Aprendizagem Automática
First Home Assignment
Group 39

António Casimiro Nº56685 (9 horas)| Bruno Lima Nº54466 (9 horas)
| João Vieira Nº45677(9 horas) | Ricardo Ramos Nº56656 (9 horas)| TP11

Professors: André Falcão | Sofia Teixeira

October 2023

Contents

1 Task

For this first home assignment, we were tasked with providing the best regression and classification models using Decision Trees and Linear models in the Superconductivity dataset, created by Kam Hamidieh. This dataset contains 81 features extracted from 21174 superconductors along with the respective critical temperature. The goal here is to predict the critical temperature based on the features extracted.

2 Objective 1 - Producing the Regression Models

Our first objective in this assignment was to produce the best regression model that predicts correctly the critical temperature of several semiconductors. With this objective in mind four models were used, the Decision Tree regression, the Linear regression, the Ridge regression and the Lasso Regression.

Firstly, we installed all the necessary libraries to create and test our models were installed and a dataframe with our dataset was created. After that we separate the independent variables from the dependent ones and start and split the data into training and testing sets, with 20% for test.

The first regression model we started to work with was the Decision Tree, where we tested the **max depth** as our stopping criteria and trough trial and error we compared several statistics to find the best model, such as Ratio of Explained Variance (RVE), Root Mean Squared Error (RMSE), Pearson Correlation Coefficient, Mean Absolute Error (MAE) and the Maximum Error (ME). The model was also validated through k-Fold cross validation with k equal to 5 and with the use of an independent validation set (IVS). We concluded that the best Decision Tree was the one with a **max depth** of 10 and we can observe the respective plot of predicted vs. actual values in the figure 1.

After that we worked with the Linear Regression model where we once again calculated all the statistics that were calculated previously with the Decision Tree. In the figure 2 we have the respective plot of predicted vs. actual values.

Last but not least we have the Ridge and Lasso regressions. To find the best regressions we tested the models several times with different alphas and saved the alpha that gave better results. In this case alpha was equal to 5 in the Ridge Regression and 0.2 in the Lasso Regression. With the alpha values defined the same statistic calculation were made. The respective plot of predicted vs. actual values can be found in figure 3 and 4.

To find the best regression model we went ahead and compared all the statistics that we obtained for the different models in the table 1.

As we can observe trough the statistic used to evaluate the regression models the best model is the Decision Tree due to having the RVE closest to 1.

Looking at all the plots of predicted vs. actual values we notice that the one that is closest to the a perfect prediction (45° red line) is also the Decision Tree model.

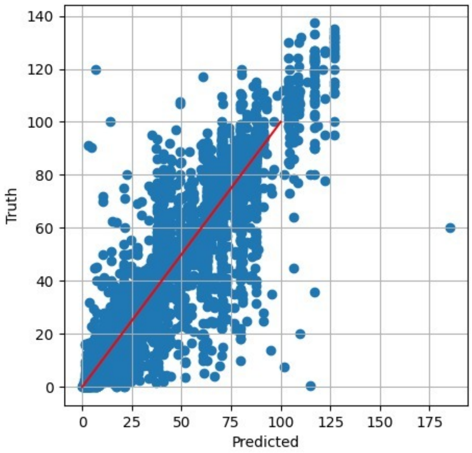

Figure 1: Predict-Truth plot for Decision Tree Regression;

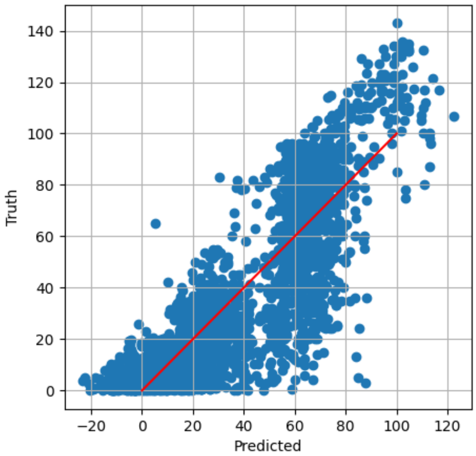

Figure 2: Predict-Truth plot for Linear Regression;

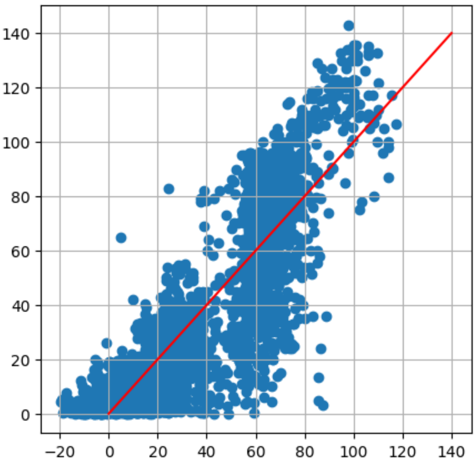

Figure 3: Predict-Truth plot for Ridge Regression;

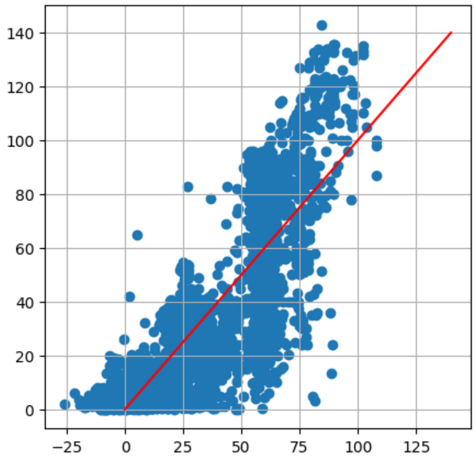

Figure 4: Predict-Truth plot for Lasso Regression;

Finally, we can conclude that the best regression model is Decision Tree due to statistic used to evaluate the regression models and the comparison with the perfect prediction.

3 Objective 2 - Producing the Classification Models

For the second objective we were tasked with producing the best binary classification model assuming as positive all instances with critical temperatures bigger or equal to 80. Naturally all other cases were considered as negative. The models used to this end were, the Decision Tree classification and the Logistic regression.

Once again we started by splitting the data into training and testing sets, with 20% for test, and, this time, splitting our dependent variable into positives and negatives as explained earlier.

	Decision Tree	Linear Regression	Lasso Regression	Ridge Regression
RVE	0.9170	0.7529	0.7077	0.7502
RMSE	9.957	17.18	18.69	17.27
Pearson	0.9576	0.8679	0.8423	0.8664
ME	71.80	84.91	78.48	84.45
MAE	6.142	13.26	14.52	13.33

Table 1: Statistics used to evaluate the regression models;

The first model we started to work with was the Logistic Regression. The confusion matrix on the figure 2. We kept a record of several statistics so that we could compare this model with the Decision Tree such as Accuracy, Precision, Recal, F1 Score and Matthews Correlation Coefficient (MCC).

Following that we started working on a classification model using Decision Tree where we choose the stopping criteria as minimum samples per leaf, with a value equal to 5. As always we recorded the necessary statistics for evaluating the model and validated it through k-Fold cross validation with k equal to 8 and with the use of an independent validation set (IVS). The confusion matrix can be found on the figure 2.

	0	1
0	3381	177
1	242	435

	0	1
0	4361	50
1	66	817

Table 2: Confusion Matrices for the Logistic Regression and the Decision Tree Classifiers

As we proceed previously, to find the best classification model we went ahead and compared all the statistics that we obtained for the different models in the table 3.

	Decision Tree	Logistic Regression
Accuracy	0.9781	0.9011
Precision	0.9423	0.7108
Recall	0.9253	0.6425
F1 Score	0.9337	0.6749
MCC	0.9206	0.6179

Table 3: Statistics used to evaluate the classification models;

Comparing the obtained statistics we can conclude with certainty that the best classification model is the decision tree due to the best results in all the parameters when comparing to the linear regression. One important detail that we need to address is that the accuracy on the Decision Tree is really high what might indicate overfitting of our model. The stopping criteria needs to be reevaluated and corrected in the future as to obtain model that aren't overfited.