**Aprendizagem Automática 23/24**

Second Home Assignment

Grupo 33

Autores: Guilherme Cepeda – 62931 horas: 20

Guilherme Rosário – 62543 horas: 0

Marco Viana – 62550 horas: 8

Introduction falar dos dados

Objective 1

The primary objective of this work was to create the best regression model to predict the variable **critical\_temp**. Before building a model, the priority was to understand the type of data contained in the dataset and assess whether it was necessary to remove any variables. We realized that the data was already cleaned, and all the 81 columns were needed to enhance the predictive power. This dataset was collected with the goal of predicting the **critical\_temp** variable values} using the 81 features from superconductors.

The algorithm we chose to create the regression model was CART (Classification and Regression Trees), **using the mean squared error (MSE) as the impurity metric**. Since it's a ***decision tree-based approach***, we did not need to perform variable scaling. We started by splitting the dataset into two parts: 75% of the data was used to train and evaluate the models, and the remaining 25% was used to create the **independent validation set (IVS).** This separation is crucial when we don't have dedicated data to test the model, as the IVS simulates the model's performance in real-world scenarios where it would be used.

To train a good model that fits the data well and can predict our variable of interest (critical\_temp), we tried to understand which hyperparameters could be useful and better suited to our case. To test different models and have a basis for comparison, we selected the stopping criteria of **maximum depth (max\_depth) and the minimum number of samples per leaf node (min\_samples\_leaf).**

We opted for the k-fold cross-validation method to assess the model quality. The dataset was divided into ten partitions (k = 10), with each fold serving as a testing set while the remaining data was used for training. Various standard performance metrics were employed to evaluate the models, with a primary focus on the Pearson correlation coefficient, which measures the linear relationship between actual and predicted values. This was complemented by an examination of mean absolute error (MAE), root mean squared error (RMSE), and maximum error (ME). These error-related metrics are of paramount importance in industry-related applications, where precision in predicting values associated with a specific outcome is a key objective.

As mentioned earlier, the trained models differed in the choice of hyperparameters and their tuning. To determine which model was most likely to have superior performance in predicting the IVS data and, consequently, a more generalized application, we conducted an iteration where we slightly varied the hyperparameter values and compared the scores between the training set and the test set. The main goal of this approach was to select models among all hyperparameter values that did not exhibit overfitting or underfitting. Thus, we varied max\_depth from 1 to 28 in increments of 3, resulting in 10 different models, and we also varied min\_samples\_leaf from 1 to 55 in increments of 6, generating another 10 models. These models were evaluated with both the training set and the test set, and we compared the Pearson correlations in a graph assessing the models' performance as we increased max\_depth (Figure 1) or min\_samples\_leaf (Figure 2).

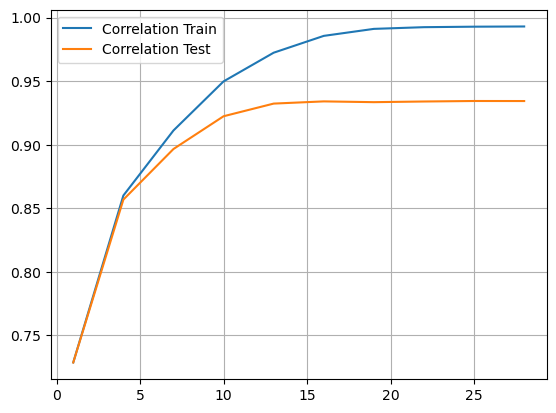
****

Figure 1 - Graph of Pearson correlation for the test set and the training set at different max\_depth values.

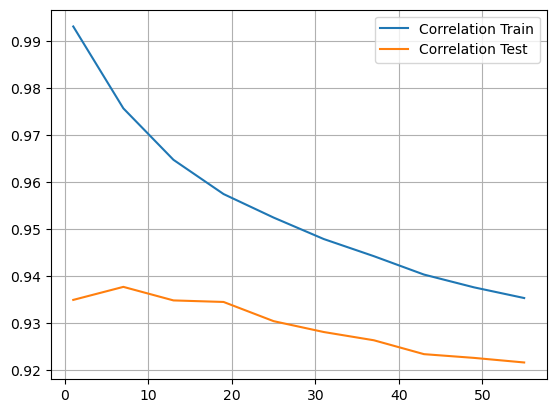
****

Figure 2 - Graph of Pearson correlation for the test set and the training set at different min\_samples\_leaf values.

After analyzing Figure 1, we can conclude that between the values max\_depth = 7 and max\_depth = 13, the correlation values of the test set begin to stabilize. There is no further benefit in increasing this hyperparameter beyond that point, as it could risk overfitting the model. However, in Figure 2, the test set's correlation line exhibits a downhill scenario in the beginning and starts to stabilize in the end proving that there is no further benefit in increasing the hyperparameter. Nevertheless, we chose to focus on values within the range of min\_samples\_leaf = 13 to min\_samples\_leaf = 25, resulting in a total of 6 models for comparison. The summary of these candidate models and their corresponding statistics is detailed in Table 2.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Val hyperparameter** | **RVE** | **RMSE** | **Correlation Score** | **ME** | **MAE** |
| **max\_depth 7** | **0.803** | **15.143** | **0.896** | **184.35** | **9.75** |
| **max\_depth 10** | **0.85** | **13.232** | **0.922** | **168** | **7.933** |
| **max\_depth 13** | **0.867** | **12.42** | **0.932** | **184.35** | **6.944** |
| **min\_samples\_leaf 13** | **0.873** | **12.177** | **0.934** | **172.196** | **7.036** |
| **min\_samples\_leaf 19** | **0.872** | **12.198** | **0.934** | **172.635** | **7.203** |
| **min\_samples\_leaf 25** | **0.865** | **12.554** | **0.93** | **180.364** | **7.467** |

**Table 1 - Summary of performance statistics for the 6 candidate models.**

**Considering that these 6 models have already been evaluated based on Pearson correlation scores, Table 1 includes various other performance statistics of the models that can now be used for our final selection. Looking at the RVE and RMSE values exhibit very slight variations. However, the model with max\_depth = 10 stands out as it has the lowest maximum error (168), which is a value we aim to minimize. Finally, the hyperparameter that instills the most confidence in producing the best model based on the tested conditions is max\_depth = 10. A new model with this hyperparameter was trained using the entire dataset and tested with the IVS, resulting in the statistics included in Table 2.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Best Model hyperparameter** | **RVE** | **RMSE** | **Correlation Score** | **ME** | **MAE** |
| **max\_depth 10** | **0.845** | **13.568** | **0.919** | **125** | **8.102** |

**Table 2 – Best model performance in the validation with IVS.**

**We can conclude that the selected model demonstrates the potential to predict the variable of interest. This is supported by its performance with the IVS data, which closely aligns with the model's performance during the testing phase and even reduces the ME. As a result, the model doesn't seem to suffer from overfitting or underfitting, showcasing effective generalization to new data.**

Objective 2