A black background with red and white triangles

Description automatically generated

02450 - Introduction to Machine Learning and Data Mining

Supervised learning: Classification and regression

Group 157

Even Johan Pereira Haslerud – s231473

Jørgen Finsveen – s231469

November 09, 2023

Table of Contents

[Author Agreement 3](#_Toc150998558)

[Introduction 4](#_Toc150998559)

[Regression Part A 4](#_Toc150998560)

[Regression Part A – 1 6](#_Toc150998561)

[Regression Part A – 2 6](#_Toc150998562)

[Regression Part A – 3 7](#_Toc150998563)

[Regression Part B 7](#_Toc150998564)

[Regression Part B – 1 7](#_Toc150998565)

[Regression Part B – 2 7](#_Toc150998566)

[Regression Part B – 3 7](#_Toc150998567)

[Classification 7](#_Toc150998568)

[Classification – 1 7](#_Toc150998569)

[Classification – 2 7](#_Toc150998570)

[Classification – 3 7](#_Toc150998571)

[Classification – 4 7](#_Toc150998572)

[Classification – 5 7](#_Toc150998573)

[Analysis of results – Discussion and Conclusion 7](#_Toc150998574)

[References 7](#_Toc150998575)

[Appendix A: Mandatory Exam Questions 8](#_Toc150998576)

[Question 1. Spring 2019 Question 13: 8](#_Toc150998577)

[Question 2. Spring 2019 Question 15: 8](#_Toc150998578)

[Question 3. Spring 2019 Question 18: 9](#_Toc150998579)

[Question 4. Spring 2019 Question 20: 9](#_Toc150998580)

[Attributes in Table 1: 10](#_Toc150998581)

[Question 5. Spring 2019 Question 22: 10](#_Toc150998582)

[Question 6. Spring 2019 Question 26: 10](#_Toc150998583)

# Author Agreement

This document is the report for Project 2 in the course "Introduction to Machine Learning and Data Mining" during the Autumn Term of 2023 at Denmark's Technical University. The authors, Even Johan Pereira Haslerud and Jørgen Finsveen have contributed equally to the project's work.

|  |  |  |
| --- | --- | --- |
| **What type of work?** | **Even Johan** | **Jørgen** |
| *Regression, part a* |  |  |
| *Regression, part b* |  |  |
| *Classification* |  |  |
| *Discussion* |  |  |
| *Exam problem 1* |  | *X* |
| *Exam problem 2* |  | *X* |
| *Exam problem 3* |  | *X* |
| *Exam problem 4* | *X* |  |
| *Exam problem 5* | *X* |  |
| *Exam problem 6* | *X* |  |
| *Tables* | *X* | *X* |
| *Figures* | *X* | *X* |
| *Report* | *X* | *X* |
| *Coding* | *X* | *X* |

# Introduction

In this report, we will build upon the first project that used the Heart Disease dataset. Specifically, we will focus on applying two methods from supervised learning: Classification and Regression. Our goal is to evaluate the performance and characteristics of different types of supervised learning models in relation to this dataset. The report will be divided into two sections, one on regression and another on classification.

# Regression Part A

The Heart Disease dataset was originally chosen for classification tasks, having 14 attributes – seven numerical and seven categorical. Among all these, the “num” attributes serves as the target variable indicating the diagnosis and extent of heart disease, which is the primary endpoint of interest for the classification part. The attribute is uniquely suited for prediction in a classification framework, as it encapsulates the core question of disease presence or severity. Conversely, the dataset lacks attributes gathered for regression analysis which suggests that models developed for such purposes might be prone to increased inaccuracies due to the dataset’s inherent structure and data collection objectives.

However, selecting an attribute from the dataset as the target for regression analysis was difficult because most attributes were either uncorrelated or only weakly correlated with one another, leading to a low predictive capability within the dataset for any given variable. To address this issue, we decided to use one of the 7 numerical attributes as our target for regression analysis while the other 14 attributes would serve as predictor variables. To determine the most suitable attribute for regression analysis among the six candidates, we needed a quantitative analysis method. We implemented a simple Ordinary Least Square (OLS) linear regression model, comparing each candidate target variable against all other variables in the dataset to assess its predictive strength.

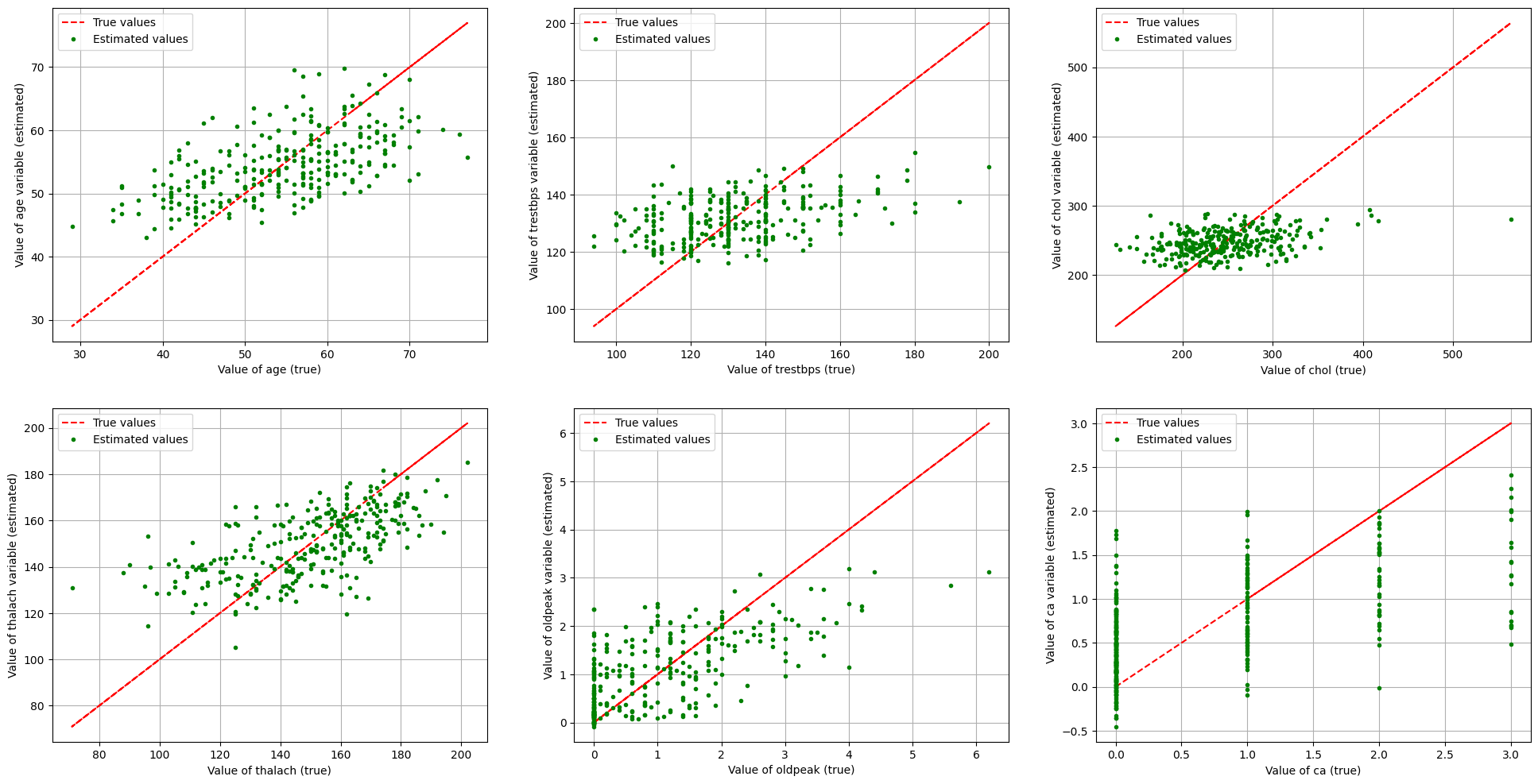


Figure 1: Six scatter plots for comparison between the given parameters.

As observed from the figure above, the dataset does not have high predictive power for the variables ‘trestbps’ and ‘oldpeak’, while the prediction of ‘ca’ and ‘chol’ is both weak and influenced by outliers. Both ‘thalac’ and ‘age’ have stronger predictive power and is therefore the most attractive attributes for the regression. One should note that some estimated valued for ‘thalac’ may be outliers. For the regression, the value which we would like to predict is ‘thalac’ as we consider this to be the most optimal attribute for this purpose based on the plots above.

Furthermore, it can also be beneficial to gather some insight in the correlation between the different attributes in the dataset when performing the regression. Should any attribute have little to no correlation with the ‘thalac’ attribute, they would not be of any use when performing predictions. The correlation between the attributes can be visualized by plotting a correlation heatmap.

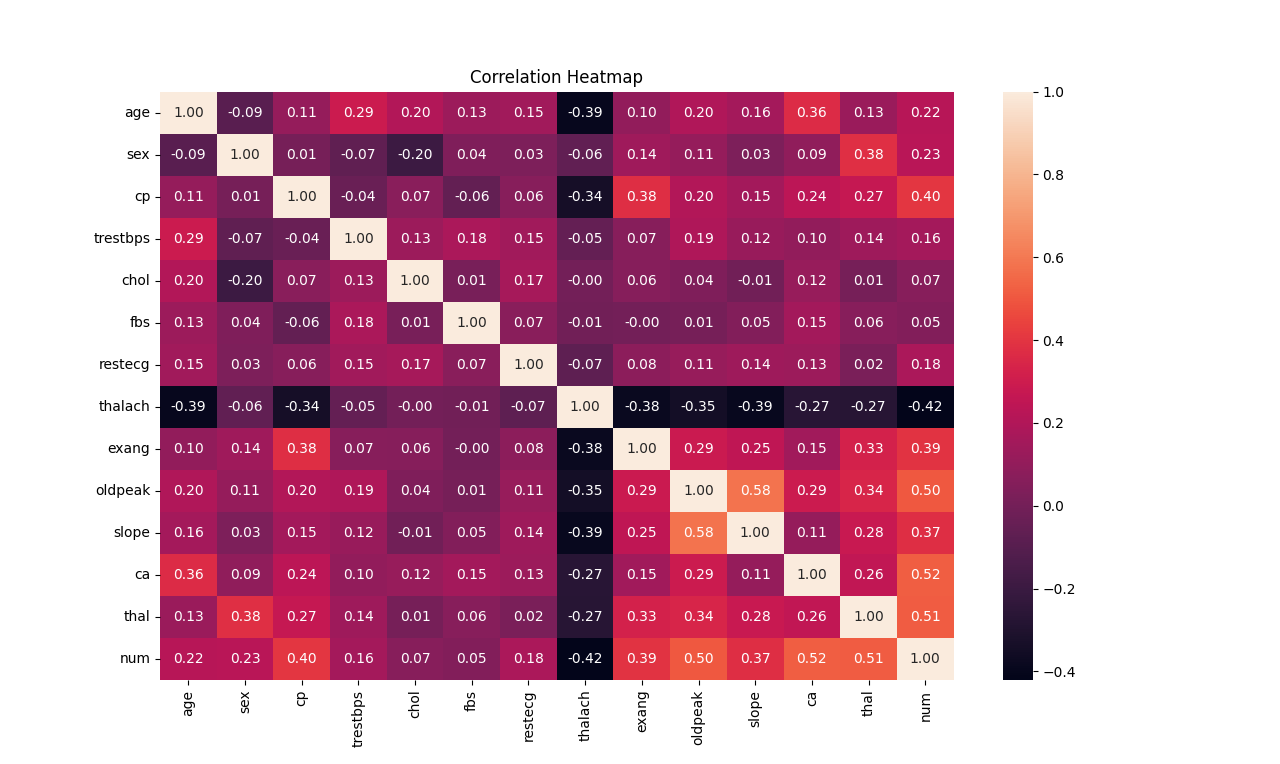


Figure 2: Heatmap

As seen in the heatmap, there are some attributes which has a very weak correlation with the ‘thalac’ attribute. One must therefore consider whether it is necessary to include these when performing the prediction. The attribute ‘chol’ has a correlation coefficient of 0.0 with ‘thalac’, rendering it to be unnecessary for the regression. Some other attributes do have a very small correlation with ‘thalac’, but since they are non-zero, we have chosen to include them in the regression regardless. The attributes which have the highest correlation with ‘thalac’ and therefore best describes this attribute are ‘age’, ‘slope’, and ‘exang’.

## Regression Part A – 1

In this initial phase of our regression analysis, we focused on predicting the 'thalac' variable, which represents the maximum heart rate achieved, using the Heart Disease dataset. We began with a set of 14 categorical variables, which we prepared for the linear regression model by employing one-of-K coding. This resulted in an expanded feature set, increasing from 16 to 21 features after encoding. This step is essential to prevent the model from making incorrect ordinal interpretations.

The goal of linear regression is to estimate the dependent 'thalach' through a linear relationship with independent variables, weighted according to their significance. The weights are determined by minimizing the mean squared error on the training data, which enables the model to predict 'thalac' effectively on new inputs.

To improve the model's accuracy, we applied a forward selection process that involved incrementally adding features that significantly reduced the error. This was evaluated through a robust 5-folder outer and 10-fold inner cross-validation scheme, which seeks to balance the model's simplicity and predictive accuracy.

We also tested additional non-linear transformations, including interaction and power terms for highly correlated variables, but they ultimately yielded negligible improvements. As a result, we decided not to complicate our model with these transformations.

To normalize the scale of our variables, which is crucial for the upcoming regularization step, we standardized our features. This ensures that each feature contributes proportionately to the model, preventing any undue influence due to scale disparities.

Our findings from the forward selection algorithm revealed that attributes like 'age', 'slope', and 'exang' most effectively describe 'thalac'. Interestingly, the inclusion of all features explained approximately 33,5% of the variance, with feature selection not significantly optimizing the model. This suggests that the full model, despite its complexity, is not overfitting and that even less-correlated features have a role in minimizing the error.

## Regression Part A – 2

If we want to work with a linear regression model that gives the lowest possible generalization error, we need to determine the optimal value of λ that would yield the lowest generalization error. To reliably estimate the generalization error for different values of λ, we performed a two-level cross-validation test. The outer layer had K1=10 folds while the inner layer had K2=10 folds. The two-level approach is useful when selecting the optimal value of λ for our model and data to minimize the generalization error. This is because we do not want to take an optimistic minimal value that occurred due to randomness.

We estimate the generalization error by computing the mean test error given by each possible λ value within the 10 folds. We then select the λ for which this error is minimum.

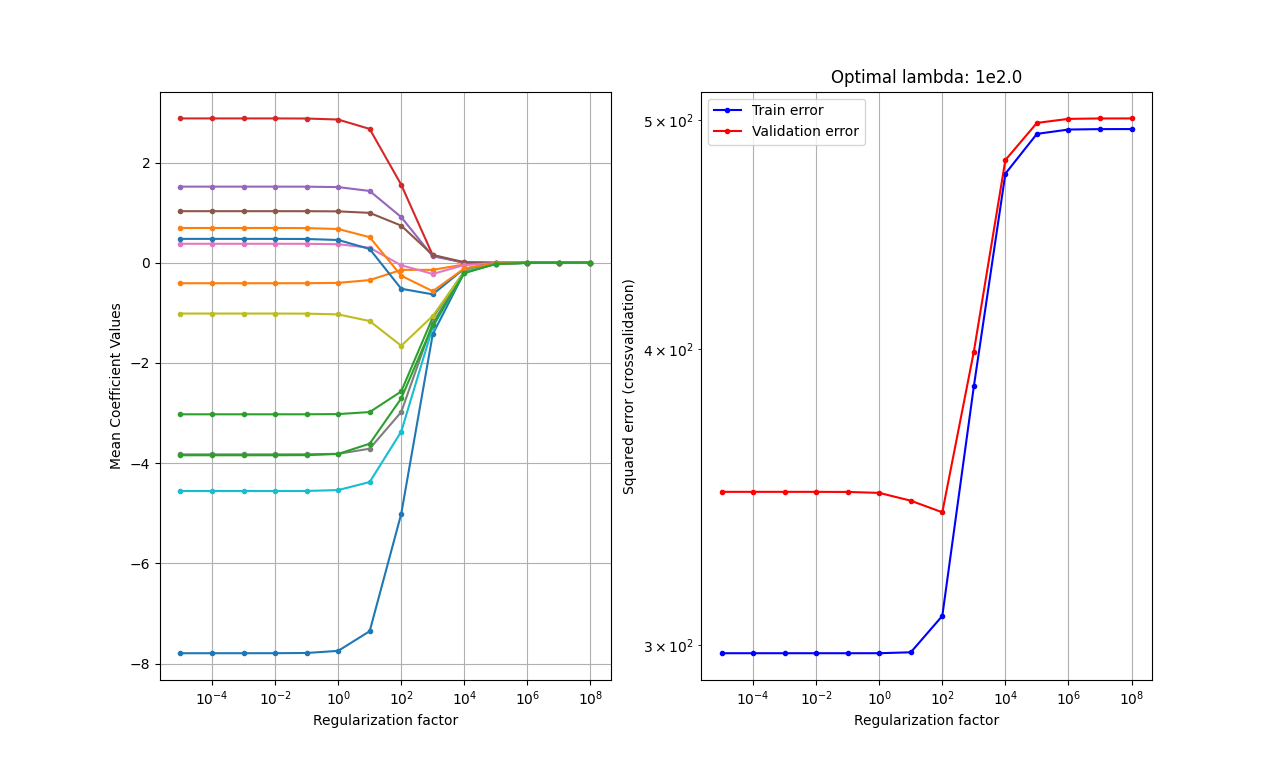
Figure 3 below to the right shows the trend of the generalization error drop and subsequent increase for increasing λ. On the other hand, figure 3 below to the left side illustrates the value of the linear model's coefficients depending on λ. It shows that as λ increases, all the weights decrease and converge towards 0.

Figure 3: Regularization

Based on Figure 3, we will select λ as . This is to minimize the mean squared errors in the linear regression model. Unfortunately, due to our poor predictor in the dataset for a regression model, when computing the overall generalization error using the two-level cross-validation, we find that using the full unregularized linear model results in a testing error of 344.63 and of 33.5%. When using the regularized linear regression model, we achieve a slight error reduction to 342.5 and an increase in to 34.06%. This makes it clear that our models cannot easily be improved through internal attribute tweaking. Instead, we need to gather new data to find and add powerful predictors.

## Regression Part A – 3

By selecting λ to be 100 and fitting the linear model on the entire dataset, we obtain the equation for the regularized linear model:

Based on our standardized data and considering the maximum heartbeat rate achieved during the hospital stay as the criterion variable, we can interpret the results effectively. The weights in the equations indicate that an increase of 1 age unit corresponds to a decrease of 6.13 units in the maximum heart rate, and that patients without heart disease also exhibit similar findings. All these findings are relevant as the human heart tends to pump blood slower with increasing age, and angiographic heart disease leads to a higher heart rate. The other parameters also show sensible results.

# Regression Part B

For this part we will compare three supervised modelling techniques, evaluate their general performance using two-level cross-validation and statistically determine whether the performance of different models is comparable or significantly different.

## Regression Part B – 1

To assess the generalization error of our regression models, we utilized a two-level cross-validation approach. For the outer and inner folds, we set the fold count to K1 = K2 = 10. The inner folds helped us fine-tune each model's complexity-controlling parameters to minimize generalization error, while the outer folds were used to assess the model’s performance across various segments of our dataset.

For the regularized linear regression, our goal was to find the optimal regularization strength, denoted as λ, that minimizes the Mean Squared Error (MSE) and balances the trade-off between variance and bias. We incorporated an intercept in the linear model and explored λ values spanning powers of ten within the range of to , which was the same range used in the previous exercise.

In parallel, we set up an Artificial Neural Network (ANN) that had a single hidden layer with a hyperbolic tangent (tanh) activation function, and an output layer suitable for regression tasks. We varied the number of hidden units from 1 to 5 to adjust the complexity of the ANN. To make sure our results were robust, we set the ANN's training iterations to a maximum of 10,000 epochs and conducted training across three replicated networks. This redundancy accounted for the possibility of the network becoming ensnared in local minima during the optimization process. We selected the ANN model with the lowest error from these trials as the optimal model.

As for the baseline model, we used a simple linear regression with no predictors. This meant that we used the mean of the dependent variable y from the training data as the prediction for all instances. This model served as a benchmark to illustrate the improvements gained from more advanced models and to ensure that any complexity added provided tangible benefits over a basic average-based prediction.

## Regression Part B – 2

As part of our analysis, we conducted a rigorous comparison of three predictive models using a two-layer cross-validation approach. We set both and to 10 folds. In this method, we employed a nested cross-validation procedure where the inner loop determined the optimal model parameters for predicting the target variable, while the outer loop evaluated the accuracy of the prediction.

We implemented this approach on the pre-processed dataset from Part A and the results are presented in Table 2. The table outlines the optimal number of hidden units (h) for the ANN, optimal λ for regularized linear regression, and the generalization errors across the ten outer folds for each model. We calculated the generalization error based on the squared loss per observation, as expressed by the following equation:

However, while executing the two-level cross-validation, we encountered discrepancies in the regularized linear regression model’s performance. Contrary to our expectations, the estimated generalization errors were significantly higher, with λ values ranging broadly from 0.01 to 100,000. This anomalous outcome suggests potential issues in the cross-validation implementation or data handling specific to this model, as such errors are uncharacteristically large.

Therefore, for a fair assessment, we adhered to the parameter tuning methodology from Part A for the linear regression model, despite leading to variations in the training/testing datasets used across the models. This discrepancy introduces a degree of uncertainty in our statistical comparison but is preferable to basing conclusions on evidently flawed error measures.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Fold** | **Baseline Error** | **Optimal λ** | **Ridge Error** | **Optimal h** | **ANN Error** |
| 1.0 | 509.78943 | 72.208090 | 297.064144 | 64.0 | 810.857297 |
| 2.0 | 592.688313 | 44.306215 | 334.506666 | 64.0 | 1022.841634 |
| 3.0 | 427.687543 | 0.000027 | 383.408093 | 64.0 | 714.970055 |
| 4.0 | 575.320216 | 16.681005 | 382.676768 | 64.0 | 676.702732 |
| 5.0 | 536.251715 | 0.01556 | 338.781417 | 64.0 | 941.837980 |

From the analysis of Table 1, it is evident that both the ANN and linear regression models perform better than the baseline model. The performance of the ANN and linear regression models is quite similar. However, detailed statistical evaluation is required to determine whether there are any significant differences in the performance of the models.

Table 2 also provides information on the optimal parameters for each model. For the ANN model, the number of hidden units, denoted by h, varies between 3 and 9, with an average of around 6. The optimal regularization strength, denoted by λ, ranges from 5 to 30, with a median of 12. This median value is in close agreement with the value of λ obtained in Part A, indicating consistency in the optimal regularization strength across different parts of our analysis.

## Regression Part B – 3

As part of our regression analysis, we evaluated three models - Artificial Neural Network (ANN), linear regression, and the baseline model. Our aim was to identify if there were any significant differences in their performance. We conducted a comprehensive statistical analysis during the final stage to determine the same.

The paired comparisons were structured as follows: ANN vs. linear regression, ANN vs. baseline, and linear regression vs. baseline. We had the flexibility to choose paired t-tests, which were a suitable choice due to the continuous nature of our outcome variable, mean squared error.

**Statistical Test Outcomes**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  | 0.0035 | 259.947 | 685.357 |  |
|  | 0.0526 | -5.513 | 605.757 |  |
| 0 | 0.0073 | -267.780 | -77.281 |  |

Following insights on the model can be gained from the Table above:

* The linear regression model significantly outperformed the baseline model, as indicated by the rejection of the null hypothesis and the confidence interval not encompassing zero.
* No significant performance difference was detected between the ANN and linear regression models, as the null hypothesis was accepted given the p-value marginally exceeded our alpha level of 0.05.
* The ANN model also demonstrated superior performance compared to the baseline model, with the null hypothesis being rejected here as well.

# Classification

## Our goal in this classification task is to predict whether a patient has angiographic heart disease or not, based on the 20 attributes of the dataset (after one-out-of-K encoding is performed). The dataset was collected for this specific purpose. Therefore, this is a binary classification problem. To deal with potential errors that may arise from large differences of scale, we will standardize the values of the 20 attributes.

## Classification – 1

We selected three methods for comparison: logistic regression, k-nearest neighbour classification (KNN), and a baseline model. The parameters for these methods were chosen based on preliminary trial runs.

**KNN:** The complexity-controlling parameter for KNN is 'k', the number of neighbours. We will explore the range of:

**Logistic Regression:** Here, λ is used as the complexity-controlling parameter. The chosen range is:

**Baseline Model:** This model predicts the most frequent class in the training data, making it a benchmark for comparing the performance of other models.

## Classification – 2

## Classification – 3

## Classification – 4

## Classification – 5

# Analysis of results – Discussion and Conclusion

# References

[1] “Project description for report 2”, 16 November 2023. [Project 2 description for 02450 Introduction to Machine Learning and Data Mining]

[2] Herlau T., Schmidt, M.N., and Mørup, M., “Introduction to Machine Learning and Data Mining,” 2021. [Official book for 02450 Introduction to Machine Learning and Data Mining].

# Appendix A: Mandatory Exam Questions

## Question 1. Spring 2019 Question 13:

**Option D:**

The ROC is created by computing different False Positive Rates (FPR) and True Positive Rates (TPR) for different threshold values for . One can create a set of coordinates by calculating for threshold values, where a coordinate is on the form (FPR, TPR). TPR can be calculated by dividing the number of predictions for a class over the threshold by the total number of observations for the same class. For FPR, one divides the number of predictions for the opposite class over the threshold by the total number of observations for the opposite class.

Start by calculating coordinates for the following thresholds based on the candidate predictions:

Upon computing the ROC curve for each candidate prediction respectively, the only candidate which match the ROC curve in question is the one corresponding to alternative D.

One should also note that the AUC can be found by calculating the area below the ROC curve. For the curve in question, the AUC is equal to 0.5, which indicates that the predictions made by the classifier are almost completely random and does not provide any accuracy. This further strengthens the claim that alternative D is correct, since one can see by the prediction plot that the class predictions are scattered across the axis and not grouped together by class.

## Question 2. Spring 2019 Question 15:

**Option C:**

By the binary split of x7=2 there are only a single observation which is for y=2. This results in the creation of a leaf node separating y=2 from the other attributes.

The total number of observations are 135. This results that there are 1 observation when x7=2, and 134 observations when x7≠2. To calculate the impurity gain, one can use Gini´s method, which will result in the following:

The correct statement is therefore option C.

## Question 3. Spring 2019 Question 18:

**Option A:**

The total number of features in the dataset is XTOT = 7. The assignment defines a single hidden layer with *nh* = 10 units. The table describes the different attributes in the dataset, and from the description, one can see that there are attributes of the following types:

* Interval: x1
* Ratio: x2, …, x7
* Nominal: x1
* Ordinal: y

In total, the attributes in the dataset are of four different attribute types, resulting in C = 4 different classes.

In order to calculate the number of parameters one would need to train to fit the neural network:

1. Number of parameters:

The number of parameters required are 124, making option A the correct alternative.

## Question 4. Spring 2019 Question 20:

**Option D:**

We can easily identify the splitting rules after setting the decision tree combinations into the classification boundary. The nodes A and C are determined by , while the other nodes B and D are determined by . This information is shown in Figure 1 below:

**D:**

**B:**

|  |  |  |
| --- | --- | --- |
| 2 | 1 | 4 |
| 1 | 3 | 4  **C:** |

**A:**

*Figure 1: Combination of the classification boundaries and decision tree.*

### Attributes in Table 1:

* **Number 1:** Congestion level 1
* **Number 2:** Congestion level 2
* **Number 3:** Congestion level 3
* **Number 4:** Congestion level 4
* **Nodes – A, B, D, and C:** These are the nodes from the structure of the given decision tree.
* **:** Is the x-coordinate for PCA 1 in Figure 4 of the Classification boundary.
* **:** Is the y-coordinate for PCA 2 in Figure 4 of the Classification boundary.

## Question 5. Spring 2019 Question 22:

**Option C:**

Both artificial neural network (ANN) and logistic regression models require performing training and testing on the same outer and inner folds. As a result, λ, and have a total of 5 possible values for both training and testing. This requirement affects the total time taken for the process, which can be calculated as follows:

This timing calculation is essential and must be carried out once for each of the five outer folds to identify the optimal model (which is the one with the lowest generalization error, ). Furthermore, it is necessary to select the best model from among the four inner folds. This selection process must be repeated five times, leading to the equation:

This ensures that each fold contributes to the determination of the most efficacious model, considering the time required for both training and testing.

## Question 6. Spring 2019 Question 26:

**Option B:**

The first step of resolving this question is to put the value of b and into the given equation to compute the , for respectively. Then we need to put the value of , for into the given softmax function to calculate the , for . According to the calculation of the option B, the approximate probabilities of the 4 classes are 0.05, 0.06, 0.15, 0.73. The class has the largest probability, so the option should be B.