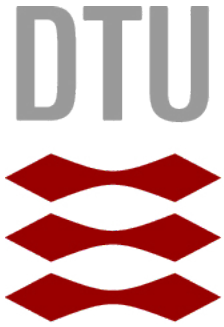

Technical University of Denmark

02450 Introduction to Machine Learning and Data Mining

Final assignment

Authors

Flavio Sarno s242991
Florencia Illanes s247222
Francesco Balducci s250200

April 10th, 2025

Contents

1 Contributions - Part 1 **1**

2 Introduction - Part 1 **1**
2.1 Literature review . . . 1
2.2 Goals . . . 1
2.2.1 Classification . . . 2
2.2.2 Regression . . . 2

3 Dataset attributes **2**
3.1 Observations tuning . . . 2
3.1.1 Missing values . . . 3
3.1.2 Standardization . . . 3
3.1.3 Correlation . . . 3

4 Data visualization **3**

5 Principal Component Analysis **7**

6 Discussion - Part 1 **9**

7 About LLM usage - Part 1 **10**

8 Contributions - Part 2 **11**

9 Introduction - Part 2 **11**

10 Regression Analysis **11**
10.1 Part A . . . 12
10.2 Part B . . . 13

11 Classification Analysis **15**

12 Discussion - Part 2 **16**
12.1 Regression Analysis . . . 16
12.2 Classification Analysis . . . 17
12.3 Previous Studies . . . 17

13 About LLM usage - Part 2 **18**

1 Contributions - Part 1

All participants in the assignment have compiled and/or reviewed the present document and all the source codes. Nevertheless, for organizational reasons, we decided to be responsible for different parts of the document, as shown in the following *(not exhaustive)* table:

	Flavio Sarno (s242991)	**Francesco Balducci (s250200)**	**Florencia Illanes (s247222)**
Introduction	30%	30%	40%
Dataset description	40%	40%	20%
Data visualization	30%	45%	25%
PCA	50%	30%	20%
Discussion	20%	20%	60%

Table 1: Table of responsibilities

2 Introduction - Part 1

The data set [1] features several engine characteristics and specifications of cars built between 1970 and 1982. The data set investigates city-cycle fuel consumption for cars built in three different geographical areas: the United States, Europe, and Japan.

The available properties are: mpg, cylinders, displacement, horsepower, weight, acceleration, model year, origin, and car name.

Generally speaking, the problem we would like to solve is to predict or compute relationships between these characteristics, mainly fuel efficiency for cars with different attributes. We will leverage machine learning techniques to tackle this task.

2.1 Literature review

This dataset was previously used in the American Statistical Association (ASA) Exposition in 1983. The data was collected by Ernesto Ramos and David Donoho, initially containing a total of 406 observations along with the same variables present in the current version [2]. The dataset provided by Ross Quinlan has 8 fewer observations due to missing values in the original database. Quinlan used this data to identify patterns in fuel consumption by combining instance-based and rule-based models. Additionally, Quinlan (1993) explored different machine learning approaches, including decision trees, linear regression, and neural networks, to improve fuel consumption prediction. His research demonstrated that combining instance-based learning with model-based methods, such as model trees, improved predictive accuracy compared to using each method separately. [3]

2.2 Goals

The features have been analyzed through Machine Learning techniques with the objective of (1) identifying and representing the relations or trends between them, (2) reducing the size of the data set while preserving the core information, (3) modeling a classification engine to predict the *origin* feature based on the rest of the data, and (4) another model to perform regression on the feature miles per gallon *(mpg)* feature, the vehicle's fuel efficiency, based on all other characteristics.

2.2.1 Classification

We would like to classify existing and unknown observations around the *origin* feature through the remaining noncategorical properties *(mpg, cylinders, displacement, horsepower, weight, acceleration, model year)*. Doing so will allow the user to obtain useful information about the relationship that exists between the characteristics of a car and the expected area of utilization.

2.2.2 Regression

The main objective of this analysis is to predict the fuel efficiency of a car, measured in miles per gallon *(mpg)*, based on various technical characteristics of the car. The selected predictive features are *number of cylinders, displacement, weight, horsepower, acceleration, and model year*. In simple terms, the goal is to model fuel consumption based on those key vehicle specifications.

3 Dataset attributes

Feature	Description	Attribute	Type
mpg	Miles per gallon (fuel efficiency)	Continuous	Ratio
cylinders	Number of engine cylinders	Discrete	Ordinal
displacement	Engine displacement (cubic inches)	Continuous	Ratio
horsepower	Engine power output	Continuous	Ratio
weight	Vehicle weight (lbs)	Continuous	Ratio
acceleration	Time to accelerate from 0 to 60 mph (seconds)	Continuous	Ratio
model_year	Model year of the vehicle	Discrete	Interval
origin	Country of origin (1: USA, 2: Europe, 3: Japan)	Discrete	Nominal
car_name	Name of the vehicle model	Discrete	Categorical

Table 2: Description and classification of dataset features based on attribute type. This dataset uses the imperial system of units [1].

The following table shows a summary of the dataset values based on the different ratio attributes, as measured by the most common statistics:

Statistic	mpg	displacement	horsepower	weight	acceleration
Count	392.00	392.00	392.00	392.00	392.00
Mean	23.45	194.41	104.47	2977.58	15.54
Std	7.81	104.64	38.49	849.40	2.76
Min	9.00	68.00	46.00	1613.00	8.00
25%	17.00	105.00	75.00	2225.25	13.78
50%	22.75	151.00	93.50	2803.50	15.50
75%	29.00	275.75	126.00	3614.75	17.03
Max	46.60	455.00	230.00	5140.00	24.80

Table 3: Summary statistics of the ratio features in the dataset.

As widely known, these metrics should not be considered as the only reliable (if any) representation of the data, but rather as a reference. Please refer to the data visualization section for in depth data illustration.

3.1 Observations tuning

The following sections will review modifications made to the dataset in order to improve its performance for the intended goals.

3.1.1 Missing values

Given that the dataset consists of **398** observations, and considering that there are only **6** observation with missing properties (all on the *horsepower* attribute), they will be excluded from the analysis in order to maintain data integrity.

3.1.2 Standardization

The various attributes happen to have very different scaling: for example, *weight* is continuous 1613.0÷5140.0 while *cylinders* is 3÷8 and discrete.

Therefore, we decided to apply feature standardization to the numerical ones (all except *car_name*) ensuring that the different magnitudes would not influence the performance of the algorithms, such as Principal Component Analysis (PCA). Standardization rescales the data such that each feature has a mean of 0 and a standard deviation of 1 and was applied through the following formula:

$$z_{ij} = \frac{x_{ij} - \mu_i}{\sigma_i} \tag{1}$$

Where:

- z_{ij} is the standardized value of feature i for observation j,
- x_{ij} is the original value of feature i for observation j,
- μ_i is the mean of feature i across all observations,
- σ_i is the standard deviation of feature i across all observations.

3.1.3 Correlation

The following table shows how related the attributes in the dataset are. By displaying correlation coefficients, it helps identify which features are strongly or weakly related to the target variable, allowing for better feature selection in machine learning models.

	mpg	cylinders	displacement	horsepower	weight	acceleration
mpg	1.0000	-0.7776	-0.8051	-0.7784	-0.8322	0.4233
cylinders	-0.7776	1.0000	0.9508	0.8429	0.8975	-0.5047
displacement	-0.8051	0.9508	1.0000	0.8973	0.9330	-0.5438
horsepower	-0.7784	0.8430	0.8973	1.0000	0.8645	-0.6892
weight	-0.8322	0.8975	0.9330	0.8645	1.0000	-0.4168
acceleration	0.4233	-0.5047	-0.5438	-0.6892	-0.4168	1.0000

Table 4: Correlation Matrix

4 Data visualization

We created a pairplot to represent each attribute value against each other in pairs. This has been crucial for getting an initial overview of the attribute values and identifying possible trends or correlations between them.

Many interesting (and sometimes expected) trends can be observed from the plot. For example, the cylinders vs. mpg plot shows that cars with more cylinders tend to consume more fuel. Similar patterns are seen with cars having higher displacement or greater weight. Other manufacturing-related trends include the fact that cars with more cylinders generally have higher (and more varied) displacement, horsepower, and weight, and take less time to accelerate from 0-60 mph. Some of these trends will be analyzed in further detail later. On the other hand, some correlations are less clear (or non-existent), particularly those related to acceleration, such as the relationship with mpg or weight.

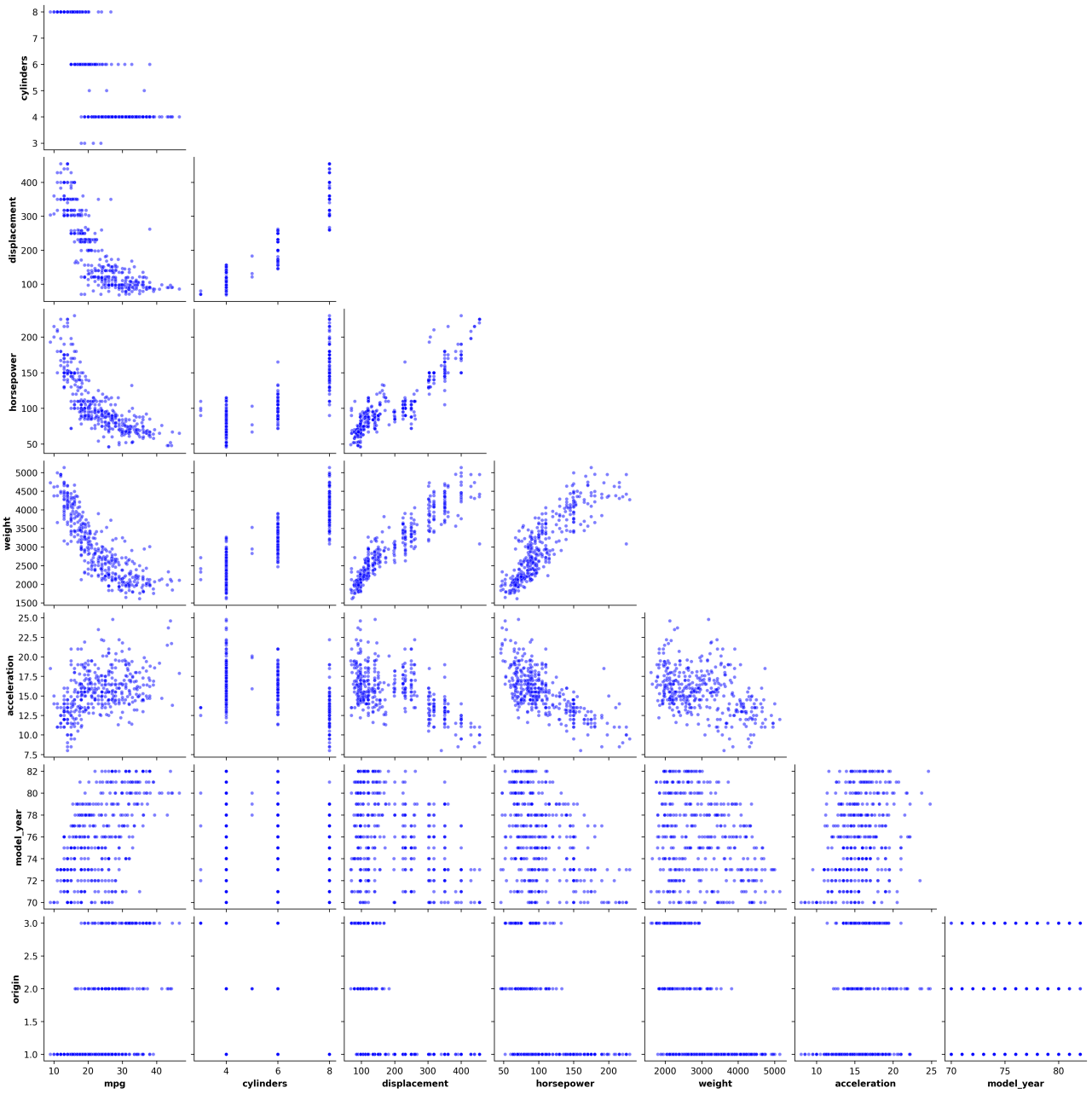

Figure 1: Pairplot showing pairwise relationships and correlation among the attributes of interest in the dataset. Only numerical attributes are displayed. Some degree of correlation is visible between almost all pairs of attributes.

Box plots are a suitable way to graphically display some of the data already presented in summary statistics table with the addition of a look at outliers, which will be analyized in the discussion seciton.

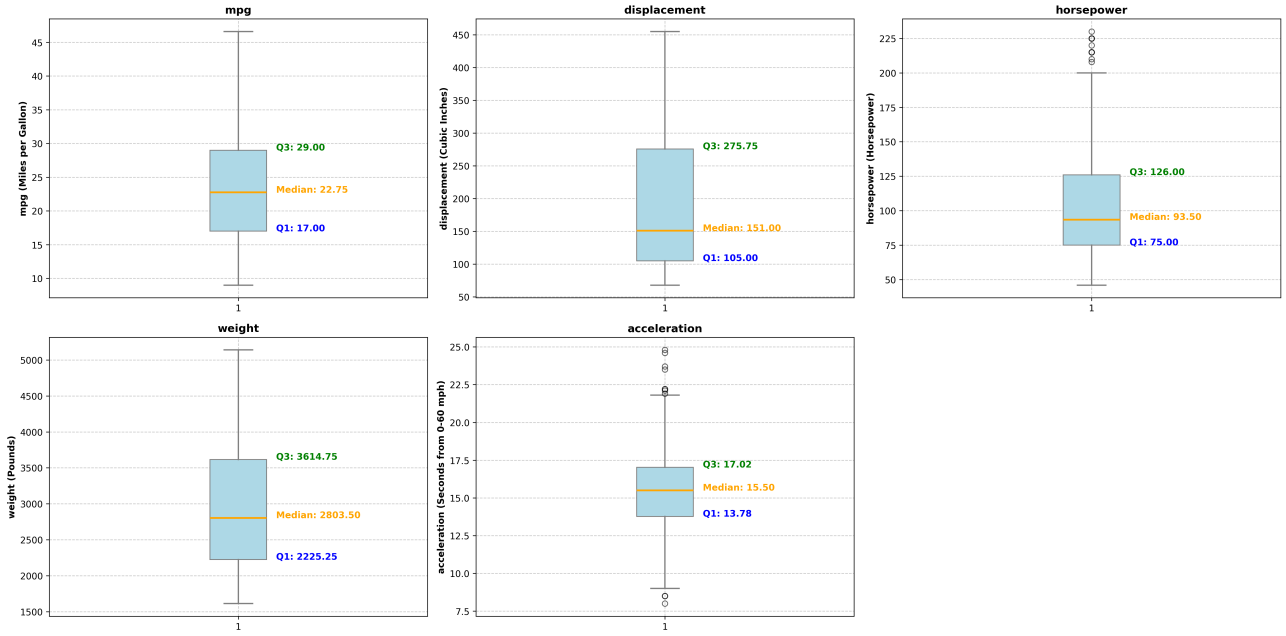

Figure 2: Box plots summarising the data set and displaying potential outliers values for each continuous attribute. In fact, *horsepower* and *acceleration* show a significant number of them.

The following histograms display the distribution of values for each continuous attribute. At first glance, only acceleration appears to follow a near-normal distribution, while the other attributes exhibit skewed distributions.

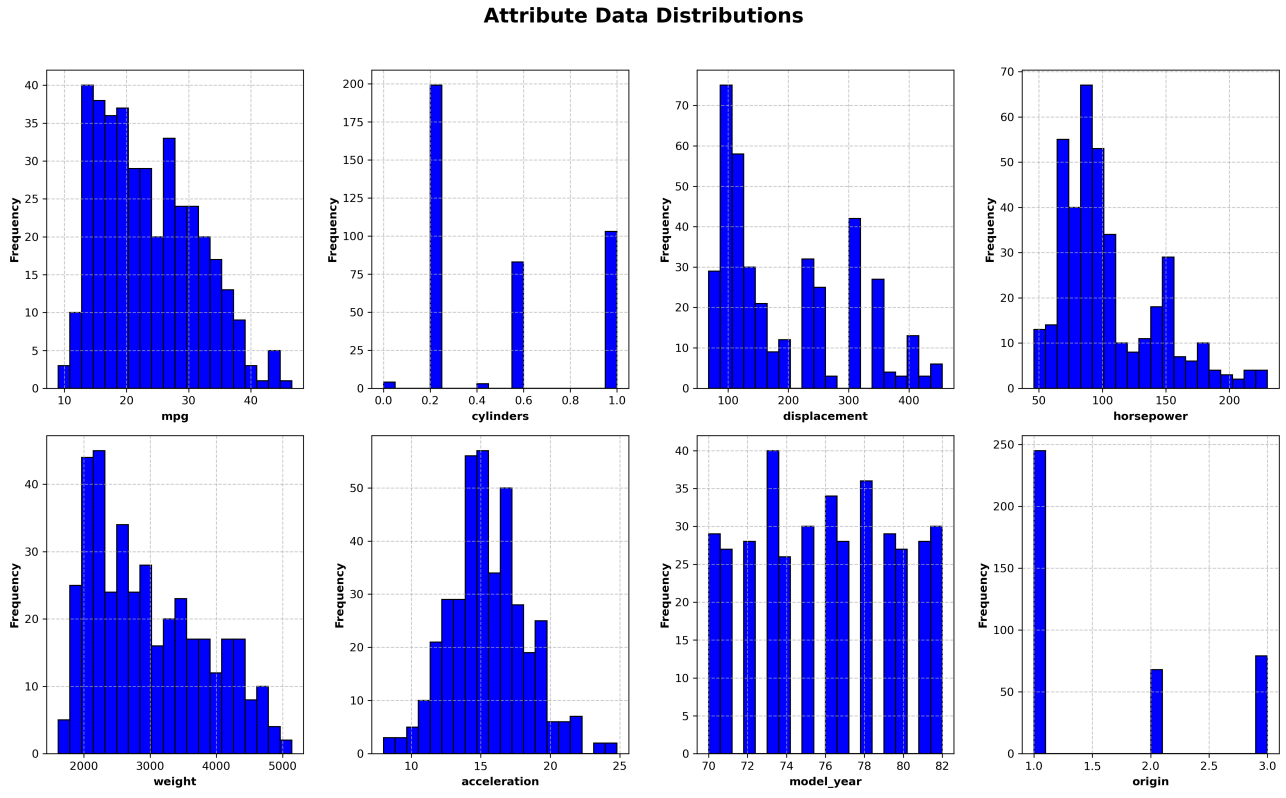

Figure 3: Histograms displaying the data distribution of all continuous attributes (plus standardized cylinders).

For the feasibility of our classification task objective, we also examined one of the attributes that, by intuition, might influence the *origin* (target label). The following boxplot clearly shows that the least efficient cars are primarily observed in the USA, while Europe and Japan tend to prefer more fuel-efficient cars, possibly due to different regulations or fuel costs. This provides an initial indication that the dataset contains information that could ultimately be useful for predicting the continent in which a car may be sold.

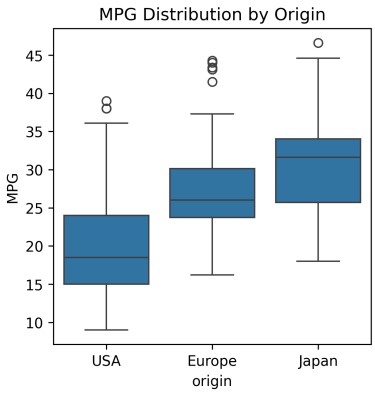

Figure 4: Boxplot of *mpg* by *origin*

Thinking about the regression task objective, we decided to visualize the predictability of MPG based on some key attributes that intuitively suggest a car may be less efficient. We expected that heavier cars, with high displacement, high horsepower, and many cylinders, would be more performant and, therefore, consume more fuel. This will also be one of the main findings of the PCA analysis.

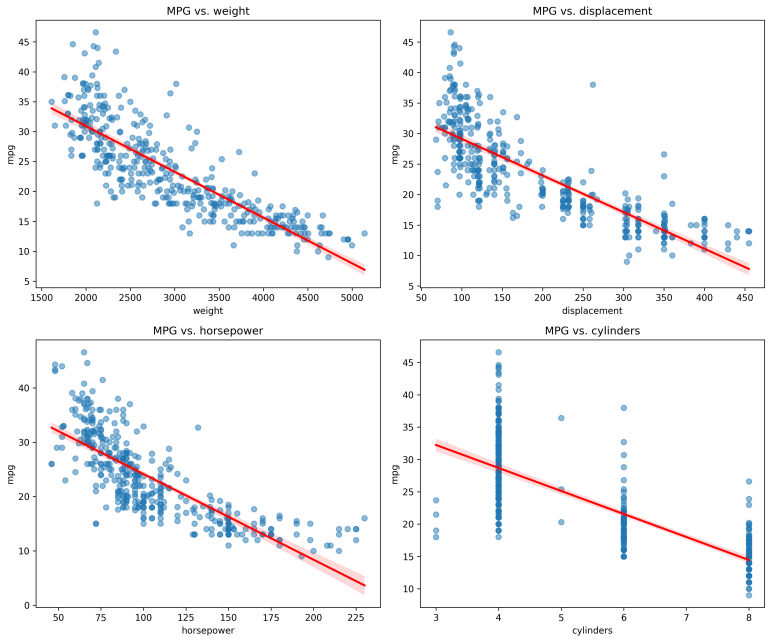

Figure 5: Comparison between MPG and some attributes that intuitively may influence it.

The heatmap below provides a more intuitive and visually appealing representation of the values obtained in the correlation table, making it easier to identify patterns and relationships between variables at a glance.

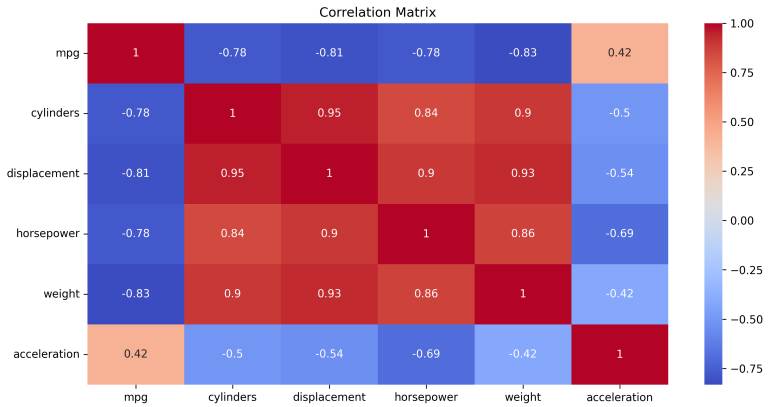

Figure 6: Correlation Matrix

5 Principal Component Analysis

Principal Component Analysis (PCA) is a powerful tool in Machine Learning that helps simplify complex datasets. When working with data, we often have many features (columns), and it's not always clear which ones are essential. Removing features without careful consideration can mean losing valuable information, while keeping everything might introduce noise or distortions due to differences in scale.

Some features might not contribute much to the overall representation of the data, while others may dominate just because of their magnitude, leading to biased results in downstream modeling. To address this, we apply Singular Value Decomposition to the standardized matrix:

$$\tilde{X} = U\Sigma V^T \tag{2}$$

Where:

- $\tilde{X}$ is the standardized data matrix of size $m \times p$, where m is the number of observations and p is the original number of features (see later).
- U is an orthogonal matrix of size $m \times m$, whose columns are the eigenvectors of $\tilde{X}\tilde{X}^T$. These represent the principal directions in the observation space.
- Σ is a diagonal matrix of size $m \times p$ containing the singular values in descending order, which indicate the contribution of each principal component to the retention of variance.
- V^T is the transpose of an orthogonal matrix V of size $p \times p$, whose columns are the eigenvectors of $\tilde{X}^T\tilde{X}$. These define the principal directions in the feature space.

As recommended, we included in $\tilde{X}$ only ratio attributes such as displacement, weight, horsepower, acceleration, and mpg, but we also decided to keep the (also standardized) discrete attribute *cylinders*. We believed it played a crucial role in representing the dataset accurately, ensuring we retained important structural information.

After obtaining the Σ and V matrices through the `dtumiltools` Python library, we can extract the first n principal components based on our threshold for cumulative explainability, which can be analyzed from the diagonal values of the Σ matrix.

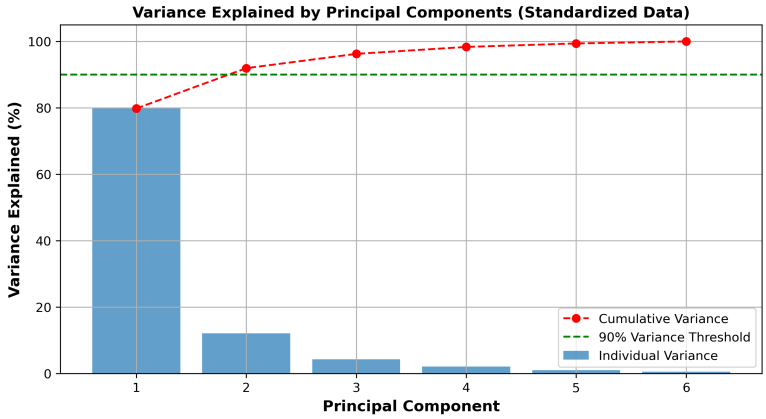

Figure 7: The bar plot represents the variance explained by each principal component (PC). The red dashed line displays the cumulative variance explained by each component. The first two principal components together account for more than 90% of the total variance, as indicated by the green threshold line. This justifies reducing the dataset to only two components, as they retain most of the original information while significantly reducing the dimensionality.

Attribute	**PC1**	**PC2**
MPG	0.39897	0.24483
Cylinders	-0.43062	-0.14831
Displacement	-0.44353	-0.10850
Horsepower	-0.43412	0.16616
Weight	-0.43010	-0.28610
Acceleration	0.29193	-0.89265

Table 5: Principal Component 1 (PC1) mainly separates heavy, high-displacement cars with more cylinders from lighter, fuel-efficient cars, while Principal Component 2 (PC2) captures differences in acceleration and weight, distinguishing fast-accelerating from slower-accelerating vehicles.

Once the n principal directions are selected, we project the entire dataset onto them using the following formula:

$$Z = \tilde{X} V_n \tag{3}$$

Where:

- Z is the transformed dataset of size $m \times n$, where each observation is now represented in the principal component space.
- $\tilde{X}$ is the standardized data matrix of size $m \times p$, where p is the original number of features.
- V_n is the matrix of the first n principal components, extracted from V, of size $p \times n$.

Thus, we obtain a new dataset of size $m \times n$, which serves as the most informative representation of the original data. This transformed dataset is now ready for further processing tailored to the intended tasks.

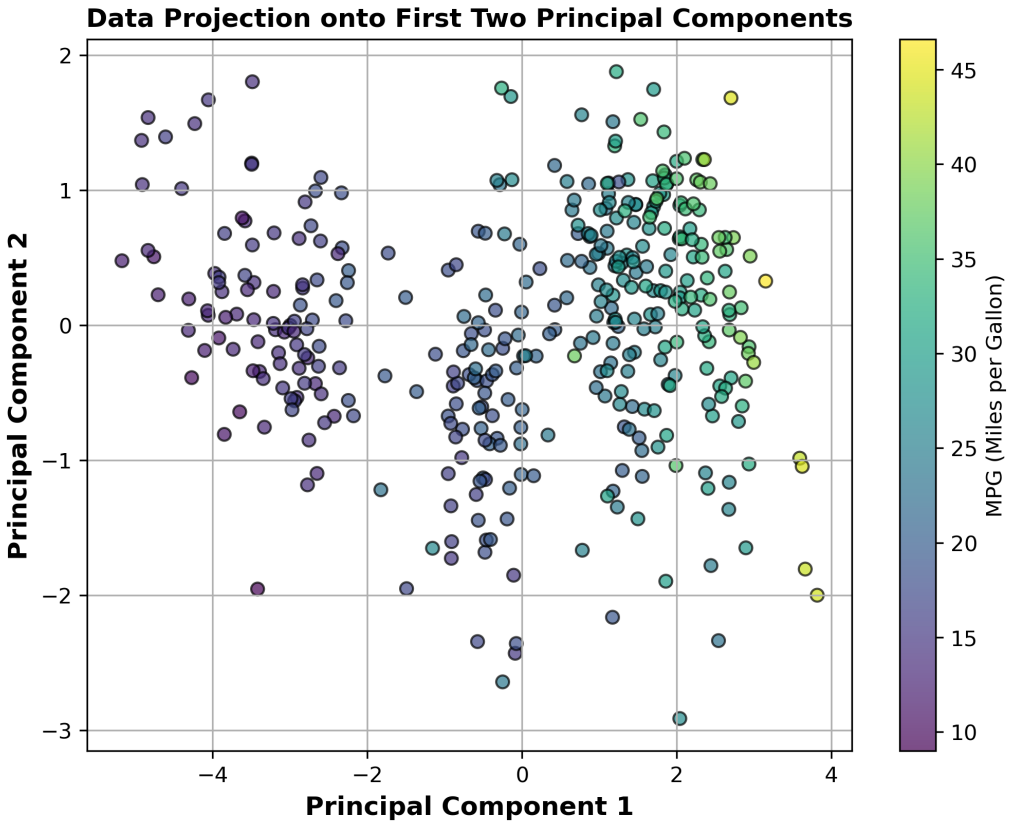

Figure 8: This scatter plot displays the dataset projected onto PC1 (horizontal axis) and PC2 (vertical axis), with a color-coding representing the *mpg* values (the target attribute of the classification task). PC1 (x-axis) mainly separates heavy, high-displacement cars on the left from lighter, fuel-efficient cars on the right, while PC2 (y-axis) captures differences in acceleration and weight, separating fast-accelerating cars at the top from slower-accelerating cars at the bottom. This projection shows how the relationship between fuel efficiency and vehicle characteristics is well preserved in the projected PCA data.

6 Discussion - Part 1

In summary, this dataset contains 392 observations after filtering out records with missing values. These observations correspond to cars and include 8 features: mpg, cylinders, displacement, horsepower, weight, acceleration, model year, origin, and car name. It is important to note that the numerical values have been standardized to ensure consistency in the analysis.

Based on the data visualization and the matrix correlation, it is observed that the features are highly correlated (as common sense might suggest), particularly *mpg*, with *displacement*, *horsepower*, and *weight*. A clear linear trend is visible among these attributes in Figure 1, while the correlation table indicates that as characteristics such as cylinders, displacement, horsepower, and weight increase, fuel efficiency (mpg) tends to decrease. This strong and negative correlation is consistent with physical intuition, as heavier cars or those with larger engines generally consume more fuel.

Additionally, the boxplots indicate that outliers are present only in horsepower and acceleration. However, these values in the dataset are not problematic because they are not improbable in the context of car characteristics. In other words, these extreme values are possible for attributes like horsepower and acceleration. Furthermore, the acceleration and horsepower values for these outlier cars do not diverge much from the IQR and represent realistic engine specifications of more powerful vehicles. Therefore, they do not represent erroneous data points that need to be adjusted or removed.

Furthermore, Principal Component Analysis (PCA) reveals that the first two principal components together ac-

count for more than 90% of the total variance. This high level of explained variance suggests that the dataset's dimensionality can be reduced to two components while preserving most of the original information. The projection of data onto these principal components also highlights a strong relationship between displacement, number of cylinders, acceleration, and weight in predicting a car's fuel efficiency.

Considering these findings, we have a hint that the *mpg* regression may lead to noteworthy results. However, the classification of a car's *origin* may present challenges, depending on how well the features differentiate the categories.

Overall, these results indicate that this dataset is suitable for achieving the regression and classification objectives.

7 About LLM usage - Part 1

Although the LLM was used to enhance students' understanding of various concepts, including data analysis and machine learning methodologies, it was not involved in the actual writing of this document, except for paraphrasing content originally written by hand.

8 Contributions - Part 2

All participants in the assignment have compiled and/or reviewed the present document and all the source codes. Nevertheless, for organizational reasons, we decided to be responsible for different parts of the document, as shown in the following *(not exhaustive)* table:

	Flavio Sarno (s242991)	**Francesco Balducci (s250200)**	**Florencia Illanes (s247222)**
Introduction	30%	30%	40%
Regression Part A	50%	30%	20%
Regression Part B	30%	45%	25%
Classification	30%	30%	40%
Discussion	20%	20%	60%

Table 6: Table of responsibilities

9 Introduction - Part 2

The data set [1] features several engine characteristics and specifications of cars built between 1970 and 1982. The data set investigates city-cycle fuel consumption for cars built in three different geographical areas: the United States, Europe, and Japan. The available properties are: mpg, cylinders, displacement, horsepower, weight, acceleration, model year, origin, and car name.
The following table shows a summary of the dataset values:

Feature	**Description**	**Attribute**	**Type**
mpg	Miles per gallon (fuel efficiency)	Continuous	Ratio
cylinders	Number of engine cylinders	Discrete	Ordinal
displacement	Engine displacement (cubic inches)	Continuous	Ratio
horsepower	Engine power output	Continuous	Ratio
weight	Vehicle weight (lbs)	Continuous	Ratio
acceleration	Time to accelerate from 0 to 60 mph (seconds)	Continuous	Ratio
model_year	Model year of the vehicle	Discrete	Interval
origin	Country of origin (1: USA, 2: Europe, 3: Japan)	Discrete	Nominal
car_name	Name of the vehicle model	Discrete	Nominal

Table 7: Description and classification of dataset features based on attribute type. This dataset uses the imperial system of units [1].

Generally speaking, the main objective of this analysis is to predict the fuel efficiency of a car, measured in miles per gallon *(mpg)*, with different linear regression methods (first project), and to classify existing and unknown observations around the *origin* feature through the remaining noncategorical properties.

10 Regression Analysis

The main objective of this analysis is to predict the fuel efficiency of a car, measured in miles per gallon *(mpg)*, based on various technical characteristics of the car. The selected predictive features are *number of cylinders, displacement, weight, horsepower, acceleration, origin, and model year*. In simple terms, the goal is to model fuel consumption based on those key vehicle specifications.

10.1 Part A

In this section, we implemented a linear regression model to **predict the mpg**. For this, all features were standardized to ensure that differences in magnitude between attributes would not distort the learning process. Standardization improves both model stability and interpretability by rescaling the data to have zero mean and unit variance. Since all input features were numerical, no one-of-K encoding was required.

To train the model, we used linear regression with L_2-regularization (Ridge regression) and performed 10-fold cross-validation over a range of λ values between 10^{-3} and 10 with 50 logarithmically-spaced steps to estimate the error (mean squared error, MSE). The optimal value was found to be:

$$\lambda^* = 1.265$$

Figure 9 shows the validation error across all λ values evaluated during cross-validation.

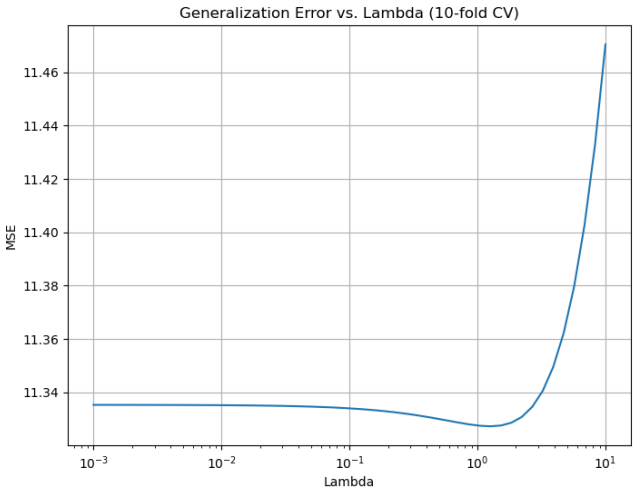

Figure 9: Mean Squared Error across different λ values during cross-validation.

This value of λ, i.e., the regularization parameter (or strength used to attenuate high magnitude weights and therefore used to lower the test error), was chosen because it minimized the validation MSE, achieving a good trade-off between underfitting and overfitting. Higher values of λ tend to oversimplify the model, leading to increased error (see Figure 9).

Based on the optimal λ, the resulting linear regression model is:

$$\begin{aligned}\text{MPG} = 23.45 &- 0.77 \cdot \text{cylinders} + 1.78 \cdot \text{displacement} \\ &- 0.69 \cdot \text{horsepower} - 5.26 \cdot \text{weight} + 0.18 \cdot \text{acceleration} \\ &+ 2.74 \cdot \text{model year} + 1.13 \cdot \text{origin}\end{aligned}$$

The regression coefficients provide insights into how each feature influences fuel efficiency (MPG). Specifically, we have:

- **Bias term (23.45)**: Expected MPG when all features are at their mean values.
- **Weight (-5.26)**: Strongest negative effect. Heavier vehicles reduce fuel efficiency by 5.26 MPG (per standard deviation increase).
- **Displacement (+1.78)**: Unexpected positive effect, likely due to multicollinearity with *cylinders* and *horsepower*.
- **Model year (+2.74)**: Newer vehicles are more efficient, reflecting technological advancements.
- **Origin (+1.13)**: Non-U.S. vehicles (Europe/Japan) are more efficient than U.S. models.

10.2 Part B

In this section, we compare the performance of three models: the regularized linear regression (RLR) model from Part A, an artificial neural network (ANN), and a simple baseline model. To ensure a fair comparison, we employed a two-level cross-validation scheme with $K_1 = K_2 = 10$ folds and tested the models on the same folds. The inner loop was used to identify the optimal hyperparameters for each model, while the outer loop was used to evaluate their generalization performance.

Artificial Neural Network (ANN). The ANN model's complexity was controlled by varying the number of hidden units in a single hidden layer, with $h \in \{1, 7, 13, 19, 25\}$. The optimal number of units selected via inner cross-validation was:

$$h^* = 25$$

This configuration yielded the lowest average generalization error, with a Mean Squared Error (MSE) of:

$$\text{MSE}_{\text{ANN}} = 6.75$$

Regularized Linear Regression (RLR). For the RLR model, model complexity was controlled via the regularization parameter λ, tested in the range $[10^{-3}, 10]$. The optimal value selected through inner cross-validation was:

$$\lambda^* = 5.67$$

With this parameter, the model achieved a generalization error of:

$$\text{MSE}_{\text{RLR}} = 8.27$$

Baseline Model. As a reference, we also evaluated a baseline model that simply predicts the mean of the target variable (MPG) from the training set. This naive strategy resulted in a considerably higher error:

$$\text{MSE}_{\text{Baseline}} = 46.71$$

Table 8 summarizes the optimal parameters and test errors for each model across the 10 outer folds. It is important to note that the optimal λ^* selected here differs from that found in Part A, as it was optimized separately within each outer training fold to minimize validation error, rather than being selected globally.

Table 8: Optimal hyperparameters and test errors per fold for each model.

Fold	Best λ	Best h	Test Error (RLR)	Test Error (ANN)	Test Error (Baseline)
1	5.6899	25	18.27	16.18	58.20
2	10.0000	25	12.21	10.65	39.53
3	1.5264	25	12.62	9.85	59.04
4	5.6899	25	18.42	14.23	54.72
5	6.8665	25	25.93	23.34	71.66
6	5.6899	25	8.27	6.75	46.71
7	10.0000	25	21.99	12.80	57.14
8	10.0000	25	21.48	18.08	71.20
9	8.2864	25	20.14	20.57	57.76
10	6.8665	25	26.31	27.96	86.78

Training and Test Performance. To compare the models, we examined both training and test errors:

- **Baseline (no features):** Training error = 60.67, Test error = 59.29
- **Linear regression without feature selection (no regularization):** Training error = 17.72, Test error = 20.28
- **Regularized linear regression:** Training error = 17.75, Test error = 20.25
- **Artificial neural network:** Training error = 14.75, Test error = 17.14

These results again highlight that both linear regression models substantially outperform the baseline. However, the ANN achieves the lowest training and test errors, making it the most accurate model overall.

Learning curves and test errors per fold. Figure 10 shows the learning curves for each fold in the ANN and the test mean squared error (MSE) per fold.

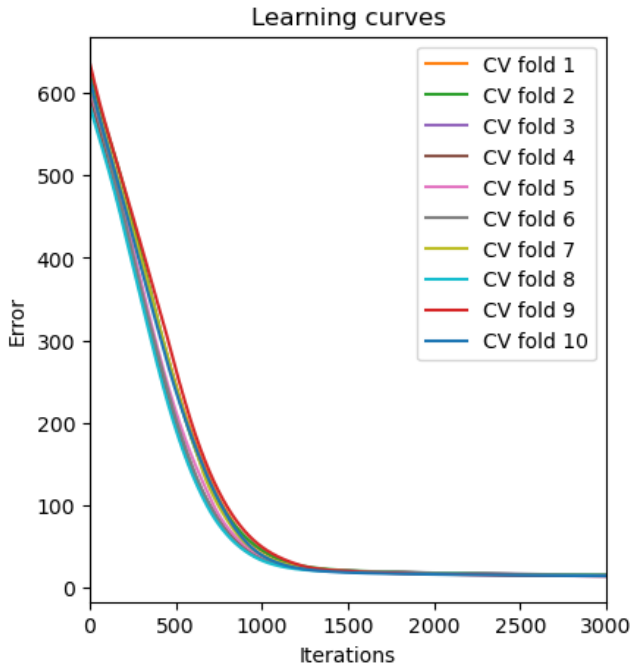

Figure 10: Learning curves for the ANN model across cross-validation folds.

As shown, the ANN converges smoothly in all folds, with a consistent decrease in error as training progresses.

The results indicate that both the RLR and ANN models significantly outperform the baseline. Moreover, the ANN achieves the lowest average MSE, suggesting it is the most effective of the three models.

To determine whether the observed differences in performance are statistically significant, we conducted a series of **paired t-tests**. The following confidence intervals and p-values summarize the comparisons (significance level $\alpha = 0.05$:

- **Baseline:** Confidence interval for MSE = [52.84, 67.60] Indicates the expected error of a naive prediction model. As anticipated, the error is high.
- **Best ANN:** Confidence interval = [12.81, 19.25] Demonstrates significantly lower error, confirming the ANN's superior performance.
- **Best RLR:** Confidence interval = [15.33, 21.77] Also significantly outperforms the baseline, though not as effectively as the ANN.
- **ANN vs. Baseline:** Confidence interval = [-50.42, -37.95], p-value = 4.02×10^{-36} Strong evidence that the ANN is significantly better than the baseline.
- **RLR vs. Baseline:** Confidence interval = [-47.60, -35.74] , p-value = 1.28×10^{-35} Indicates a statistically significant improvement of RLR over the baseline.
- **ANN vs. RLR:** Confidence interval = [-3.95, -1.09] , p-value = 0.06×10^{-2} Suggests that the ANN performs slightly better than RLR, although the difference is not statistically significant at the 5% level.

Overall, the ANN emerged as the best-performing model in terms of generalization error. Both the ANN and RLR models demonstrated statistically significant improvements over the baseline, reinforcing the value of using machine learning models for this predictive task.

11 Classification Analysis

The objective of this classification task was to predict the *origin* of a vehicle—United States, Europe, or Japan—based on its non-categorical attributes: *mpg, cylinders, displacement, horsepower, weight, acceleration*, and *model year*. This is a *multi-class classification problem* with three possible classes, and we compared three approaches:

- **Multinomial Logistic Regression** (LogReg) with L_2 regularization, where $\lambda \in [0.001, 10]$
- **Artificial Neural Network** *(as "method 2")* with $h \in \{1, 7, 13, 19, 25\}$ hidden units
- **Baseline Model**, which always predicts the majority class in the training set

We implemented a two-level cross-validation procedure with $K_1 = K_2 = 10$ folds to ensure robust model evaluation. The inner folds were used for hyperparameter selection, while the outer folds measured generalization performance. All features were standardized within each fold using training data statistics only.

The hyperparameters act as the control knobs of each model: for Logistic Regression, we tuned the regularization strength λ; for ANN, we selected the optimal number of hidden units h. The results across outer folds—including the best hyperparameters and test error rates—are summarized in Table 9.

Fold	h^* (ANN)	Test Error (ANN)	λ^* (LogReg)	Test Error (LogReg)	Test Error (Baseline)
1	25	0.125	0.193	0.175	0.375
2	25	0.125	0.281	0.125	0.275
3	13	0.256	0.110	0.205	0.385
4	19	0.205	0.001	0.308	0.462
5	13	0.256	0.409	0.282	0.333
6	19	0.179	0.494	0.231	0.385
7	25	0.333	0.233	0.256	0.410
8	25	0.256	0.193	0.231	0.333
9	25	0.128	0.001	0.231	0.359
10	13	0.256	0.024	0.231	0.436

Table 9: 10x10 Cross-Validation Results

To determine whether the performance differences among the models were statistically significant, we conducted **pairwise comparisons using McNemar's test** (Setup I) on the outer-fold predictions. The statistical results are summarized in Table 10.

Comparison	Error Difference	95% CI	p-value
ANN vs. Baseline	-0.163	[-0.214, -0.112]	1.32×10^{-9}
LogReg vs. Baseline	-0.148	[-0.200, -0.096]	6.47×10^{-8}
ANN vs. LogReg	-0.015	[-0.054, 0.023]	0.59

Table 10: Statistical Comparison of Classification Models (McNemar's Test)

The results show that both ANN and Logistic Regression significantly outperform the baseline model. While ANN achieves slightly lower test error than LogReg, this difference is not statistically significant ($p = 0.59$).

Finally, we trained a **logistic regression model** on the full dataset using the optimal regularization value $\lambda = 0.596$, selected from Table 9. The model was trained using `L2` regularization with the `lbfgs` solver. The resulting weights—shown in Table 11—indicate the influence of each feature in classifying a sample into each of the three origin categories.

As illustrated in Figure 11, the model relies heavily on features such as *displacement*, *weight*, and *horsepower* to differentiate between regions. This is consistent with domain knowledge—U.S. cars typically have

Feature	United States (0)	Europe (1)	Japan (2)
mpg	-0.752	0.494	0.258
cylinders	-1.583	0.803	0.779
displacement	8.800	-4.885	-3.916
horsepower	-0.875	-1.576	2.451
weight	-2.823	3.975	-1.153
acceleration	0.212	-0.401	0.189
model_year	0.573	-0.845	0.273

Table 11: Logistic Regression Weights for Each Region Class

larger engines and are heavier, whereas Japanese cars are often lighter with higher horsepower relative to their size. Weight is also a distinguishing feature, as it varies significantly across manufacturing regions.

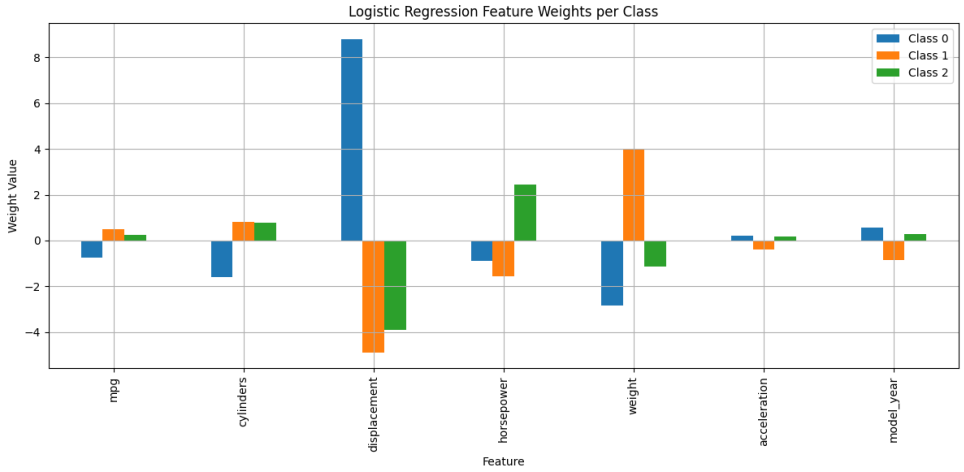

Figure 11: Feature Weights for Each Class (Car Origin)

12 Discussion - Part 2

12.1 Regression Analysis

The regression analysis revealed how various physical characteristics of vehicles influence fuel efficiency, measured in miles per gallon (MPG). The results demonstrated both negative and positive impacts of certain features.
For the negative impact we have:

- **Weight** (-5.26): As expected, heavier vehicles tend to consume more fuel.
- **Horsepower** (-0.69): More powerful vehicles also reduce MPG, aligning with common expectations.
- **Cylinders** (-0.77): Vehicles with more cylinders, typically associated with larger engines, similarly reduce fuel efficiency.

For the negative impact we have:

- **Model Year** (+2.74): The model year was found to have a positive effect on fuel efficiency. This likely reflects advancements in fuel-efficient technologies over time.
- **Displacement** (+1.78): An unexpected positive effect was observed for displacement, which might be influenced by multicollinearity with other engine-related features, such as cylinders and horsepower.

These factors confirm the intuition of how the technical characteristics of the car can have an influence on the fuel efficiency.

For the technical insights in the regression analysis, we employed two-level nested cross-validation with $K_1 = K_2 = 10$ to identify optimal regularization strengths (λ) and the number of hidden units (h) in the Artificial Neural Network (ANN). This approach ensured robustness against overfitting while balancing model complexity. Both Regularized Linear Regression (RLR) and ANN significantly outperformed the baseline model, which predicts the mean value.

Statistical tests confirmed that the ANN outperformed RLR with a p-value of 0.018, while both models were significantly better than the baseline with $p < 0.001$.

12.2 Classification Analysis

A three-class classification problem was established to predict the *origin* of a vehicle (United States, Europe, or Japan) based on its physical and performance features. Three approaches were compared:

- **Baseline**: Predicts the majority class (USA) for all samples
- **Multinomial Logistic Regression**: With L2 regularization ($\lambda \in [0.001, 10]$)
- **ANN**: Single hidden layer with $h \in [1, 7, 13, 19, 25]$

The average test error rates across the folds showed that both learning models significantly outperformed the baseline:

Moreover, McNemar's test showed that ANN and LogReg are better models than baseline ($p < 0.01$), and actually the difference between them was not statistically significant ($p = 0.59$). This suggests that while both learning models are effective for this classification task, their performance is relatively similar.

Finally, we trained a final logistic regression model on the full dataset using the optimal λ, and examined the learned weights. Features such as *weight*, *displacement*, and *horsepower* played a major role in distinguishing the origin of the vehicles, consistent with expectations based on regional manufacturing characteristics. For this, the linear regression obtained was completely different from the one to predict the mpg, since they have different approaches.

In general, both regression and classification tasks highlighted the effectiveness of regularized models and neural networks when used thoughtfully with cross-validation. Although artificial neural networks (ANNs) generally delivered slightly better results, logistic regression remained a strong competitor due to its interpretability, particularly in understanding feature importance. Moreover, the significance of features varied across tasks, with the importance patterns differing from those observed in the MPG prediction task, demonstrating that feature relevance can change based on the specific prediction goal.

12.3 Previous Studies

Finally, in Quinlan's seminal 1993 paper, *"Combining Instance-Based and Model-Based Learning"* [1], the Auto MPG dataset was utilized for regression analysis. Specifically, Quinlan employed both *Linear Regression* and *Artificial Neural Networks (ANN)*, with a 10-fold cross-validation method. The results showed that the average error for the linear regression model was 2.61, with a relative error of 19.4%, while the ANN yielded an average error of 2.02 and a relative error of 12.5%.

This analysis demonstrates that the ANN outperforms the linear regression model, although it's important to note that Quinlan's evaluation metrics differ.

13 About LLM usage - Part 2

Although the LLM was used to enhance students' understanding of various concepts, including data analysis and machine learning methodologies, it was not involved in the actual writing of this document, except for paraphrasing content originally written by hand.

References

[1] R. Quinlan. *Auto MPG*. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5859H. 1993.

[2] American Statistical Association. *Before 1993*. 2019. URL: https://community.amstat.org/jointscsg-section/dataexpo/dataexpobefore1993.

[3] Steven L. Salzberg. "Book Review: C4.5: Programs for Machine Learning by J. Ross Quinlan". In: *Machine Learning* 16 (1994), pp. 235–240. DOI: 10.1023/A:1022645310020.

[4] Tue Herlau, Mikkel N Schmidt, and Morten Mørup. "Introduction to machine learning and data mining". In: *Lecture notes of the course of the same name given at DTU (Technical University of Denmark)* (2023).

[5] Carnegie Mellon University. *Cars Dataset Description*. n.d. URL: http://lib.stat.cmu.edu/datasets/cars.desc.