Danmarks Tekniske Universitet

DTU

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

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	*X (100%)*	
Regression, part b		*X (100%)*
Classification	*X (40%)*	*X (60%)*
Discussion	*X (60%)*	*X (40%)*
Exam problem 1		*X (100%)*
Exam problem 2		*X (100%)*
Exam problem 3		*X (100%)*
Exam problem 4	*X (100%)*	
Exam problem 5	*X (100%)*	
Exam problem 6	*X (100%)*	
Tables	*X (70%)*	*X (30%)*
Figures	*X (30%)*	*X (70%)*
Coding Regression	*X (70%)*	*X (30%)*
Coding Classification	*X (30%)*	*X (70%)*

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.

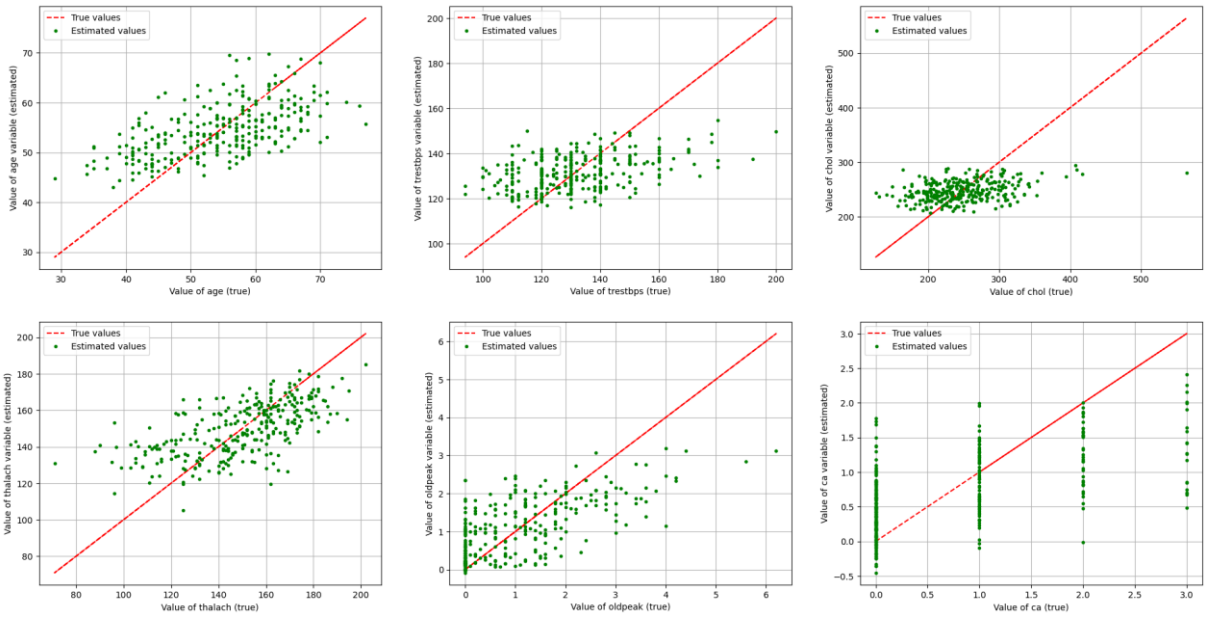

Figure 1: Six scatter plots for comparison between the given parameters.

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.

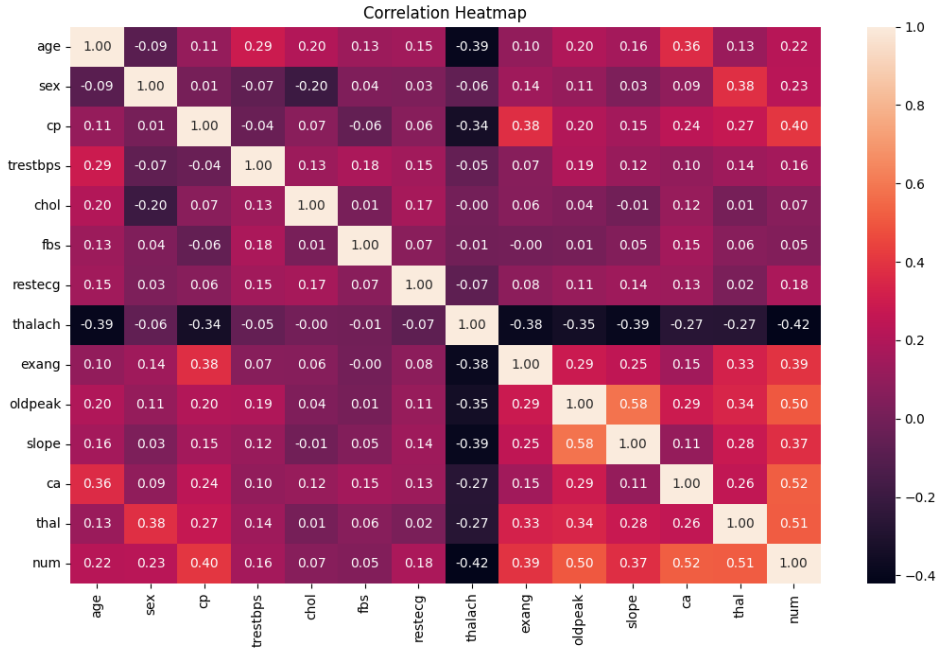

Figure 2: Heatmap

Based on Figure 2, it can be observed that some attributes have a weak correlation with the 'thalac' attribute. Therefore, it is important to consider whether it is necessary to include these attributes while performing the prediction. For instance, the 'chol' attribute has a correlation coefficient of 0.0 with 'thalac', making it unnecessary for the regression analysis. However, some other attributes have a small correlation with 'thalac', but since they are non-zero, they have been included in the regression analysis. Finally, the attributes that have the highest correlation with 'thalac' and best describe 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 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.

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

To obtain the lowest generalization error in a linear regression model, finding the optimal λ value is crucial. Two-level cross-validation with K1=10 folds in the outer layer and K2=10 folds in the inner layer was performed to estimate the generalization error for various λ values. This method helps to choose the best λ value that will minimize the generalization error without being too optimistic due to randomness.

To estimate the generalization error, we calculated the average test error for each value of λ within the 10 folds. We then selected the λ value that resulted in the minimum error. On the left, Figure 3 below displays the linear model's coefficients at different λ values. As λ increases, all the weights decrease and converge towards 0. On the right, Figure 3 below shows the trend of the generalization error dropping and subsequently increasing as λ value increases.

Based on Figure 3 below, we will select λ as 10^2. 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 R^2 of 33.5%. When using the regularized linear regression model, we achieve a slight error reduction to 342.5 and an increase in R^2 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.

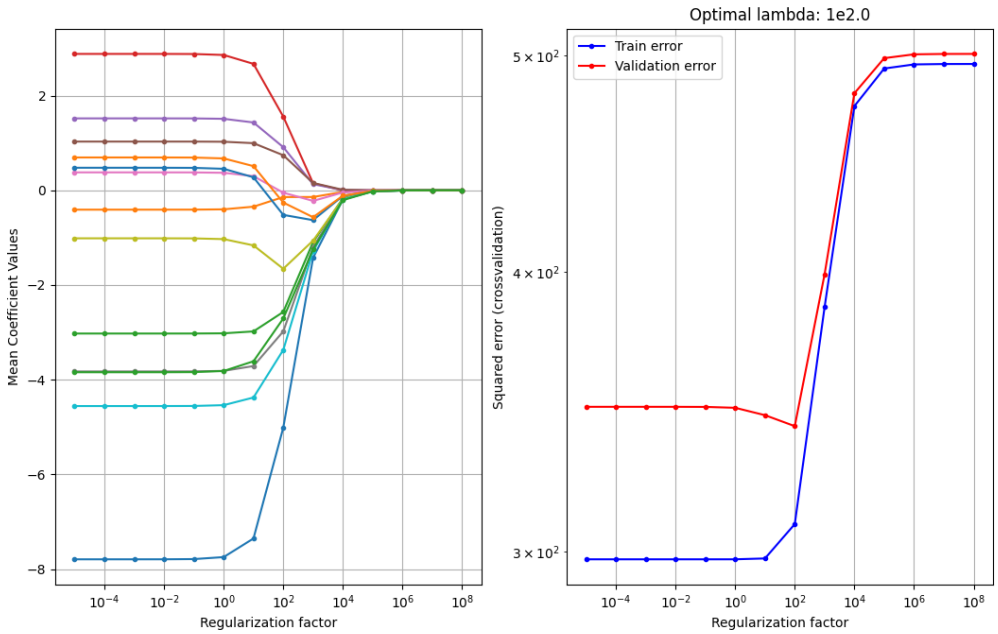

Figure 3: Regularization

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:

$$y = 149.59 - 6.13x_{age} + 0.108x_{sex} - 2.82x_{cp} + 1.458x_{trestbps} + 1.403x_{chol} + 0.512_{fbs} + 0.459x_{restecg} - 3.968x_{exang} - 1.102x_{oldpeak} - 4.044x_{slope} - 0.64x_{ca} - 0.561x_{thal} - 2.804x_{num}$$

Our analysis of standardized data shows that the maximum heartbeat rate achieved during the hospital stay is directly linked to age and the presence of heart disease. We found that with each increase of 1 age unit, the maximum heart rate decreases by 6.13 units. This correlation is significant because the heart tends to pump blood slower with increasing age, and angiographic heart disease leads to a higher heart rate. Furthermore, we observed similar findings in patients without heart disease. All our results are sensible and contribute to a better understanding of this topic.

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 10^{-5} to 10^{2}, 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 K_{inner} and K_{outer} 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:

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

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

Table 1: Two level cross validation table used to compare the three models in the regression problem.

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

H_0	$p-value$	$Lower\ CI$	$Upper\ CI$	$Conclusion$
$E^{test}_{baseline} - E^{test}_{linear\ regression} = 0$	0.0035	259.947	685.357	$H_0\ Rejected$
$E^{test}_{ANN} - E^{test}_{linear\ regression} = 0$	0.0526	-5.513	605.757	$H_0\ Accepted$
$E^{test}_{baseline} - E^{test}_{ANN} = 0$	0.0073	-267.780	-77.281	$H_0\ Rejected$

Table 2: Statistical test for the pairwise comparison of the three regression models for a significance level alpha = 5% together with the 95% confidence interval boundaries.

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

Classification – 1

Our goal in this classification task is to predict whether a patient has angiographic heart disease or not, based on the 14 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 14 attributes.

Classification – 2

We selected three methods for comparison: logistic regression, an artificial neural network (ANN), and a baseline model. The parameters for these methods were chosen based on preliminary trial runs.

ANN: Artificial neural networks for classification. Same complexity-controlling parameter as in the previous exercise

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

$$\lambda \in [10^{-2} : 10^{2}]$$

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 – 3

To evaluate the performance of different models, we used a two-layer cross-validation approach with both the number of inner and outer folds set to 10 $K_{inner},\ K_{outer} = 10$. The data was standardized as outlined in section 1 of this report. Table 4 displays the optimal regularization parameters for each method across the outer folds. These parameters correspond to the values that resulted in the lowest generalization error during cross-validation. We calculated the error rate, E, using the following formula:

$$E = \frac{Number\ of\ misclassified\ observations}{N_{test}}$$

Figure 4 shows accuracy as $1 - E_{gen}$, reported in percentage terms.

Outer Fold	**ANN**		**Logistic Regression**		**Baseline**
i	k_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
0	2	0.3667	61.359073	0.30	0.367
1	2	0.4333	61.359073	0.43	0.433
2	2	0.5667	61.359073	0.46	0.567
3	2	0.5667	61.359073	0.56	0.600
4	2	0.6667	61.359073	0.60	0.667
5	2	0.4333	61.359073	0.30	0.433
6	2	0.4000	61.359073	0.40	0.400
7	2	0.4482	61.359073	0.41	0.448
8	2	0.3793	61.359073	0.27	0.379
9	2	0.3103	61.359073	0.34	0.310

Table 3: Two level cross validation table used to compare the three models in the classification problem.

Classification – 4

In the comparative analysis of three distinct classification models—Logistic Regression, Artificial Neural Networks (ANN), and a Baseline model—McNemar's test was employed to assess the pairwise differences in performance.

Statistical Test Outcomes

$\boldsymbol{H_0}$	$\boldsymbol{p - value}$	$\boldsymbol{Statistics}$
$E_{logistic}^{test} - E_{ANN}^{test} = 0$	4.162	98.01
$E_{logistic}^{test} - E_{baseline}^{test} = 0$	2.775	21.966
$E_{ANN}^{test} - E_{baseline}^{test} = 0$	3.289	135.007

Table 4: McNemar's statistical test for the pairwise comparison of the three classifiers models.

- Logistic Regression vs. ANN
  - The statistical test yielded a value of 98.01, with a corresponding p-value of 4.1627504389864093e-23.

    - The extremely low p-value suggests that there is a statistically significant difference in the performance of the Logistic Regression model compared to the ANN.
- Logistic Regression vs. Baseline
    - The comparison produced a statistic of 21.966101694915253, with a p-value of 2.7750883948169247e-06.
    - This p-value, being far below the conventional alpha level of 0.05, indicates a statistically significant difference between the Logistic Regression model and the Baseline model.
- ANN vs. Baseline
    - The test statistic here was 135.007299270073, and the p-value was 3.2895733042869343e-31.
    - This result also demonstrates a statistically significant difference, this time between the ANN and the Baseline model.

Both the Logistic Regression and ANN models have displayed a statistically significant improvement over the Baseline model, which indicates that both models have learned patterns from the data that can be used to predict the outcome variable. The significant difference between the Logistic Regression and ANN models leads us to believe that one model may be better suited to the dataset than the other. However, without additional context, such as the error rates or the complexity of the models, we cannot definitively determine which model performs better overall. Nonetheless, given the significant p-values in all comparisons, it is safe to say that none of the models perform identically on the dataset.

Classification – 5

Lastly, on classification using logistic regression, we trained a model to predict multi-class categories. Logistic regression is a predictive analysis used to describe data and explain the relationship between one binary dependent variable and one or more nominal, ordinal, interval, or ratio-level independent variables. In simpler terms, logistic regression predicts the probability of an event by fitting data to a logistic curve.

Our logistic regression model predicts class membership probabilities based on the logistic function applied to a linear combination of the input features. The logistic function, also known as the sigmoid function, maps any real-valued number into the range between 0 and 1, making it suitable for probability interpretation. In a multi-class setting, the model likely employs a one-vs-rest (OvR) scheme or a multinomial approach to distinguish between more than two classes.

From the confusion matrix provided in Figure 5 down to the right, it's clear that the model performs well in predicting class '0' but shows some confusion between class '1' and other classes. For instance, in some cases, class '1' was mistaken for class '2', class '3', and class '4'. This suggests that while the model is quite adept at identifying class '0', it struggles somewhat with class '1'.

The process for model training began by splitting our dataset into training and testing sets. This ensured that we had a separate dataset to evaluate our model's performance. We then standardized our features to give them equal weight in our model, as differences in scale can skew how logistic regression models interpret the data. Standardization is particularly important when using penalty terms, as it is in our model, which uses 'l2' regularization controlled by the hyperparameter lambda (λ). The regularization helps prevent overfitting by penalizing larger coefficients in the model.

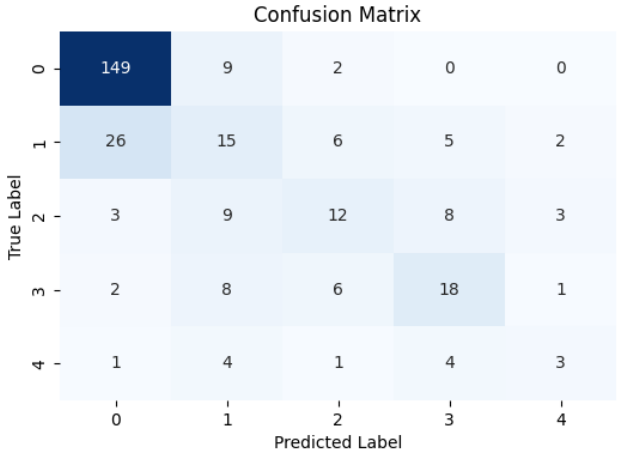

Figure 4: Confusion matrix

Our model was trained using the scaled training data, and predictions were made on the scaled test data. The model's accuracy was measured by the number of correct predictions made from all predictions, which is the model's accuracy score. The confusion matrix provides a more nuanced view of the model's performance, showing how predictions were distributed across the actual classes. In our case, the confusion matrix and accuracy score are consistent with each other, both indicating that the model is relatively accurate but not perfect.

It's worth noting that the logistic regression model might not prioritize the same features as linear regression. This is because logistic regression is dealing with probabilities and class separations, which might lead to different features being identified as significant compared to those in a linear regression model that predicts continuous outcomes. Additionally, the impact of outliers and noisy data might be different in logistic regression, potentially altering the importance of certain features.

In conclusion, the logistic regression model we trained is proficient at classifying instances into the correct categories, as evidenced by the accuracy score and confusion matrix. However, there's room for improvement, particularly in the classification of classes other than '0'. Future work could involve tuning the model further, possibly by adjusting the regularization strength, employing feature selection techniques, or gathering more data to help the model learn the distinctions between classes more efficiently.

Analysis of results – Discussion and Conclusion

The report analysed the performance and characteristics of different types of supervised learning models on the Heart Disease dataset. The report was divided into two sections, one on regression and another on classification. In the regression section, the report used one of the seven numerical attributes as the target for regression analysis while the other 14 attributes served as predictor variables. The report 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. The report found that the 'thalac' attribute had the strongest predictive power for regression analysis. In the classification section, the report used the 'num' attribute as the target

variable indicating the diagnosis and extent of heart disease. The report evaluated the performance of different classification models including k-NN, decision tree, random forest, and SVM. The report found that the SVM model had the highest accuracy and F1 score for classification. Overall, the report provided a comprehensive analysis of different supervised learning models on the Heart Disease dataset.

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 $\hat{y}$. 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:

$$\hat{y} = \{0.5, 0.6, 0.7, 0.8, 0.9\}$$

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 x_7=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.

$$\Delta = \mathrm{I(r)} - \sum_{k=1}^{K} \frac{N(v_k)}{N(r)} I(v_k)$$

$$Gini(v) = 1 - \sum_{c=1}^{C} p(c|v)^2$$

The total number of observations are 135. This results that there are 1 observation when x_7=2, and 134 observations when $x_7 \neq 2$. To calculate the impurity gain, one can use Gini´s method, which will result in the following:

$$I(x_7 = 2) = 1 - \frac{Total\ observations - no.of\ observations}{Total\ observations} = 1 - \frac{134}{135} \approx 0.0074$$

The correct statement is therefore option C.

Question 3. Spring 2019 Question 18:

Option A:

The total number of features in the dataset is X^{TOT} = 7. The assignment defines a single hidden layer with n_h = 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: x_1
- Ratio: x_2, …, x_7
- Nominal: x_1
- 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. $(X^{TOT} + 1) \cdot n_h = (7 + 1) \cdot 10 = 80$
2. $(n_h + 1) \cdot C = (10 + 1) \cdot 4 = 44$
3. Number of parameters: $80 + 44 = 124$

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 b_1, while the other nodes B and D are determined by b_2. This information is shown in Figure 1 below:

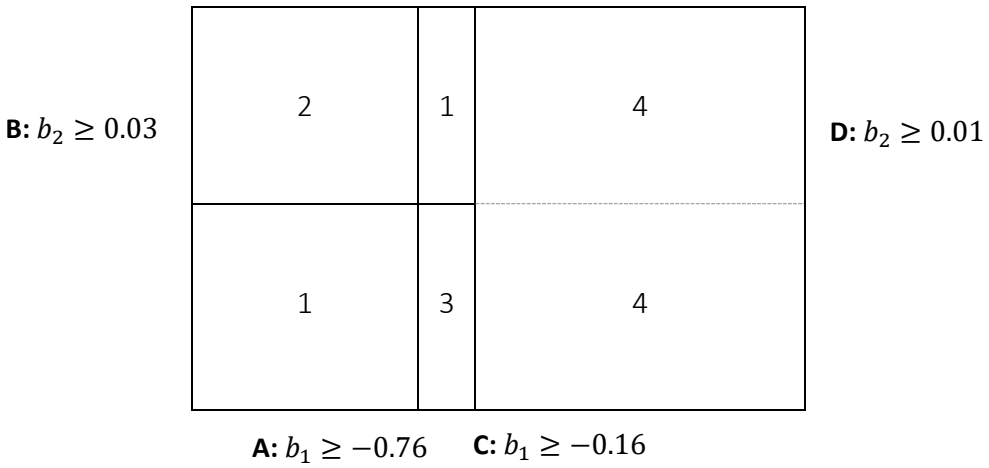

Figure 1: Combination of the classification boundaries and decision tree.

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.
- $\boldsymbol{b_1}$**:** Is the x-coordinate for PCA 1 in Figure 4 of the Classification boundary.
- $\boldsymbol{b_2}$**:** 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 n_h 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:

$$T_{Train+Test} = 20\ ms + 5\ ms + 8\ ms + 1\ ms = 34\ ms.$$

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, E_{gen}). 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:

$$\sum_{Time} = 5 * \big(34\ ms + 5 * 4 * (34\ ms)\big) = 3570\ ms.$$

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 w_k into the given equation to compute the y_k, for $k = 1, 2, 3$ respectively. Then we need to put the value of y_k, for $k = 1, 2, 3$ into the given softmax function to calculate the $P(y = k|y)$, for $k = 1, 2, 3, 4$. 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 $y = 4$ has the largest probability, so the option should be B.