Danmarks
Tekniske
Universitet

Project 2

Authors

Filippo Bosi - s220015 (section 2 & 5)
Davide Venuto - s220331 (section 3)
Aleksander Nagaj - s220350 (section 4)

April 21, 2022

Contents

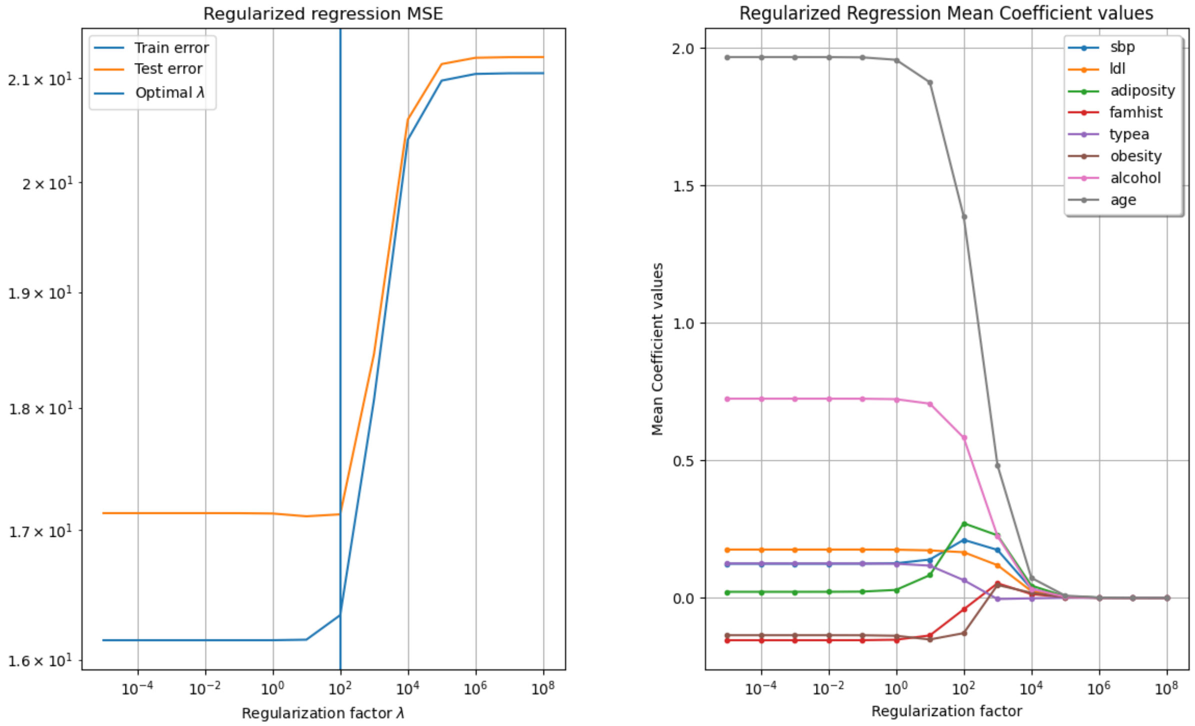
1 Dataset

All work in this report is based on the *South African Heart Disease* data set. It was previously described and profoundly analyzed in the **Project Report 1**.

2 Regression part A

The aim of this section was to predict the attribute *tobacco*, based on all the available attributes.

The reason for choosing to predict this particular attribute lies in the way the *tobacco* variable is calculated. It represents the cumulative amount of tobacco consumed by the subjects belonging to the data set during their entire lifetime. Hence, we expected that the *age* variable would have a large influence on the output of the linear regression model for predicting the *tobacco* attribute.



(a) Mean squared error over λ.

(b) Mean Coefficient values over λ.

Figure 1: Regularization factor λ.

At this point we have applied two feature transformations. First, we removed *tobacco* and *chd* attributes (and their relative values) from the data matrix to allow for the prediction of the *tobacco* variable. Later, we carried out the standardization of the remaining data matrix columns to enable the regularization of the data.

We then introduced a regularization parameter λ within the range of values $[10^{-5}, 10^{8}]$ with a logarithmic step. In this range, the generalization error, estimated using $K = 10$-fold cross-validation, first decreases and then increases. In our case this drop, even if present, is not very appreciable; consequently there is an uncertainty linked to the value assumed by λ. Here, we used the basic version of cross-validation to find the optimal value of lambda, which corresponds to $\lambda = 100$ (Figure 1a). However, it is possible to further improve this algorithm by implementing a two-level cross-validation in order to select a more accurate optimal value of λ, as we did for Section 3.

After analyzing the results from the generalized error perspective, we noticed that the regularization does not seem to affect our model significantly. The regularized linear regression model performed only moderately better with respect to the one without the optimal λ. We have also tried to double-check this outcome by changing the variable to be predicted with the other attributes of the data set, and the result did not change. This means that our linear regression model does not get any significant benefit by introducing a regularization parameter.

In general, a new observation can be predicted inside the linear model by making use of all the attributes in the data set. Upon deeper analysis, it becomes clear that there are two main attributes that influence the variable *tobacco*: *age* and *alcohol* (Figure 1b).

However, it is not possible to exclude the contribution of the other variables even if their weight is much lower compared to those two. In fact, in future non-linear regression analysis (e.g. Artificial Neural Network) these attributes could still play an important role. For instance, they may be characterized by a non-linear scheme that the linear model cannot catch, but an ANN can recognize.

Finally, as expected from the beginning of this study, the results produced by the regularized linear regression model for predicting *tobacco* seem reasonable from the weight of the attributes perspective, as they agree with the intuitive reasoning of how this variable is calculated.

3 Regression part B

3.1 Model and Parameters selection

In this second section we are going to compare three models: Regularized Linear Regression from the previous section, an Artificial Neural Network (ANN) and a Baseline. Our aim, as previously, is to predict the variable *tobacco* and then compare the performance of the three models based on their E^{gen}.

In order to implement the three models a standardization of the attributes was provided at first (subtracting the mean and dividing by standard deviation). Then a two-layer cross-validation with ($K_1 = K_2 = 10$) folds was implemented. For choosing the range of parameters h, λ that best fit the models with the data, some first trial run were iterated. In trial runs we set $K_1 = K_2 = 5$ and set a limitation of iterations to keep low the computation time and allow us to perform more evaluations.
Findings bring us to choose the range of h from $[1, 4]$ and λ in the interval $[10^{-5}, 10^{8}]$ with a log step, as in the previous section.

3.2 Comparison Analysis

Using the range of h and λ previously found, we proceeded with a comparison between the models (Table 1).

Using a two-layer cross-validation ($K_1 = K_2 = 10$), for each fold i, we extract the optimal value [1] of hidden units (h_i) and of regularization strength (λ_i). The chosen models with h^* and λ^* were evaluated. E_i^{test} was calculated for each fold i on D_i^{test}. For the error measure a Mean Square Error (MSE) was used (subsection 3.2).

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

In the last row the mean of E_i test was provided, it describes the approximated E^{gen} for each model and give us information about the error that we could expect and hope to find[2] if the model would be fit on a new data set. From the last line, it could be seen that the 'best performing model', which present the min E^{gen} of 16.93 and therefore fit the data in the best way, was the Linear Regression model E^{gen}. The second best-performing model, with a E^{gen} of 17.14 was the Artificial Neural network (ANN). At the end, as expected, the baseline had the worst performance. This is expected result since the baseline is a simplistic model. It takes Y^{train} as an input, calculates the mean, then uses that mean as an output for each Y_i test. Baseline is often used as a lower bound performance for testing more complex models.

[1] The optimal value taken in analysis was the value of h^*, λ^* that shows the min Gen Error in each inner loop.The model with h^* *lambda** is selected and then will be tested on a new test data set in the outer loop that gives us the error E_i^{test} shown in the table.

[2] we introduce the world "hope", therefore all this evaluation are done on one single dataset, which could show particular and singular correlation among the data (i.e from how data were collected) that are not general characteristic of the type of data taken in analysis

Outer fold	ANN		Linear Regression		Baseline
i	h_i^*	E_i^{test}	λ	E_i^{test}	E_i^{test}
K0	1	39.65	10.0	39.49	49.58
K1	1	14.39	10.0	15.78	15.81
K2	1	13.95	100.0	13.39	17.15
K3	1	26.08	10.0	24.98	28.84
K4	1	12.47	100.0	12.71	16.97
K5	1	11.3	10.0	11.13	15.85
K6	2	13.57	100.0	11.91	17.69
K7	1	12.24	10.0	11.8	13.16
K8	1	14.71	10.0	14.36	19.36
K9	1	13.06	10.0	13.71	16.03
Mean	**1**	**17.14**	**10**	**16.92**	**21.05**

Table 1: Two-level cross-validation table of the three models.

Analysing the Table 1, where all E^{gen}, h^* and λ^* of each outer fold were shown, it is worth noticing how, in some i folds, the model's E^{gen} appeared with a high volatility (i.e. K_0 and K_3) and in general the Generalization Error changed from fold to fold. In opposition the parameter h^* seemed to be more stable, it could be seen a change only in fold 6, from 1 to 2 hidden units. Regarding the second parameter λ^* two changing were seen from 10 to 100 in fold 2 and 6. However, changes in parameters didn't seem linked to any increase or decrease of E^{gen}.

3.3 Statistical evaluation with pairwise tests

For evaluating the models each against one another we decided to implement setup I methodology and perform a paired t-test. We have taken this decision because the data taken in analysis is quite old, meanwhile some medical technologies underlie data collection changed. For this reason a more general statistical evaluation and a following application of the results on newer data sets would not be useful. Therefore our aim was to evaluate how each model performs against the others in the data set. For performing the t-test, confidence level $\alpha = 0.05$, we are using the same k-fold[3] cross-validation used for the findings in Table 2, to get a coherent statistical evaluation.

[3] Since we are using k splits of the data, we expect that each f i is bit less performing than f_D trained on all the data set; therefore we expect that our estimation of E^{gen} will be a little overshoot.

Table 2: P-value comparison and rejection table for $\alpha = 0.05$

p-value comparison of E_{gen}		
$E_{\text{rlr}} - E_{\text{Base}}$ $p = 1.828e^{-7}$	$E_{\text{ANN}} - E_{\text{Base}}$ $p = 3.741e^{-7}$	$E_{\text{ANN}} - E_{\text{rlr}}$ p=0.278
$\times_{rejectH_0}$	$\times_{rejectH_0}$	$\checkmark_{acceptH_0}$

Analysing the p-value Table 2 it could be immediately seen how the null hypothesis of both the pairwise tests that concern baseline were rejected.It means that there is a strong statistical evidence that in both the cases the compared E^{gen} appears not to be equivalent. In other words the ANN and Linear Regression models outperformed the baseline. This statement appear to be particular evident by the analysis of the Confidence Intervals. How it could be see Figure 2 in both the cases the CI did not include the 0 and was strongly negative, the last statement means the baseline present an higher error. Turning to the last performed test,the p-value shown was higher than α. Which means we could not reject the null hypothesis:ANN and Linear Regression had equivalent performances in terms of E^{gen}. The same result were shown by the Confidence Interval Figure 2 as expected the interval contain the 0, and split it in 2 sections. However, it has to be reminded that the shown CI is only one realization of the defined statistic of CI. It means that in 95 of 100 cases the experienced θ will fall in this interval, but it could not give stronger statistical evidences for E^{gen} of one ANN being better than E^{gen} of Linear Regression, though the Interval seem to be slightly more in the positive region. In conclusion, in line with expectations, the baseline model was overall over-performed by ANN and Linear Regression. In contrast, ANN and Linear Regression seemed to perform in a similar way, there wasn't any statistical evidence to confute this Hypothesis.

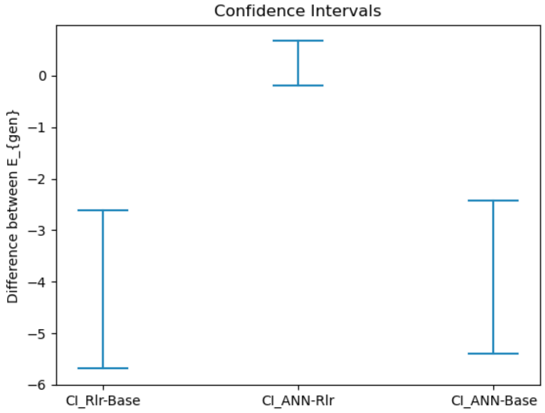


Figure 2: confidence intervals of E^{gen}=E^{gen}_A-E^{gen}_B

4 Classification

4.1 Models and Parameters selection

The *SA Hearth Disease* dataset consists of an obvious candidate for the dependent variable – the *chd* (coronary hearth disease) binary variable.[4]. It has been decided to use all the remaining attributes as an independent X **input vector**, since there wasn't performed any attribute significance analysis beforehand.

The setup used for classification was analogous to the one used in section 3. For each fold ($K_1 = K_2 = 10$), the **training set** was standardized as well as for the inner hyperparameter validation fold.

It has been decided to analyze two models: **Logistic Linear Regression** and **Decision Tree**, and compare their performance with a **Baseline** which for classification was given as:

$$\text{Baseline} = arg\max\{n_0, n_1\}$$

where

$$n_c = \sum_{i=i}^{N^{\text{train}}} \delta_{y_i^{\text{train}}, c}$$

are the number of observations assigned to each class c. In our case it is either 0 or 1. Logistic regression was standardized based on the parameter $\lambda \in \mathbb{N}$ which was chosen from the range $[10^{-5}, 10^{5}]^{\mathbb{N}}$ with a logarithmic step. The model at first fits the weighted linear function to a data, just as the linear regression from the previous section, then squashes it into a $[0,1]^{\mathbb{R}}$ range using the *sigmoid* function:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Then output of such function can be used as a parameter for a *Bernoulli* probability density:

$$Ber\left(y|\theta = \sigma(z)\right) = \theta^{y}(1-\theta)^{1-y}$$

which **expected value** is simply $\sigma(z)$. Weights and subsequently regularization strength λ has an influence on the model behaviour. However, unlike the linear regression, this time different variable is predicted (binary *chd*) and therefore attributes had been assigned different weights.

[4] 1 - subject has a disease, 0 - subject is healthy

Decision tree was parametrized by the maximum depth $d \in \mathbb{N}$ from range $[2, 21]$.

4.2 Comparison Analysis

Having defined range of potential parameters for both models and a Baseline, we performed $K = 10$ fold cross-validation and reported optimal λ^* and d^* parameter values as well as **test generalization errors** E_i^{test} (Table 3). For classification the error $E \in [0,1]^{\mathbb{R}}$ was defined as following:

$$E = \frac{\{\text{number of missclassified observations}\}}{N^{\text{test}}}$$

Outer fold	Logistic LR		Decision Tree		Baseline
i	λ^*	E_i^{test}	d_i^*	E_i^{test}	E_i^{test}
K0	10	0.28	3	0.34	0.32
K1	10	0.32	3	0.21	0.34
K2	10	0.17	5	0.24	0.28
K3	10	0.33	2	0.37	0.41
K4	100	0.24	3	0.24	0.30
K5	10	0.22	3	0.22	0.28
K6	10	0.28	3	0.33	0.35
K7	10	0.26	3	0.35	0.37
K8	10	0.20	3	0.17	0.28
K9	0	0.39	2	0.43	0.52
Mean	**10**	**0.27**	**3**	**0.29**	**0.35**

Table 3: Two-level cross-validation table of the three classification models.

As expected, both Logistic Regression and Decision Tree models outperformed the Baseline. For Logistic Regression the most common $\lambda^* = 10$ and a $\bar{E}^{\text{test}} = 0.27$. There wasn't a single fold where this model error rate was greater that Baseline. Slightly worse yet still feasible results were created by the Decision Tree which once scored even worse than a reference. On average, the optimal *maximum tree depth* $d^* = 3$ and $\bar{E}^{\text{test}} = 0.29$ which placed this model in between the remaining ones. It is worth noticing that the last $K = 9$ fold presented a significant challenge for all of the models resulting in significantly worse outcomes than any other fold and outlying parameters. It shows the significance and superiority of cross-validation over hold-out method. Training model on different subsets of the same data set allows to mitigate a bias caused by the arbitrary choice of a training and a test set.

4.3 Statistical evaluation with pairwise tests

In this section we implemented a McNemar test to compare pairwise the three models, in therms of their θ^5. Aim of the following tests is to evaluate in pairs if one model outperform the other. In different words if the null hypothesis H_0 could be rejected. The test was implemented inside the same k cross-validation split of the previous section, for maintain a statistical coherence. After selecting the type of cross validation, n predictions $\hat{y}_i$ were obtained for each model and the n confusion matrix Figure 4b was calculated. The next step was to define $\hat{\theta}$ (subsection 4.3) as the difference between θ_A and θ_B, utilized to define null Hypothesis H_0 in the following McNemar test.
The results of the tests are shown in Table 4 regarding the p-value and in Figure 3 concerning the Confidence Intervals (CI). Analysing the results of p-value it could be immediately seen how the Baseline approach was outperformed by both Logistic Regression and Classification Tree. It could be also seen from the Confidence Intervals both of them did not include and were quite far from the 0 value. Finally, analysing the last remaining test, Classification Tree against Logistic regression, a high p-value is shown. It means the null hypothesis could not be rejected. Looking at the Confidence Interval it could also be noticed how it is split by the zero in two parts. However, it only gives us a stronger confirmation of what was said before. In conclusion, it could be said that how expected the simplest Baseline was outperformed by the more complex models. Furthermore, there is no statistical evidence that neither Classification Tree nor Logistic Regression performed better than another.

$$\hat{\theta} = \frac{n_{12} - n_{21}}{n}$$

Where:

n_{12} classifier A is correct, B is wrong.

n_{12} classifier B is correct, A is wrong.

n total number of predictions.

Table 4: P-value comparison and rejection table for α=0.05

p-value comparison of $\theta = \theta_A - \theta_B$		
$\theta_{\text{LR}} - \theta_{\text{Base}}$ p=0.006	$\theta_{\text{Tree}} - \theta_{\text{Base}}$ p=0.002	$\theta_{\text{Tree}} - \theta_{\text{LR}}$ p=0.723
$\times_{rejectH_0}$	$\times_{rejectH_0}$	$\checkmark_{acceptH_0}$

[5] θ is the probability the classifier is correct $\theta = n_{\text{correct}}/n$

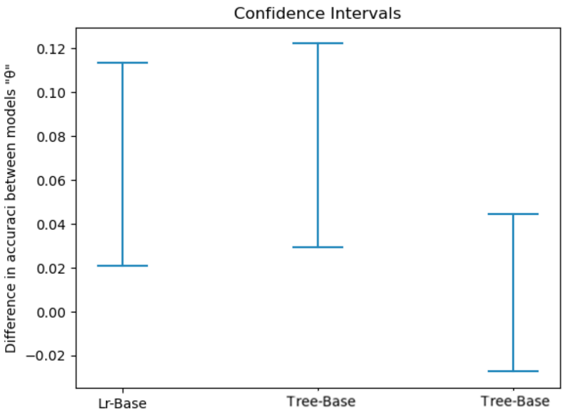
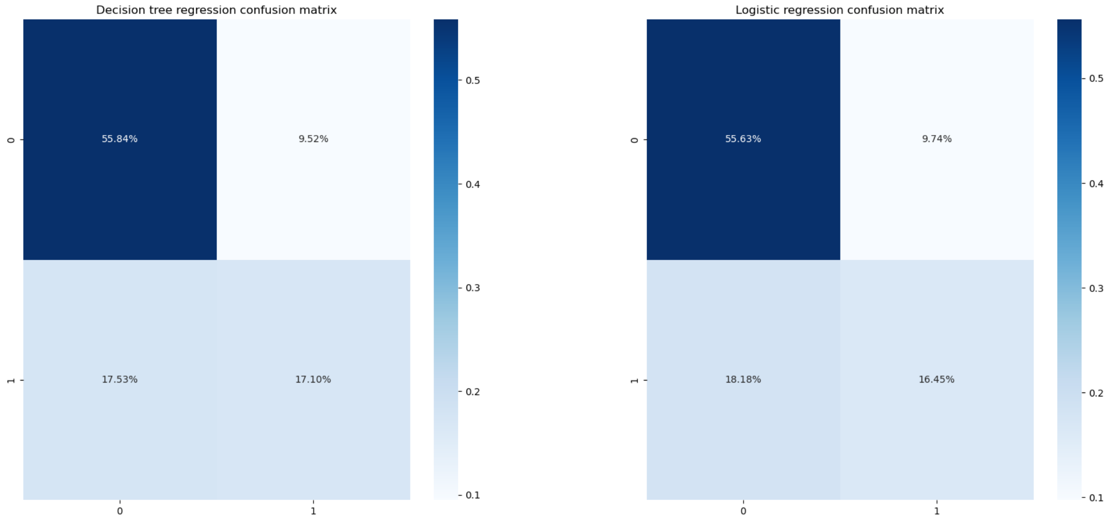


Figure 3: confidence intervals of θ from the compared classifiers and the empirical observed θ.



(a) Confusion Matrix Class Tree.

(b) Confusion Matrix Log Regression.

Figure 4: Confusion Matrix.

5 Final Discussion

As far as concerns the regression task, consisting in the prediction of the attribute *tobacco*, we have learned that, in our data set, the regularization does not significantly affect the outcome. However, we have confirmed our initial beliefs that the variable *tobacco* is largely influenced by the *age* attribute.

Moreover, thanks to the t-test (Section 3.3), we have verified that both the Regularized Linear Regression and the ANN models outperform the Baseline.

Considering now the classification task based on the McNemar test (Section 4.3), we have validated that the Logistic Regression model, as well as the Decision Tree model, perform better with respect to the Baseline. Nonetheless, the generalized error of these two more complex models is not outstanding, leaving ample room for further improvements.

5.1 Comparison of results

The previous analysis performed by (Hastie et al, 2009) [1] on the *South African Heart Disease* data set was focused on the classification of the Coronary Heart Disease (*chd*) attribute. We attempted to achieve the very same result for our project, since we believe it represents the most meaningful variable of the data set.

In their paper, the authors fitted a linear logistic regression model by maximum likelihood to the data set. Later, as a preliminary analysis, they excluded some of the attributes from the model. Subsequently, the authors implemented a nonlinear model by the means of natural splines. The interesting outcome of this analysis was that the previously excluded terms could be included again in the final nonlinear model.

Our study focused more on comparing different types of models rather than finding out which attributes could be dropped out from a model.

Hence, it is not possible to compare the results, in terms of generalization errors, of our Logistic Regression with respect to the one performed by the authors due to the lack of this aspect in their paper. Nonetheless, we can notice that in both projects non-linear models have been implemented. In fact, due to the nature of our data set, we can not exclude the presence of non-linear patterns that a linear model is not able to pick up.

The overall results achieved by our study are not completely satisfactory. By means of the Logistic Regression model for the classification of *chd* we obtained a generalization error of about 27% which is however better than the Baseline error. A way to improve this outcome, could be represented by trying to identify and remove outliers. The process of excluding outliers for both classification and regression tasks could be beneficial in order to reduce the final error rate.

In the end, even if the achieved result were not outstanding, our models performed better than the baseline, and we have gained a lot of experience for the implementation of classification and regression models.

References

[1] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning: Data mining, inference, and prediction. second edition.," pp. 122–124, 146–148, February 2009.

A Exam Questions

A.1 Question 1

C is correct. From the point at $\theta \approx 0.75$ where FPR reaches 0.5, TPR increases to 0.75 and stays at that level until FPR reaches 1, because the observation corresponding to lowest $\hat{y}$ is a false negative.

A.2 Question 3

A is correct. The ANN is composed by:

- Input units, $i = 7$
- Hidden units, $h = 10$
- Output units, $o = 4$

The number of parameters that has to be trained to fit the neural network is given by the sum of the connections between layers and the biases in each layer.

$$\begin{aligned} numParameters &= (i \times h + h \times o) + (h + o) = \\ &= (7 \times 10 + 10 \times 4) + (10 + 4) = 124 \end{aligned} \tag{1}$$

A.3 Question 4

The correct option is D. From the Classification boundary plot it is possible to notice a clear decision line which classifies all the observation with $b_1 \geq -0.16$ in Congestion level 4. From the decision tree, node C is the only one responsible for the classification of Congestion level 4. Option D is the only one that contains the correct expression for C (C: $b_1 \geq -0.16$). Therefore, D is the correct answer.

A.4 Question 5

C is correct.

Single model total training time is given as:

$$T = K_1(t_{\text{test}} + t_{\text{train}} + K_2(|N|(t_{\text{test}} + t_{\text{train}})))$$

Where $|N|$ is number of distinct parameters to be tested in inner validation loop.

Combining times for both ANN and Logistic Regression,

$$\begin{aligned}T &= K_1(t_{\text{test}_{\text{ANN}}} + t_{\text{train}_{\text{ANN}}} + K_2(|N|(t_{\text{test}_{\text{ANN}}} + t_{\text{train}_{\text{ANN}}})))\\ &+ K_1(t_{\text{test}_{\text{LR}}} + t_{\text{train}_{\text{LR}}} + K_2(|N|(t_{\text{test}_{\text{LR}}} + t_{\text{train}_{\text{LR}}})))\\ &= 5(5 + 1 + 20 + 8 + 4(5(20 + 5) + 5(8 + 1)))\\ &= 3570\end{aligned}$$