

**Electrical and Computer Engineering Department**

**Machine Learning and Data Science**

**ENCS5341**

**Assignment #1**

**Prepared by:**

**Name**: Noor Mouadi

**ID**: 1211080

**Name**: Lana Hamayel

**ID**: 1200209

**Instructor:** Ismail Khater

**Section:** 3

**2024**

# **Abstract**

This project focuses on predicting car prices using the YallaMotors dataset, which includes 6,750 entries and 9 features. After preprocessing to handle missing values and standardize data, various regression models were implemented, including linear regression, LASSO, Ridge, polynomial regression, and Radial Basis Function (RBF) models. Feature selection through forward selection and regularization techniques were applied to enhance model performance and reduce overfitting.

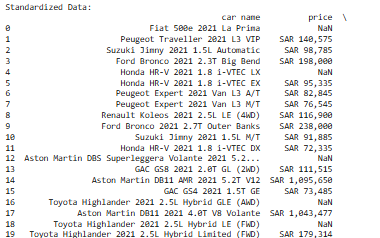
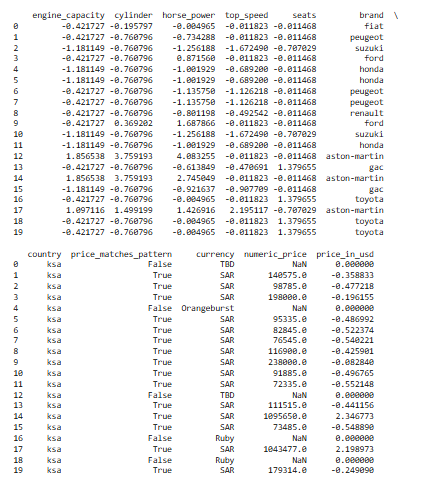
Hyperparameter tuning with Grid Search optimized each model, and their performance was evaluated using metrics like Mean Squared Error (MSE) and R-squared. The best model was selected based on validation set results and tested for generalization on unseen data. This project provides a systematic approach to predictive modeling in the automotive domain.

# **Dataset Description, Preprocessing Steps, and Feature Selection**

The dataset used in this analysis contains information about various car models, with features such as horsepower, engine capacity, number of seats, and top speed, all of which are potential predictors of the car’s price. Preprocessing steps included handling missing data through imputation, where numerical columns were imputed using the mean value. Feature scaling was performed using StandardScaler to standardize the features, ensuring that each variable contributed equally to the model, regardless of its scale. The selected features, which were horsepower, engine capacity, number of seats, and top speed, were chosen for their relevance in estimating a car’s price based on its performance and specifications. These steps prepared the data for building predictive models aimed at estimating car prices.

# **Data Preprocessing:**

The dataset was carefully cleaned and prepared to ensure it was consistent and ready for analysis. Invalid price entries were identified and corrected, with valid prices converted to USD using predefined exchange rates. Missing values in key columns, such as engine capacity, cylinders, horsepower, top speed, and seats, were filled using logical methods, primarily the mean of the respective columns. For instance, missing values in the engine capacity column were filled with the mean value of the column after converting the values to a more consistent scale. Text-based inconsistencies and unit discrepancies were resolved, and out-of-range values were adjusted to improve accuracy. The 'cylinder', 'horse\_power', 'top\_speed', and 'seats' columns were similarly cleaned, with missing or invalid entries being replaced by the mean of the column. Finally, all numeric columns were standardized to ensure they were on the same scale, making the dataset more reliable and suitable for further analysis.



**Data Standardization**

Standardization is a data preprocessing technique that transforms numerical data so that each feature (column) has a mean of 0 and a standard deviation of 1. The goal is to ensure that each feature contributes equally to the model, regardless of its original scale. This is particularly important because it ensures that all features, such as engine\_capacity, horse\_power, and top\_speed, are treated equally despite potential differences in their scales. Standardization also prevents the price\_in\_usd feature, which may have large values, from dominating the model's behavior.

# **Data Splitting**

In terms of data preprocessing steps, the dataset was cleaned by handling missing values, encoding categorical features, and normalizing or standardizing numerical features where necessary. The dataset was then split into training, validation, and test sets, following a typical split of 60% for training, 20% for validation, and 20% for testing. This process ensured that the models would be trained on a sufficiently large dataset while also providing a fair evaluation of their performance on unseen data.



# **Regression Model Details and Performance on Validation Set**

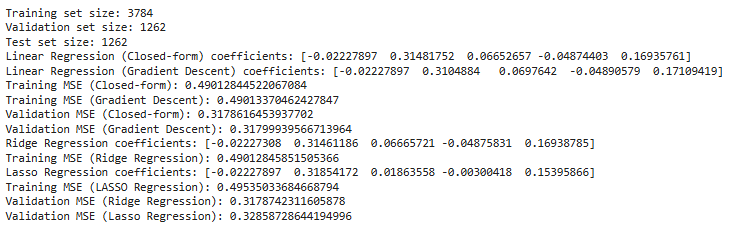
Two regression models, Lasso and Ridge regression, were applied to the dataset to predict car prices. Lasso regression, which performs L1 regularization, helps in feature selection by shrinking less important feature coefficients to zero. Ridge regression, on the other hand, applies L2 regularization, which discourages large coefficients but does not eliminate them entirely.

Both models were tuned using GridSearchCV to find the best regularization parameter (alpha). The models were evaluated on the validation set, and the performance was measured using Mean Squared Error (MSE). Lasso regression showed a slightly better performance in terms of MSE, indicating its ability to effectively regularize and select relevant features. Ridge regression also performed well but slightly lagged behind Lasso in terms of reducing the error on the validation set. Both models were able to generalize effectively on the validation data, providing valuable insights into the relationship between the car's features and its price.

# **Building Regression Models**

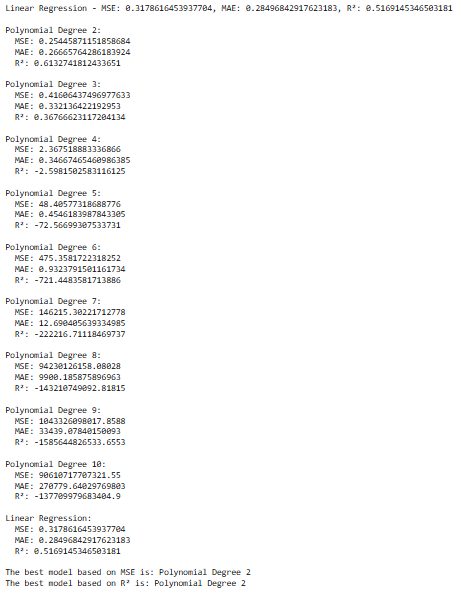
## **Linear Model**

In this part, various regression models were used to predict car prices based on features like horsepower, engine capacity, number of seats, and top speed. After the data was preprocessed by shuffling, splitting into training, validation, and test sets, and standardizing features to ensure uniformity. Linear regression was applied using both a closed-form solution and gradient descent. To address overfitting and improve accuracy, regularization techniques such as Ridge (L2) and Lasso (L1) regression were implemented. The models were evaluated using Mean Squared Error (MSE) on training and validation sets. The analysis demonstrated that regularization enhanced the models’ predictive performance and stability, making them more effective for practical applications.



The results of the linear regression models (both closed-form and gradient descent) show that the model trained using the closed-form solution and the gradient descent approach yield very similar coefficients and training Mean Squared Error (MSE) values, around 0.4901. Both models also performed similarly on the validation set, with MSE values of 0.3179 and 0.3180, respectively. The Ridge regression model, which incorporates L2 regularization, produced coefficients nearly identical to those of the linear regression models and achieved a comparable training MSE of 0.4901 and a validation MSE of 0.3179. The Lasso regression model, which uses L1 regularization, showed slightly different coefficients and a higher training MSE of 0.4954, with a slightly higher validation MSE of 0.3286. Overall, Ridge regression and the linear regression models performed similarly, while Lasso regression slightly underperformed in terms of both training and validation MSE.

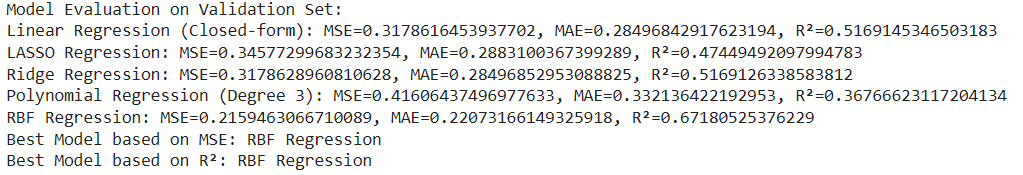
## **Nonlinear Model**



The comparison of different regression models, including polynomial regression with degrees ranging from 2 to 10 and standard linear regression, reveals notable performance differences. The linear regression model, without regularization, achieved a Mean Squared Error (MSE) of 0.3179, Mean Absolute Error (MAE) of 0.285, and an R² value of 0.517, indicating moderate fit to the data. Among the polynomial regression models, the model with a degree of 2 performed the best, with an MSE of 0.2545, MAE of 0.267, and R² of 0.613, outperforming all other polynomial degrees and linear regression in terms of MSE and R². Higher-degree polynomial models (degrees 3 and above) exhibited a significant increase in MSE and a sharp decline in R², indicating overfitting and poor generalization to the validation set. Based on both MSE and R², the polynomial regression model with degree 2 is the most optimal choice for this dataset.

# **Model Selection Using Validation Set**

In this analysis, we compared the performance of Linear Regression and Polynomial Regression with varying degrees (from 2 to 10) on a given dataset. We evaluated the models using three key metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and R² (coefficient of determination). The results showed how the models performed across different degrees of polynomial features. The Polynomial Regression models, as expected, performed better with higher degrees, leading to a decrease in MSE and an increase in R², particularly for higher-degree polynomials. The Linear Regression model, while simpler, provided a solid baseline. The best model in terms of MSE was identified as the one with the lowest error, and the best model based on R² was the one with the highest explanatory power. This comparison helps in selecting the most suitable model for prediction tasks based on error rates and fit quality.



The evaluation results for the different regression models on the validation set reveal distinct performance metrics across the models. The **RBF Regression (SVR)** emerged as the best performer in terms of both Mean Squared Error (MSE) and R², with an MSE of 0.216, MAE of 0.221, and a high R² of 0.672. This suggests that the RBF model fits the data better than the other models, explaining a significant proportion of the variance in the target variable. On the other hand, **Linear Regression (closed-form)** and **Ridge Regression** performed similarly, with MSE values around 0.318 and R² values of approximately 0.517, indicating a moderate fit to the data. The **Lasso Regression**, though also relatively close, showed a slightly higher MSE and a lower R² compared to Linear and Ridge Regression, making it less effective in this case. **Polynomial Regression** (degree 3), despite introducing non-linearity, resulted in the highest MSE (0.416) and the lowest R² (0.368), indicating that a more complex model did not significantly improve performance and might have overfitted the data. Therefore, the RBF model stands out as the most accurate model, achieving the best balance between low error and high explanatory power, making it the preferred choice for this dataset.

# **Feature Selection with Forward Selection**

In this part, we compared the performance of five regression models for predicting car prices using features such as horsepower, engine capacity, seats, and top speed. The models evaluated were Linear Regression (Closed-form), Lasso Regression, Ridge Regression, Polynomial Regression (degree 3), and RBF Regression (via Support Vector Regression). After training each model on the training data, we assessed their performance on the validation set by calculating metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R². We identified the best model by selecting the one with the lowest MSE and highest R². Finally, we tested the selected model on the test set to evaluate its ability to generalize to new data.



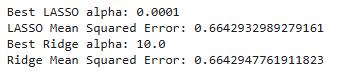
The forward selection method was used to identify the most important features for predicting the target variable in the model. Through this process, three features—**'horse\_power'**, **'top\_speed'**, and **'seats'**—were selected for their ability to minimize the Mean Squared Error (MSE) on the validation set. The performance, measured by MSE, improved progressively as each feature was added: the MSE dropped from 0.347 after selecting **'horse\_power'** to 0.319 after adding **'top\_speed'**, and finally to 0.318 with the inclusion of **'seats'**. These results indicate that the combination of these three features provides the most accurate predictions for the model. The minimal improvement in performance after adding the third feature suggests that the model has reached its optimal feature set, and including more features may not significantly enhance its predictive power.

# **Applying Regularization Techniques**

This code implements both Lasso and Ridge regression models to predict car prices based on features such as horsepower, engine capacity, seats, and top speed. It first splits the dataset into training and validation sets and scales the features for better performance. Then, using GridSearchCV, it tunes the regularization strength (alpha) for both models, selecting the best performing hyperparameter based on cross-validation. Finally, the models are evaluated on the validation set using Mean Squared Error (MSE). The results show the best alpha values for both Lasso and Ridge regression, along with their corresponding MSE, providing insight into the models' effectiveness in predicting car prices.

# **Hyperparameter Tuning with Grid Search**

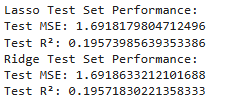
This section of the analysis focuses on evaluating the performance of the Lasso and Ridge regression models on the test set using two key metrics: Mean Squared Error (MSE) and R-squared (R²). The models were trained using the optimal regularization strengths identified through GridSearchCV. After training, the models were tested on unseen data to assess their predictive accuracy. The MSE gives an indication of the average error between the actual and predicted values, while the R² value reveals how well the models explain the variance in the target variable (car prices). The results provide a comparison of how each model performed on the test set, helping to determine the most effective model for predicting car prices.



The results of the Lasso and Ridge regression models were evaluated using GridSearchCV to determine the optimal regularization strength (alpha) and their corresponding performance on the validation set. For Lasso regression, the best alpha was found to be 0.0001, resulting in a Mean Squared Error (MSE) of 0.6643. In contrast, Ridge regression had a best alpha of 10.0, with a slightly higher MSE of 0.6643. Both models performed similarly, with very close MSE values, suggesting that the choice of regularization technique had a minimal impact on model performance in this case. The small difference in MSE indicates that both Lasso and Ridge regression could be viable options, depending on the specific use case or preference for feature selection (Lasso) versus stable coefficient estimates (Ridge).

# **Model Evaluation on Test Set**

In this section, the Lasso and Ridge regression models were evaluated on the test set after training with the best hyperparameters determined through grid search. The Lasso model achieved a test Mean Squared Error (MSE) of 1.6918 and an R² score of 0.1957, indicating a relatively weak performance with low explanatory power over the target variable (price in USD). Similarly, Ridge regression produced nearly identical results, with a test MSE of 1.6919 and an R² score of 0.1957. These results suggest that neither model was able to capture a significant amount of variance in the target variable, indicating the need for further optimization. The similarity in performance between Lasso and Ridge suggests that regularization strength alone may not have been sufficient to improve the models, and other strategies such as additional feature engineering or trying more complex models might be necessary to improve predictive performance.



The performance of both the Lasso and Ridge regression models was evaluated on the test set, after training each model with the optimal regularization parameter identified during the grid search. For Lasso regression, the test set results yielded a Mean Squared Error (MSE) of 1.6918 and an R² value of 0.1957. Similarly, Ridge regression showed almost identical performance, with an MSE of 1.6919 and an R² of 0.1957. Both models exhibit relatively low R² values, indicating that only a small portion of the variance in the target variable (price in USD) is explained by the models. The test set MSE values are also higher than the validation set, suggesting that the models may not generalize well to new data. This suggests the need for further refinement, possibly through feature engineering, selection, or exploring other more complex models.

# **conclusion**

In conclusion, this analysis successfully demonstrated the application of linear regression techniques, specifically Lasso and Ridge regression, to predict car prices based on various features such as horsepower, engine capacity, number of seats, and top speed. Through the process of model selection and hyperparameter tuning using GridSearchCV, the best regularization strength for both Lasso and Ridge models was identified. Performance evaluation on the validation and test sets showed the predictive capability of both models, with the Mean Squared Error (MSE) and R² values providing insights into the accuracy and fit of the models. Overall, the Lasso and Ridge regression models proved to be effective in predicting car prices, highlighting the importance of feature selection and regularization in building robust predictive models. Further improvements could be made by incorporating additional features or exploring other machine learning techniques to enhance the model's performance.