak correlation between these attributes and the age of the subjects. Thus, we conclude that predicting age confidently using this dataset is not feasible.

Sex Prediction

In contrast, the prediction of the subject's sex based on the available attributes showed more promising results. The Decision Tree Classifier used for this task achieved a balanced accuracy of approximately 69%, with a macro F1 score of 0.554. While these results are not exceptionally high, they indicate a reasonable level of predictability for sex.

The analysis revealed that certain features, such as 'pregnant' and 'query on thyroxine,' are significant predictors of sex. This makes sense biologically, as pregnancy is a female-specific attribute, and thyroid disorders, which are more prevalent in females, are often monitored and treated differently in males and females. The 'query on thyroxine' feature also stood out due to its higher incidence in women, who are more prone to thyroid conditions.

Therefore, while not perfect, the model shows a fair degree of accuracy in predicting sex, making it a viable task using this dataset.

Objective 3: Feature Importance

Diagnoses

For the classification of thyroid diagnoses, the top features identified include: Pregnant, Thyroid surgery, I131 treatment, Hypopituitary, TSH measured, TSH, TT4 measured, TBG measured, TBG, Referral source (WEST).

Sex

For predicting the sex of the subjects, the significant features identified were: Query on thyroxine, Pregnant, Thyroid surgery, Lithium, Goitre, Hypopituitary, TT4 measured, Diagnoses, Referral source (STMW), Referral source (SVHC).

These features include both direct indicators of sex (e.g., pregnancy) and related medical conditions and treatments that show different prevalence between sexes.

Age

For predicting age, the significant features identified through Sequential Feature Selection (SFS) include: I131 treatment, Goitre, Psych, TSH measured, T3 measured, T3, Referral source (STMW), Referral source (SVHD), Referral source (SVI), Referral source (WEST).

These features, while significant, collectively did not provide a strong predictive power for age, reflecting the complexity and variability of age-related patterns in the dataset.

Conclusion

In conclusion, the Decision Tree Classifier stands out as the best model for classifying thyroid diagnoses, providing high accuracy and reliability. Predicting sex is moderately successful, leveraging specific thyroid-related attributes and demographic factors. However, predicting age remains challenging due to weak correlations between the features and age. Feature importance analysis has highlighted the most influential factors for each prediction task, offering valuable insights into the dataset's structure and informing future improvements to the models.