Group 43

Pedro Gomes – 58167 – 10 hours

Pedro Marques – 48674 – 10 hours

# Introduction and Goals

Having a dataset that comprises a version of an archive of thyroid diagnoses

obtained from the Garvan Institute, consisting of 6420 records from 1984 to

early 1987, and comprised of a list of subjects, each with 29 varied attributes connected with a "diagnoses" attribute;

Our task was as follows, delineated in 3 sub-objectives:

**O1:** Find the best possible classification models for this dataset, with the classification classes consisting of the following, based on the "diagnoses" attribute (in parenthesis):

hyperthyroid conditions (A, B, C, D)

hypothyroid conditions (E, F, G, H)

binding protein (I, J)

general health (K)

replacement therapy (L, M, N)

discordant results (R)

healthy (-)

other (all other cases fall into this class)

**O2:** Two of the attributes are Age and Sex of the subject. We should decide if we can confidently predict these attributes (separately) given the other attributes.

**O3:** We should collect the most significant features in the best models obtained in objectives 1 and 2. These would be up to 3 distinct models, one for classification, and two for predicting the Age and the Sex of the subject.

# Data Processing

The data is comprised of a total of 31 columns:

['sex', 'on thyroxine', 'query on thyroxine', 'on antithyroid medication', 'sick', 'pregnant', 'thyroid surgery', 'I131 treatment', 'query hypothyroid', 'query hyperthyroid', 'lithium', 'goitre', 'tumor', 'hypopituitary', 'psych', 'referral source', 'diagnoses'], and an extra column "record identification" which is the individual number identifying the specific record. This last column will not be used in the models, as it is irrelevant for classification.

The 'referral source' column is also dropped, as it is also irrelevant to the classification.

The 'diagnoses' column is out target for classification.

Here are the steps we took when preprocessing the data, before feeding it to the models:

1. Remove trailing colons from column names;
2. Replace non-numeric values with "NaN";
3. Replace binary "true/false" columns with 1/0 numerical values;
4. Convert appropriate columns to numeric values, nominally ['age', 'TSH', 'T3', 'TT4', 'T4U', 'FTI', 'TBG'];
5. Drop the record identification column from the data (we preserve it elsewhere but do not feed it to the models;
6. We then ensure that all the remaining data in the dataset is numeric;
7. Then we define the mapping for new classes, as outlined in O1, and map the 'diagnoses' to these new classes. We also encode the new class labels;
8. The data is now ready to be used in the models.

# Variable Selection

We examined the dataset to identify the continuous and discrete features and also took care of excluding the irrelevant features, like the 'referral source' since it was not relevant for none of the 3 prediction models we did, as well as the record identification, which we removed but saved it in a variable for future identification.

We proceeded to encode categorical variables to ensure all the features were numeric and suitable for modeling, we also removed outliers from the age variable to ensure a cleaner and more suitable dataset for training.

The continuous features we had were 'TSH, 'T3', 'TT4', 'T4U', 'FTI', 'TBG' The discrete features we identified were, 'sex', 'on thyroxine', 'query on thyroxine', 'on antithyroid medication', 'sick', 'pregnant', 'thyroid surgery', 'I131 treatment', 'query hypothyroid', 'query hyperthyroid', 'lithium', 'goitre', 'tumor', 'hypopituitary', 'psych', 'TSH measured', 'T3 measured', 'TT4 measured', 'T4U measured', 'FTI measured', 'TBG measured', 'diagnoses'

# Model Results

O1- For classification we used a “Decision Tree Classifier” model:

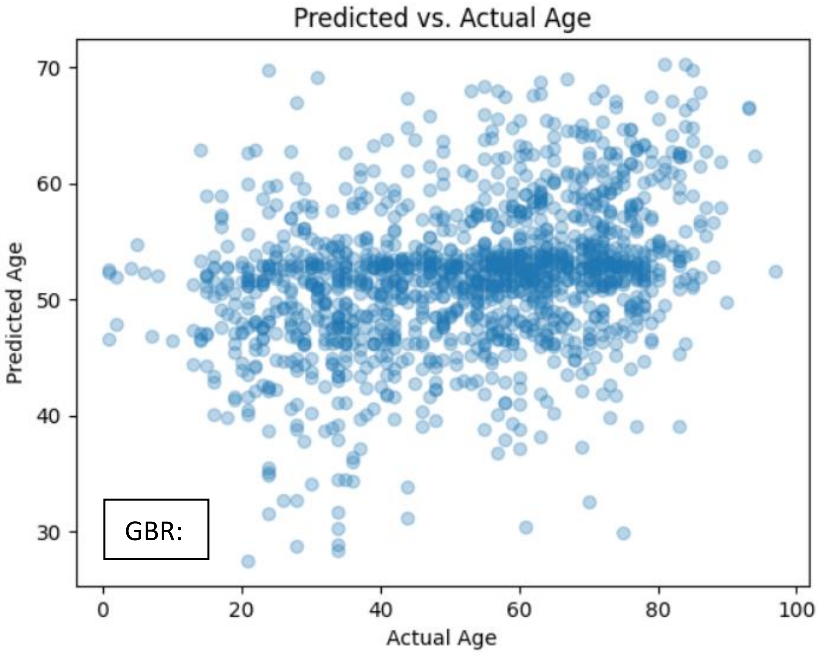
| classes | precision | recall | f1-score | support |
|---|---|---|---|---|
| binding protein | 0.70 | 0.71 | 0.70 | 55 |
| discordant results | 0.79 | 0.71 | 0.75 | 31 |
| general health | 0.79 | 0.83 | 0.81 | 72 |
| healthy | 0.97 | 0.96 | 0.97 | 1114 |
| hyperthyroid conditions | 0.75 | 0.72 | 0.73 | 25 |
| hypothyroid conditions | 0.93 | 0.98 | 0.95 | 81 |
| other | 0.77 | 0.82 | 0.79 | 33 |
| replacement therapy | 0.95 | 0.93 | 0.94 | 57 |
| macro avg | 0.83 | 0.83 | 0.83 | 1468 – total support |
| weighted avg | 0.93 | 0.93 | 0.93 | 1468 – total support |

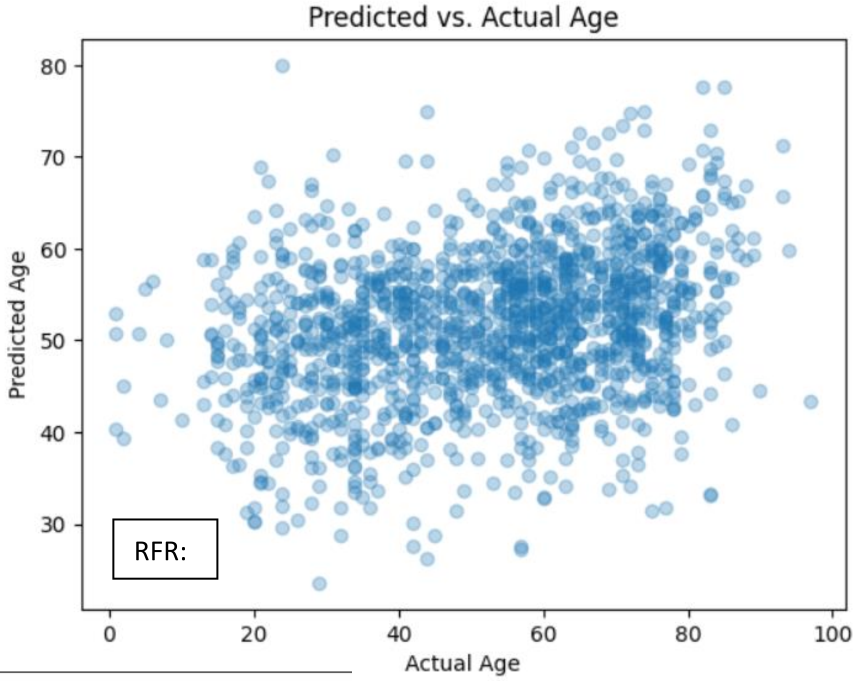
| | | |
|---|---|---|
| accuracy | 0.93 | 1468 – total support |

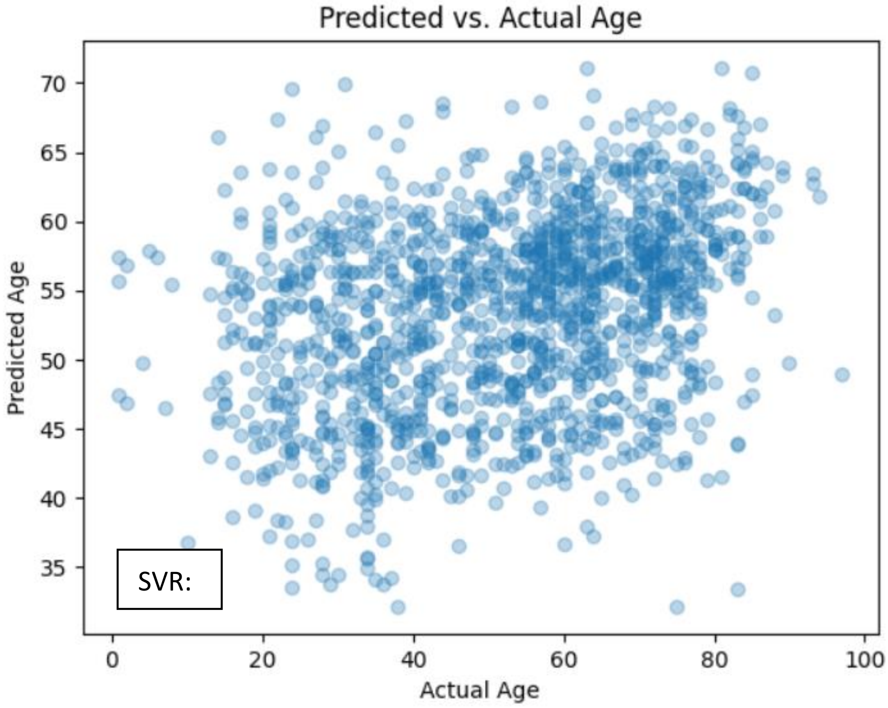
O2- For predicting Ages we tested 3 models, Gradient Boosting Regressor (GBR), Random Forest Regressor (RFR), and SVR.

| Model | MSE | RMSE | Accuracy | R2 |
|---|---|---|---|---|
| GBR | 317.062 | 17.806 | 0.108 | 0.108 |
| RFR | 332.903 | 18.245 | 0.063 | 0.063 |
| SVR | 320.178 | 17.893 | 0.099 | 0.099 |

Here are the graphs of predicted age versus real age:


Predicted vs. Actual Age
Predicted Age
Actual Age
70
60
50
40
30
0
20
40
60
80
100
GBR:


Predicted vs. Actual Age
Predicted Age
Actual Age
80
70
60
50
40
30
0
20
40
60
80
100
RFR:


Predicted vs. Actual Age
Predicted Age
Actual Age
70
65
60
55
50
45
40
35
0
20
40
60
80
100
SVR:

For sex prediction, we used a Random Forest Classifier. Here are the statistics:

MSE: 0.379

RMSE: 0.615

Accuracy score: 0.685

# Conclusion

**O1-** The model we used, a Decision Tree Classifier, was able to reliably predict the class of a report, with an average accuracy of around 93%.

**O2-** Predicting the age of the subjects using the other attributes proved to be unreliable. Of all the models we tested, the best one was the Gradient Boosting Regressor, sporting only an accuracy of around 10%, and a mean squared error of over 317.

Predicting the sex of the subjects, however, was considerably more reliable, with our Random Forest Classifier model achieving an accuracy of around 70%, and a mean square error of 0,379.

**O3-** in O1, our best model (DTC) had these as the most important features, with these importances:

| | |
|---|---|
| TSH | 0.259567 |
| FTI | 0.189511 |
| T3 | 0.181832 |
| T4U | 0.086070 |
| TT4 | 0.079606 |

In O2, our best model (GBR) for age had these as the most important features, with these importances:

| | |
|---|---|
| T3 | 0.368265 |
| T4U | 0.147864 |
| TSH | 0.092270 |
| FTI | 0.057891 |
| TSH measured | 0.052879 |

Our best model for sex (RFC) had these as the most important features, with these importances:

```
age    0.161265

T4U    0.150866
TT4    0.145415
TSH    0.141446
FTI    0.133971
```