Aprendizagem Automática 2022/2023

First Home Assignment

Group number: 36

Student names and ids:

Luís Pires 54471,

Inês Martins 54462,

Miłosz Włoch 59437

How many hours each student contributed to the project:

Luís Pires: 10 hours

Inês Martins: 10 hours

Miłosz Włoch: 10 hours

—–---

Aims of the assignment:

- Providing the best possible regression and classification models using **Decision Trees and Linear models**
- Examining different hyperparameters and selecting the best one considering its simplicity as well
- Full testing and validation cycle of model selection and evaluation

Preprocessing:

Before we selected the best model, we had to perform a variable selection. We excluded 'subject#' column as it gives as only ID of examined person and 'age', 'sex', 'test_time' because they give us bias based on what we were able to examine from the input data. Because of the fact that 'total_UPDRS' and 'motor_UPDRS' are dependent values, they have not been included in X (independent values) as well. According to p-values, AIC and BIC statistics we selected 7 columns that are: 'Jitter(Abs)', 'Shimmer:APQ5', 'Shimmer:APQ11', 'NHR', 'HNR', 'DFA', 'PPE' of 22 all columns. Only these were important to train and validate our models.

Objective 1:

- ### A summary of the performed work

  We decided to use the Decision Tree regression model using criterion 'squared error' to train the Parkinson data that were used in the assignment. We examined different hyperparameters to choose the best model of all. Every model used N-fold cross validation to validate. We decided to choose a split value of 10 as it gave us some prominent results in comparison to other values of split. Our data were divided into train and test data with a ratio of 80% to 20%. We plotted the results of statistics as: 'RVE', 'RMSE', 'Pearson correlation', 'Maximum Error', 'Mean absolute error' to make up our minds about which model would be the most appropriate. This way it was easier to notice how values changed over increasing than to look at the raw results.

- ### A table with models tested and relevant statistics

  We examined different hyperparameters as min_samples_leaf and max_depth. These are the results of examined statistics for models:

min_samples_leaf	RVE	RMSE	Correlation	Max Error	MAE
1	**0.142**	**7.507**	**0.390**	**24.641**	**6.126**
2	**0.143**	**7.507**	**0.390**	**24.641**	**6.125**
3	**0.143**	**7.507**	**0.390**	**24.641**	**6.125**
4	**0.144**	**7.503**	**0.391**	**24.641**	**6.122**
5	**0.144**	**7.502**	**0.391**	**24.641**	**6.123**
6	**0.145**	**7.497**	**0.392**	**24.641**	**6.118**
7	**0.145**	**7.496**	**0.392**	**26.045**	**6.120**

8	**0.145**	**7.497**	**0.392**	**26.045**	**6.121**
9	**0.145**	**7.497**	**0.392**	**26.045**	**6.121**
10	**0.146**	**7.498**	**0.392**	**26.045**	**6.121**
11	**0.146**	**7.493**	**0.393**	**26.045**	**6.115**
12	**0.145**	**7.494**	**0.392**	**26.045**	**6.115**
13	**0.145**	**7.496**	**0.392**	**26.045**	**6.117**
14	**0.145**	**7.498**	**0.391**	**26.045**	**6.122**
15	**0.145**	**7.495**	**0.392**	**26.045**	**6.121**
16	**0.146**	**7.492**	**0.392**	**26.045**	**6.118**
17	**0.149**	**7.480**	**0.396**	**26.045**	**6.105**
18	**0.146**	**7.490**	**0.396**	**26.045**	**6.117**
19	**0.148**	**7.485**	**0.394**	**26.045**	**6.111**
20	**0.147**	**7.487**	**0.394**	**26.045**	**6.113**

max_depth	**RVE**	**RMSE**	**Correlation**	**Max Error**	**MAE**
1	**0.017**	**8.037**	**0.137**	**19.567**	**6.866**
2	**0.058**	**7.865**	**0.246**	**20.444**	**6.688**
3	**0.115**	**7.627**	**0.341**	**21.629**	**3.394**
4	**0.129**	**7.565**	**0.366**	**24.871**	**6.262**
5	**0.142**	**7.507**	**0.390**	**24.641**	**6.123**
6	**0.146**	**4.493**	**0.407**	**26.585**	**6.004**
7	**0.127**	**7.577**	**0.408**	**27.103**	**5.952**
8	**0.091**	**7.791**	**0.400**	**28.604**	**5.997**
9	**0.025**	**8.005**	**0.380**	**30.243**	**6.131**
10	**-0.0287**	**8.222**	**0.370**	**30.913**	**6.224**

For max depth that is between 10 and 20 the values for RVE are negative and were discarded.

Discussion and conclusion

In this objective we understood the workings of linear regression and the relationship between dependent and independent values in such models. In the case of conducted research we chose the decision tree regressor model with min_samples_leaf that equals 17 and max_depth of 7. It is clear that the pruning of the decision tree regressor was necessary to avoid overfitting.

The low values for RVE were strange. Even though very high RVEs would suggest overfitting, such a low value suggests other problems with our model or the data chosen for prediction. In contrast, the values for MAE and RMSE were low which indicates that while our model wasn't the best at predicting the exact values for UPDRS, it didn't deviate too much. Because of the studies, we can also expect that the data are too imbalanced and as a result difficult to predict.

The best model statistics are:

The RVE is: 0.15059807315844986

The rmse is: 7.572738020898977

The Correlation Score is is: 0.4119 (p-value=2.443828e-49)

The Maximum Error is is: 23.656204999999996

The Mean Absolute Error is: 6.0293854625261485

Objective 2:

- ### A summary of the performed work

We decided to use the Decision Tree classifier model using criterion 'gini' to train the data used in the assignment. The dependent value needed to be converted into two unique values as we were expected to do the binary classification. What else we had to do in the preprocessing phase was to check the condition of the 'total_UPDRS' column. If any value of this variable was greater than '40' we assigned '1' as a label (positive) , otherwise it was '0' (negative). We examined different hyperparameters to choose the best model of all as it was in the first objective. Every model used N-fold cross validation to validate. We decided to choose a split value of 20 as it gave us some prominent results in comparison to other values of split. Our data was divided the same way as before. We plotted the results of statistics that are: 'Precision', 'Recall', 'F1 score' and 'Matthew's correlation coefficient''. We decided that they were going to give us enough information for choosing the right model. Although the second task was about classification the process of it was very similar to the first task with some minor changes.

A table with models tested and relevant statistics

min_samples_leaf	Precision	Recall	F1	MCC
1	0.869	0.861	0.866	0.224
2	0.873	0.834	0.853	0.221
3	0.868	0.865	0.867	0.218
4	0.870	0.857	0.863	0.221
5	0.866	0.875	0.870	0.213
6	0.868	0.869	0.869	0.221
7	0.868	0.893	0.881	0.241
8	0.869	0.894	0.882	0.248
9	0.867	0.903	0.885	0.244
10	0.866	0.904	0.885	0.239
11	0.866	0.909	0.887	0.242
12	0.864	0.912	0.888	0.235
13	0.865	0.922	0.892	0.249
14	0.863	0.922	0.892	0.241
15	0.865	0.927	0.894	0.255
16	0.863	0.925	0.893	0.249
17	0.861	0.931	0.895	0.242
18	0.862	0.933	0.896	0.248
19	0.862	0.935	0.897	0.252
20	0.863	0.939	0.899	0.267

max_depth	Precision	Recall	F1	MCC
1	0.831	1.000	0.907	0.000
2	0.831	1.000	0.907	0.000
3	0.831	0.999	0.907	0.045
4	0.843	0.978	0.905	0.172
5	0.843	0.983	0.908	0.184

6	**0.846**	**0.981**	**0.908**	**0.206**
7	**0.847**	**0.968**	**0.903**	**0.187**
8	**0.850**	**0.957**	**0.900**	**0.199**
9	**0.852**	**0.949**	**0.898**	**0.198**
10	**0.852**	**0.941**	**0.894**	**0.191**
11	**0.854**	**0.919**	**0.885**	**0.178**
12	**0.856**	**0.912**	**0.883**	**0.191**
13	**0.859**	**0.896**	**0.877**	**0.193**
14	**0.860**	**0.891**	**0.875**	**0.196**
15	**0.865**	**0.877**	**0.871**	**0.210**
16	**0.868**	**0.875**	**0.871**	**0.223**
17	**0.865**	**0.874**	**0.870**	**0.212**
18	**0.867**	**0.866**	**0.866**	**0.215**
19	**0.866**	**0.864**	**0.865**	**0.207**
20	**0.868**	**0.863**	**0.865**	**0.216**

• Discussion and conclusion

In this objective we understood the workings of the decision tree classifier model. For this objective the model chosen was a decision tree classifier with min_samples_leaf = 20 and max_depth = 6. Here we also find pruning to be necessary to avoid overfitting and to provide the best results.

In contrast to what we get in the regression, in the classification we receive some very good results. It can be said that our model is successful in predicting positive or negative cases of the disease. Even though the data is imbalanced our best model has almost 83% accuracy. The Mathew's correlation coefficient was not taken into account because the data was not balanced enough.

The best model statistics are:
The Precision is: 0.8322
The Recall is: 0.9886
The F1 score is: 0.9037
The Matthews correlation coefficient is: 0.1989
The Accuracy is: 0.8272