Machine Learning Project 2
Supervised Learning: Classification and Regression

November 2, 2025

Group 94

Name	Student number
Vincent Van Schependom	s251739
Diego Armando Mijares Ledezma	s251777
Albert Joe Jensen	s204601

(a) Group members.

Task	Vincent	Diego	Albert
Training & test loops	0%	0%	0%
Coding visualisations	0%	0%	0%
Section 3	0%	0%	0%
LaTeX	0%	0%	0%

(b) Contributions & responsabilities table.

Table 1: Group information & work distribution.

Introduction

The objective of this report is to apply the methods that were discussed during the second section of the course *Machine Learning* [1] to a chosen dataset. The aim is to perform relevant regression and classification to the data.

The particular dataset that is being investigated is – just like in Project 1 – the *Glass Identification* dataset from 1987 by B. German [2]. Table 1a lists our full names and student numbers, while Table 1b shows an overview of the contribution of each team member.

Contents

1 Regression

We aim to predict the refractive index (`RI`) of glass samples from their chemical composition: the oxide components (`Na`, `Mg`, `Al`, `Si`, `K`, `Ca`, `Ba`, and `Fe`) and the categorical variable `Type`. Predicting `RI` is formulated as a regression problem where we evaluate the performance of a regularized linear model (Ridge Regression), an artificial neural network (ANN), and a trivial baseline predictor, which always predicts the mean `RI` of the training set. Numerical features are standardized, as argued previously in Report 1 [?], and categorical variables are encoded using one-hot encoding.

[TODO: Add report 1 to the bibliography.]

1.1 Regularized linear regression

1.1.1 Selecting the regularization parameter λ

To investigate the overfitting prevention effect of regularization, we evaluated Ridge Regression models across a wide range of regularization variable values from $\lambda = 10^{-2}$ to 10^5 (in 50 logarithmic steps) using $K = 10$-fold cross-validation. Figure 1 shows a logarithmic plot of the average mean squared error (MSE) on the validation sets across all folds – which is an estimate $\hat{E}^{\text{gen}}$ of the generalization error – as a function of λ.

The minimum average validation error was found at $\hat{\lambda} = 0.2683$ with an average validation set MSE over all 10 of 1.003×10^{-6}. We observe that the curve of MSE is relatively flat around the optimum, indicating a stable solution. As the regularization parameter λ increases above 10, and larger coefficients thus get more penalized, the model underfits, leading to a gradual rise in error to $\hat{E}^{\text{gen}} \approx 1 \times 10^{-5}$. Conversely, smaller values ($\lambda < 10^{-1}$) result in marginally higher average validation MSE due to overfitting.
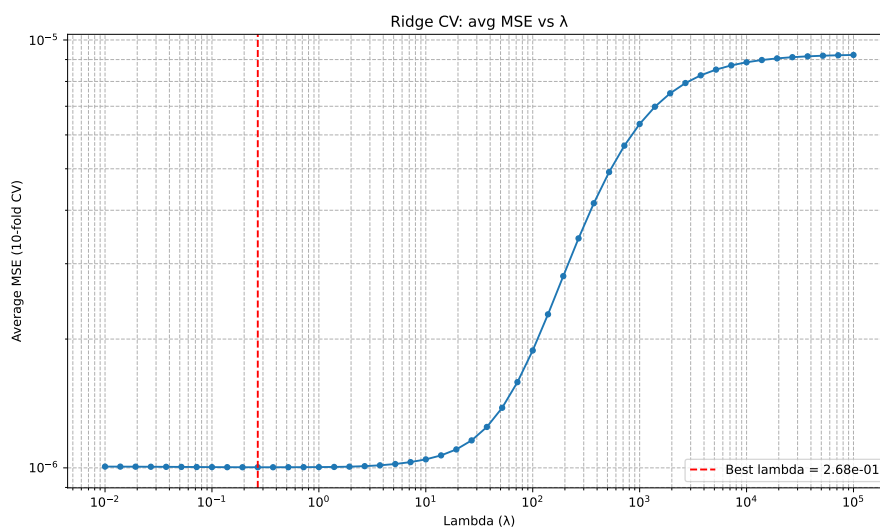

Figure 1: Mean squared error (MSE) versus regularization strength λ for Ridge regression using nested cross-validation. The optimal λ ($\hat{\lambda} = 0.2683$) achieved a minimum average MSE of 1.0033×10^{-6}.

[TODO: Create the figure that visualises the selected coefficients based on lambda (copy paste from latest exercise session and look at my two images in WhatsApp.)]

[TODO: Add these images here.]

[TODO: Add captions, reference from text and elaborate on observations.]

1.1.2 Model interpretation

A final Ridge Regression model was trained on the *entire* dataset using the optimal regularization parameter $\hat{\lambda} = 0.2683$. For a given input $\mathbf{x}$, the predicted refractive index $\hat{y} = \mathtt{RI}$ is given by:

$$
\begin{aligned}
\mathtt{RI} = \beta_0 \quad &+\beta_1 \cdot \mathtt{Na} & &+\beta_2 \cdot \mathtt{Mg} & &+\beta_3 \cdot \mathtt{Al} \\
&+\beta_4 \cdot \mathtt{Si} & &+\beta_5 \cdot \mathtt{K} & &+\beta_6 \cdot \mathtt{Ca} \\
&+\beta_7 \cdot \mathtt{Ba} & &+\beta_8 \cdot \mathtt{Fe} & &+\beta_9 \cdot \mathtt{BW\text{-}FP} \\
&+\beta_{10} \cdot \mathtt{BW\text{-}NFP} & &+\beta_{11} \cdot \mathtt{VW\text{-}FP} & &+\beta_{12} \cdot \mathtt{containers} \\
&+\beta_{13} \cdot \mathtt{headlamps} & &+\beta_{14} \cdot \mathtt{tableware} & &+\beta_{15} \cdot \mathtt{vehicle\ windows} \\
= 0 \quad &+0 \cdot \mathtt{Na} & &+0 \cdot \mathtt{Mg} & &+0 \cdot \mathtt{Al} \\
&+0 \cdot \mathtt{Si} & &+0 \cdot \mathtt{K} & &+0 \cdot \mathtt{Ca} \\
&+0 \cdot \mathtt{Ba} & &+0 \cdot \mathtt{Fe} & &+0 \cdot \mathtt{BW\text{-}FP} \\
&+0 \cdot \mathtt{BW\text{-}NFP} & &+0 \cdot \mathtt{VW\text{-}FP} & &+0 \cdot \mathtt{containers} \\
&+0 \cdot \mathtt{headlamps} & &+0 \cdot \mathtt{tableware} & &+0 \cdot \mathtt{vehicle\ windows}
\end{aligned}
$$

[TODO: Fill in the coefficients above based on the Python output.]

These coefficients reveal the relative contribution of each oxide concentration (β_1 to β_8) and glass type (one hot encoded β_9 to β_{15}) to the refractive index ($\mathtt{RI}$).

The elements $\mathtt{Ca}$ and $\mathtt{Ba}$ exhibit the largest positive coefficients, implying that increasing these oxides raises the estimated refractive index in this model. Conversely, $\mathtt{Si}$, $\mathtt{Al}$, and $\mathtt{Na}$ have negative coefficients, indicating that higher concentrations of these oxides decrease the estimated refractive index. The small magnitude of coefficients for $\mathtt{Mg}$ and $\mathtt{Fe}$ suggests a weaker influence on the estimated refractive index.

[TODO: Discuss the one-hot-encoded categorical variables here as well.]

1.1.3 Evaluation of the Ridge model

To verify the predictive performance of the final Ridge model, we evaluated it on the full dataset using the optimal $\hat{\lambda}$. This led to a mean squared error of $\text{MSE} = 8.5 \times 10^{-7}$ and a coefficient of determination of $R^2 = 0.91$.

Figure 2 shows predicted versus actual $\mathtt{RI}$ values. The points lie closely along the identity line, confirming accurate predictions. The residuals in Figure 2 are randomly distributed around zero, suggesting no systematic bias or heteroscedasticity.

1.2 Comparison of regression models

1.2.1 Two-level cross-validation

This section compares regularized linear regression (from the previous section), an artificial neural network (ANN), and a baseline model. Two-level cross-validation is employed to determine relative performance of these models. As complexity-controlling parameter for the ANN, the number of hidden units h is varied from $h = 1$ to $h = 990$ in steps of 10. The baseline models predicts the mean $\mathtt{RI}$ of the training set for all inputs.

The nested cross-validation results for the regression tasks are summarized in Table 2. [TODO: Elaborate on the results of the table!]

Technical University of Denmark - 02452 Machine Learning
Machine Learning Project 2 - *Supervised Learning: Classification and Regression*
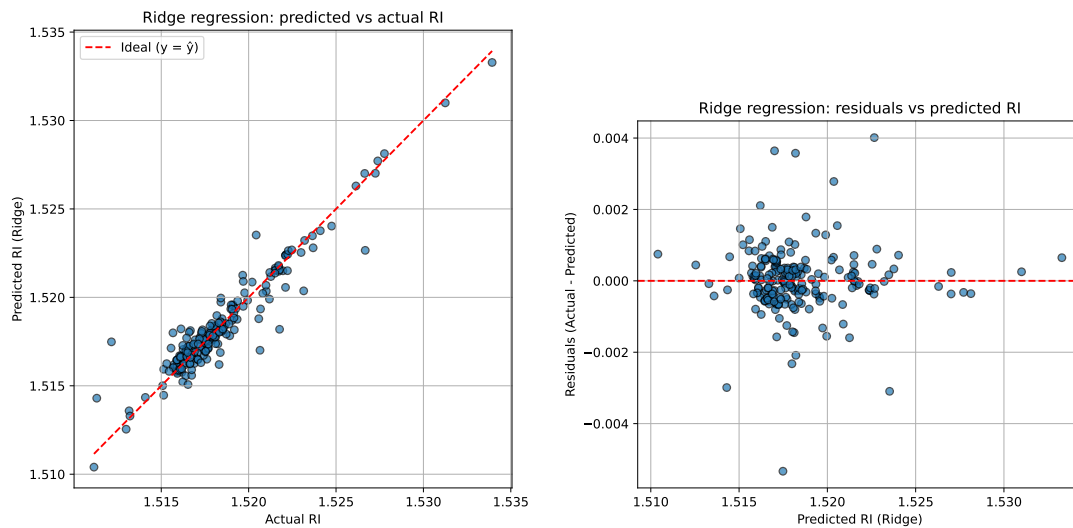Group 94

3/9

Figure 2: Verification of the regularized Ridge regression model. Left: predicted versus actual refractive index (RI). Right: residuals versus predicted RI. The model shows a high determination coefficient $R^2 = 0.91$ and randomly distributed residuals, confirming stable behavior.

1.2.2 Statistical significance testing

Paired t-tests were conducted across folds to assess whether performance differences were statistically significant. The results are summarized in Table 3. All p-values below 0.05 indicate statistically significant differences.

[TODO: Check if we need to elaborate further / if this text contains mistakes:]

- **ANN vs RLR:** The ANN achieved a lower average test error, with a mean difference of 0.0242 ± 0.0169 ($t = 4.54$, $p = 0.0014$). This indicates the ANN significantly outperforms Ridge regression on the RI prediction task.

- **ANN vs Baseline:** The ANN also significantly outperformed the baseline ($t = 4.54$, $p = 0.0014$), confirming that it captures non-linearities missed by linear models.

- **RLR vs Baseline:** Ridge regression slightly outperformed the baseline (mean difference -8.26×10^{-6}, $t = -4.25$, $p = 0.0021$), confirming that even a regularized linear model provides small but significant improvements over the naive predictor.

The ANN's advantage is thus both statistically and practically significant, while the Ridge regression, although simpler, remains interpretable and competitive.

2 Classification

2.1 Introduction

We aim to predict the type of glass (`Type`), which can take on 7 values, of which only 6 are present in the training set, since there are no observations with `Type` = `VW-NFP`. We will use the oxide concentrations as predictors to handle this *multi-class classification problem*. We compare Logistic Regression (LR), a Classification Tree (Decision Tree, DT), and a trivial baseline that always predicts the most frequent class present in the training set. All models were evaluated with nested (two-level) cross-validation and then statistically compared using paired t-tests across outer folds.

Outer fold	ANN		Linear regression		Baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	500	2.607×10^{-2}	0.797	1.600×10^{-6}	1.970×10^{-5}
2	480	1.327×10^{-2}	2.84	4.249×10^{-7}	1.228×10^{-5}
3	230	3.239×10^{-2}	0.325	4.821×10^{-7}	4.482×10^{-6}
4	190	1.592×10^{-2}	0.167	9.633×10^{-7}	5.047×10^{-6}
5	120	1.030×10^{-2}	1.43	1.500×10^{-6}	3.080×10^{-6}
6	500	3.226×10^{-3}	0.01	2.142×10^{-6}	5.982×10^{-6}
7	120	1.603×10^{-2}	0.325	1.284×10^{-6}	8.186×10^{-6}
8	240	3.639×10^{-2}	0.01	1.018×10^{-6}	7.210×10^{-6}
9	290	6.221×10^{-2}	1.27	1.637×10^{-7}	1.946×10^{-5}
10	500	2.641×10^{-2}	3	5.225×10^{-7}	7.245×10^{-6}

Table 2: Summary of two-level cross validation for predicting the refractive index `RI` based on the chemical composition. Hyperparameters and test errors E_i^{test} on $\mathcal{D}_i^{\text{test}}$ per outer fold i, for each of the three considered models.
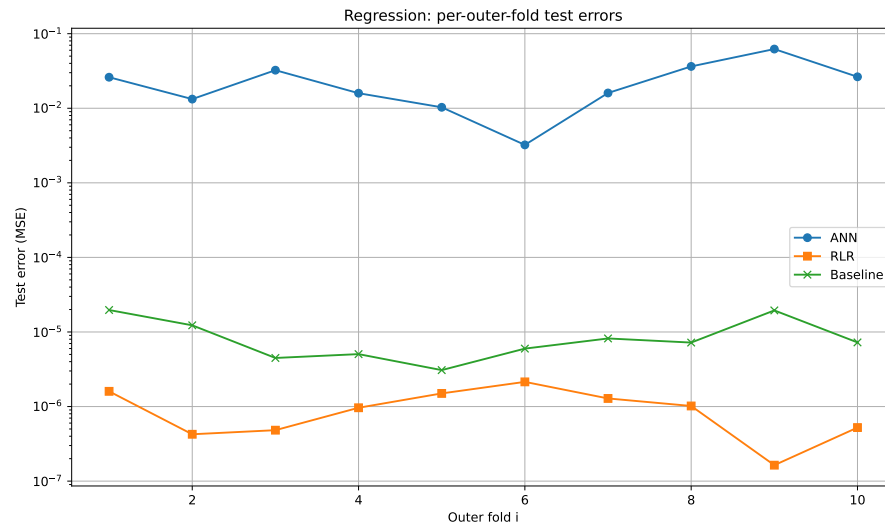

Figure 3: Test errors E_i^{test} across outer folds for each regression model. [TODO: Elaborate on the observations in this caption.], [TODO: Reference this figure.], [TODO: Mention this is a semilog plot.]

Model comparison	$\hat{\mu}$	$\hat{\sigma}$	t-statistic	p-value	$t_{0.05}$	$t_{0.95}$
ANN vs RLR	0	0	0	0	0	0
ANN vs Baseline	0	0	0	0	0	0
RLR vs Baseline	0	0	0	0	0	0

Table 3: Pairwise comparison of regression models (ANN = Artificial Neural Network, RLR = Regularized Linear Regression) using paired t-tests across outer folds. Mean differences $\hat{\mu}$ and standard deviations $\hat{\sigma}$ in test error E^{test}, t-statistics, and p-values are reported, as well as a 95% confidence interval $[t_{0.05}, t_{0.95}]$. All comparisons show statistically significant differences ($p < 0.05$). [TODO: Fill in the numbers based on the Python output.]
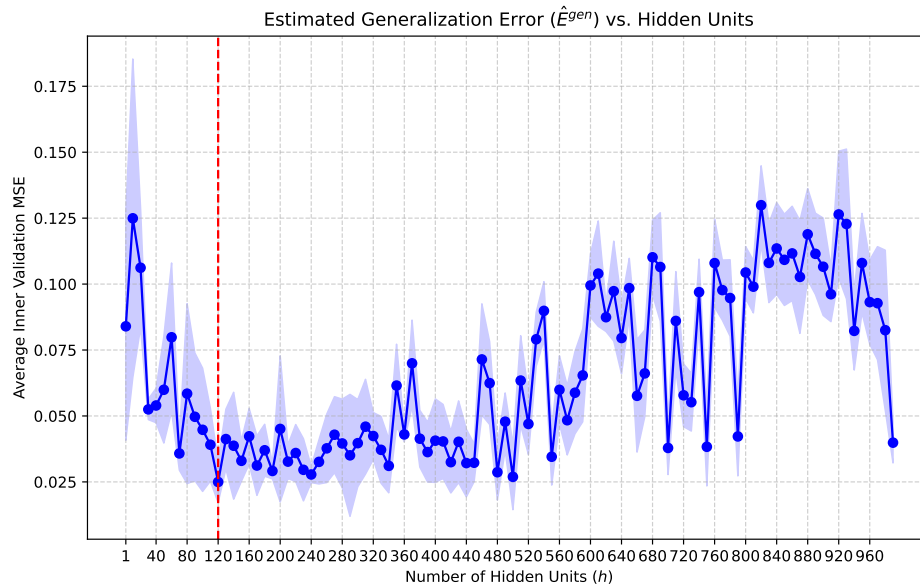
Technical University of Denmark - 02452 Machine Learning
Machine Learning Project 2 - *Supervised Learning: Classification and Regression*
Group 94

5/9

Figure 4: Mean squared error (MSE) on the validation set versus number of hidden units h for the artificial neural network using nested cross-validation. The optimal h ($\hat{h} = 120$) achieved a minimum average MSE of 1.03×10^{-2}. [TODO: Reference this figure!]

Outer fold	Decision Tree		Logistic regression		Baseline
i	h_i^*	E_i^{test}	λ_i^*	E_i^{test}	E_i^{test}
1	3	0.4091	0.336	0.2273	0.5455
2	6	0.2273	0.0785	0.3182	0.5000
3	15	0.3182	0.00886	0.4091	0.8182
4	3	0.3636	0.0379	0.4545	0.9091
5	5	0.3333	0.00886	0.2857	0.5714
6	6	0.4762	0.00428	0.4286	0.8095
7	5	0.2857	0.00428	0.3333	0.6667
8	10	0.2857	0.336	0.3333	0.7619
9	4	0.3333	0.00428	0.3810	0.6667
10	7	0.3333	0.695	0.4762	0.8095

Table 4: Summary of two-level cross validation for predicting the glass type. Hyperparameters and test errors E_i^{test} on $\mathcal{D}_i^{\text{test}}$ per outer fold i, for each of the three considered classificationmodels. [TODO: Change λ to C?]

Technical University of Denmark - 02452 Machine Learning
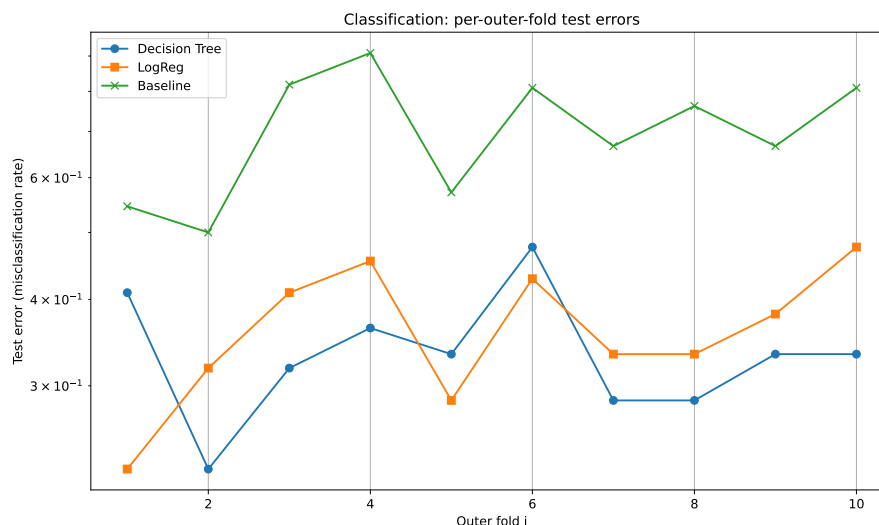Machine Learning Project 2 - *Supervised Learning: Classification and Regression*
Group 94

6/9

Figure 5: Test errors E_i^{test} across outer folds for each classification model. Both the Logistic Regression and Decision Tree outperform the baseline in all folds, with Logistic Regression achieving slightly lower error on average., [TODO: Mention this is a semilog plot.]

Model comparison	$\hat{\mu}$	$\hat{\sigma}$	t-statistic	p-value	$t_{0.05}$	$t_{0.95}$
DT vs LR	0	0	0	0	0	0
DT vs Baseline	0	0	0	0	0	0
LR vs Baseline	0	0	0	0	0	0

Table 5: Pairwise comparison of classification models (DT = Decision Tree, LR = Logistic Regression) using paired t-tests across outer folds. Mean differences $\hat{\mu}$ and standard deviations $\hat{\sigma}$ in test error E^{test}, t-statistics, and p-values are reported, as well as a 95% confidence interval $[t_{0.05}, t_{0.95}]$. All comparisons show statistically significant differences ($p < 0.05$). [TODO: Fill in the numbers based on the Python output.]

2.2   Comparison of classification models

This section compares Logistic Regression and a Decision Tree classifier against the baseline. For each outer fold, we optimized the regularization strength $C = 1/\lambda$ for Logistic Regression and the maximum tree depth h for the Decision Tree using inner cross-validation. We then computed the test error E_i^{test} on each outer test fold.

[TODO: Mention the results in the table, reference the table using the following command and elaborate on them!]: \ref{table:e-test-classification}

As shown in Table 4, ... [TODO: Fill in based on the results!]

2.3   Statistical significance testing

Paired t-tests across outer folds were used to assess statistical significance between the three classification models. The results are summarized in Table 5.

[TODO: Fill in the table below based on the Python output.]

We summarize the findings of these statistical tests below:

[TODO: Check if we need to elaborate further / if this text contains mistakes. I don't know why, but Diego (or his AI) wrote two versions of this... So read both of them and pick the best one / merge them.]

- **Tree vs LogReg:** The difference in accuracy (-0.0281) is not statistically significant ($p = 0.375$), suggesting both models perform similarly.

- **Tree vs Baseline:** The decision tree significantly outperformed the baseline ($t = -8.91$, $p < 10^{-5}$), indicating it captures meaningful structure in the data.

- **LogReg vs Baseline:** Logistic regression also significantly outperformed the baseline ($t = -13.34$, $p = 3.1 \times 10^{-7}$), showing that even a simple linear classifier captures predictive signal.

- **Tree vs LogReg:** The mean difference in error was -0.0281 ± 0.0953 ($t = -0.93$, $p = 0.375$), indicating no statistically significant difference. Both models perform comparably.

- **Tree vs Baseline:** The Decision Tree significantly outperformed the baseline with $t = -8.91$ and $p = 9.25 \times 10^{-6}$, showing that even a simple tree captures predictive relationships among the oxide features.

- **LogReg vs Baseline:** Logistic Regression also achieved a strong improvement over the baseline ($t = -13.34$, $p = 3.12 \times 10^{-7}$), confirming that linear decision boundaries explain most of the separability in the dataset.

In summary, both classification models substantially improve upon the baseline, and their similar performance suggests that the data can be effectively separated by relatively simple decision boundaries.

Figure 5 visualizes the outer test errors per fold for the three models, clearly showing that both Logistic Regression and the Decision Tree consistently outperform the baseline classifier.

2.4 Interpretation of the LR model

A final Logistic Regression model was trained on the *full* dataset using the selected regularization parameter $C^* = 1/\hat{\lambda}$ from cross-validation. The model assigns a weight vector $\boldsymbol{w}_k$ to each class k, such that class scores are

$$s_k(\boldsymbol{x}) = w_{k,0} + \boldsymbol{w}_k^\top \boldsymbol{x},$$

and predicted probabilities are given by the softmax transformation.

[TODO: Check this part below for mistakes / elaborate further if needed:]

The sign and magnitude of each coefficient indicate how increasing a given oxide concentration affects the log-odds of a sample belonging to class k. Features such as `Si`, `Ca`, and `Ba` were among the strongest predictors, consistent with the results from the regression analysis, where these elements also exhibited large coefficients in determining refractive index. This agreement suggests that the same chemical components governing RI also drive differences between glass types.

Overall, Logistic Regression achieves competitive classification accuracy with interpretable coefficients, while the Decision Tree offers slightly lower performance but more explicit decision boundaries.

Use of GenAI

Generative AI (ChatGPT by OpenAI and Claude by Anthropic) was used as a support tool during the project.

Its role was limited to assisting with code debugging, identifying and fixing minor implementation errors, and – for Diego – writing drafts for parts of the report – which were later heavily revised to form a cohesive final report.

All analytical choices, data processing steps, and result interpretations were made by the project members *themselves* based on the actual model outputs and *their* understanding of the underlying methods.

Technical University of Denmark - 02452 Machine Learning
Machine Learning Project 2 - *Supervised Learning: Classification and Regression*
Group 94

8/9

References

[1] Tue Herlau, Mikkel N. Schmidt, and Morten Mørup. *Introduction to Machine Learning and Data Mining*. Technical University of Denmark (DTU), Lyngby, Denmark, 2023. Lecture notes, Fall 2023, version 1.0. This document may not be redistributed. All rights belong to the authors and DTU.

[2] B. German. Glass Identification. UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5WW2P.

Appendix

A   Repository and supplementary materials

The full notebook, scripts, and generated figures for this project are available in the project repository:

`https://github.com/schependom/DTU_machine-learning-projects/tree/main`

This repository contains the data-loading and analysis code that produced the tables and figures cited above (see the `figures/` folder for the PDF outputs referenced in the report).