UNIVERSITI KEBANGSAAN MALAYSIA

The National University of Malaysia

Assignment 1

Machine Learning

STQD6024

Hazim Fitri Bin Ahmad Faudzi

P152419

Prof. Madya Dr. Mohd Aftar Bin Abu Bakar

Instruction

Choose suitable dataset from this repository:

https://archive.ics.uci.edu/

Make sure the dataset that you choose have **more than 10 variables**, and the response variable should be quantitative.

By using the **forward, backward, and forward-backward regressions**, together with **k-fold cross-validation**, identify the **best model** for that dataset.

Dataset Selection

This project aims to identify the best Linear Regression model using forward selection, backward elimination, and stepwise regression, evaluated through k-fold cross-validation. The Forest Fires dataset from the UCI Machine Learning Repository was selected for this purpose, as it includes a clear quantitative response variable, the burned area (m^2) of forest.

The dataset contains information on forest fires in Montesinho Natural Park, located in the northeast region of Portugal, collected from 2000 to 2003. It includes 10 numerical variables and 2 categorical variables, offering a mix of meteorological and spatial data relevant for modelling fire behaviour. A table summarizing each variable, its type, and description is provided as below:

Variable	Type	Description
X	Numeric	x-axis spatial coordinate within the Montesinho park
Y	Numeric	y-axis spatial coordinate within the Montesinho park
Month	Nominal	Month of the year
Day	Nominal	Day of the week
FFMC	Numeric	Fine Fuel Moisture Code (FFMC) from Canadian Forest Fire Weather Index (FWII) system.
DMC	Numeric	Duff Moisture Code (DMC) is the moisture content of loosely impacted organic layers.
DC	Numeric	Drought Code (DC) indicates long-term drought effect.
ISI	Numeric	Initial Spread Index (ISI) related to the rate of fire spread.
Temp	Numeric	Temperature in Celsius Degrees
RH	Numeric	Relative Humidity (RH)
Wind	Numeric	Wind speed in km/h
Rain	Numeric	Rainfall in mm/m2
Area	Numeric	Burned area of the forest in hectares

Method

This study uses three linear feature selection techniques which is Forward Selection, Backward Elimination, and Stepwise Regression. These techniques are used to identify the optimal subset of predictors for modelling forest fire area using a linear regression.

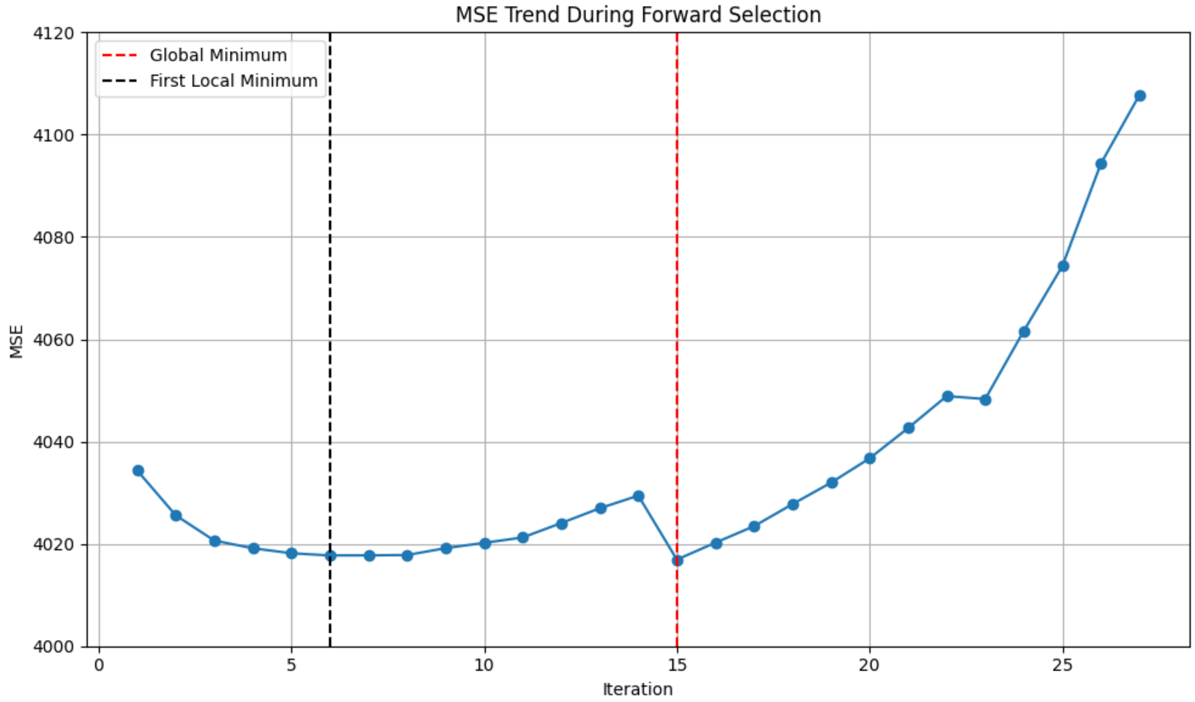
The analysis is based on the Forest Fires dataset from the UCI Machine Learning Repository. Prior to model training, data preprocessing was performed. Categorical variables such as month and day were transformed using one-hot encoding, and all numeric features were standardized to ensure uniform scaling across variables. The target variable, area, was used in its original form for model fitting.

To assess model performance reliably, k-fold cross-validation (k=10) was applied throughout the selection process. This method splits the data into 10 subsets, using nine for training and one for validation in each round, ensuring a robust estimate of the model's generalization error. A general concept of linear model selection techniques used is as follows:

- In Forward Selection, variables were added one at a time based on which provided the greatest reduction in Mean Squared Error (MSE).
- In Backward Elimination, the model started with all predictors and removed the least useful ones iteratively.
- Stepwise Regression combined both strategies, allowing the model to add or remove features at each iteration based on improvement in cross-validated MSE.

Each method was evaluated by tracking MSE across iterations to identify the point at which model performance peaked, balancing accuracy and simplicity.

Forward Selection



The above graph visualizes the performance of a linear regression model as features are added using the forward selection technique. The model's performance improves during the early stages, reaching its first local minimum at iteration 6 with an MSE of 4017.74. This is marked by the black dashed vertical line. After this point, although more features are added, the improvements in MSE become marginal or even negative, suggesting that the added complexity does not contribute meaningfully to predictive power and may introduce noise or overfitting.

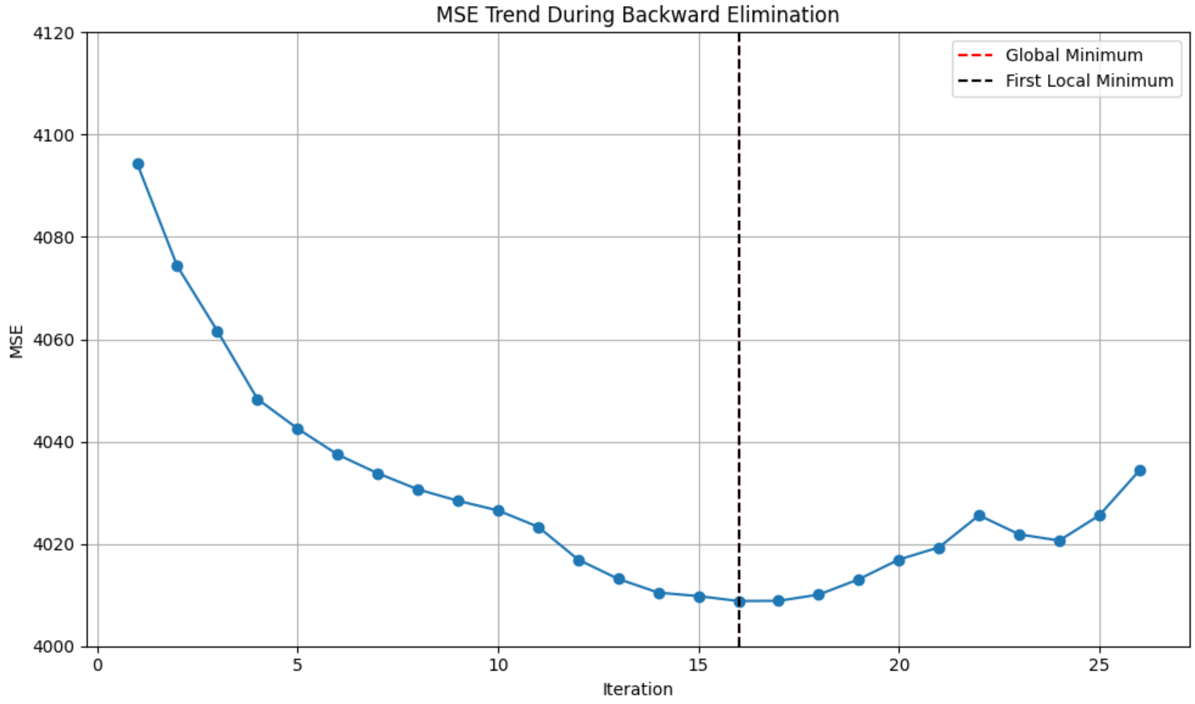
The red dashed vertical line marks the global minimum, which occurs at iteration 15 with an MSE 4016.92. While this point achieves the lowest overall MSE in the entire selection process, it comes at the cost of increased model complexity. The model at this stage includes more features, which could potentially lead to overfitting and reduced interpretability. In many real-world applications, especially when interpretability and robustness are important, stopping at the first local minimum is considered a more balanced and efficient choice.

At the first local minimum, the selected model is:

$$area = \beta_0 + \beta_1 X + \beta_2 month\_dec + \beta_3 month\_jun + \beta_4 month\_oct + \varepsilon$$

This equation indicates that the forest fire area is predicted based on the X coordinate and three specific months (December, June, and October).

Backward Selection



The above graph illustrates the progress of the model's Mean Squared Error (MSE) as features are gradually removed from a full model using backward selection. The trend clearly demonstrates a rapid drop in MSE during the early stages, indicating that the initial removals helped reduce model complexity and eliminate noise without compromising performance. The curve reaches its global minimum at iteration 16, where the MSE is 4008.82. This is denoted by the black dashed vertical line.

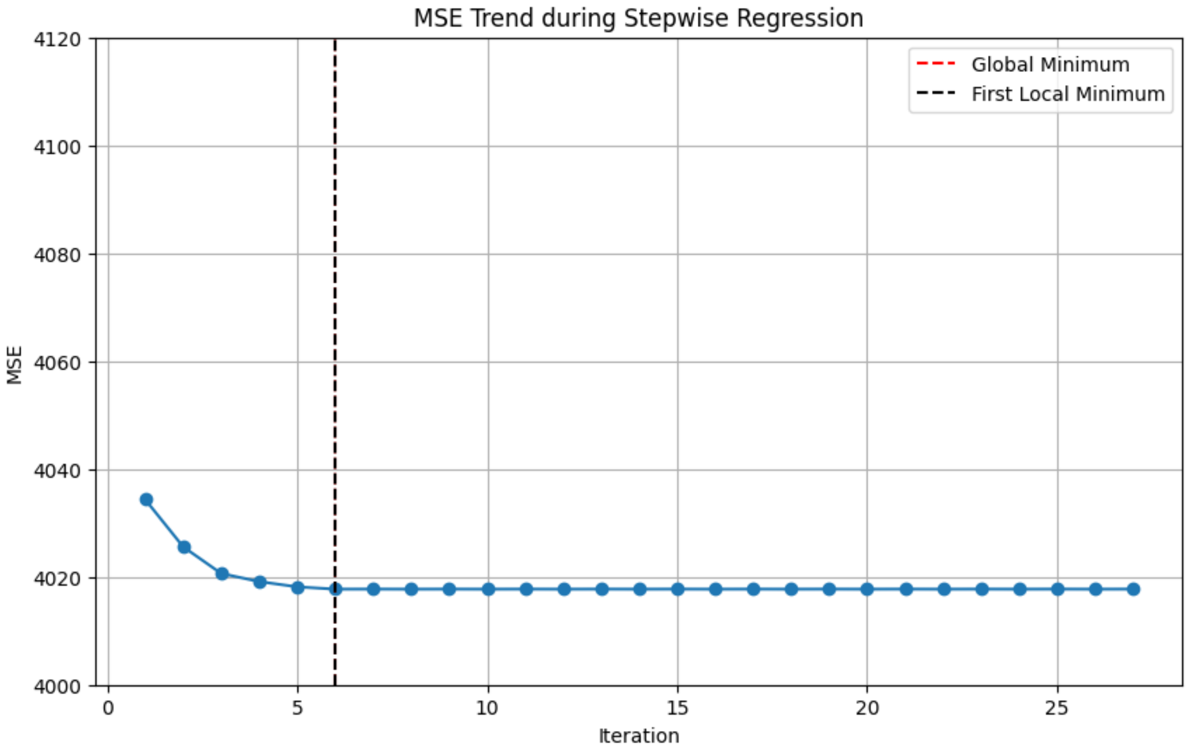
Interestingly, the first local minimum also occurs at the same iteration, suggesting that no better-performing model appears later in the process. This reinforces the validity of the selected model as the most effective balance between simplicity and performance.

The final selected model at this point is:

$$area = \beta_0 + \beta_1\, X + \ \beta_2\, DMC + \ \beta_3\, temp + \ \beta_4 month\_dec + \beta_4 month\_dec \\ + \beta_5 month\_jun + \beta_6 month\_mar + \beta_7 month\_oct + \beta_8 month\_sep + \varepsilon$$

This model includes the X coordinate, DMC (Duff Moisture Code), temperature, and five specific months. The inclusion of multiple month indicators suggests that seasonal patterns play a significant role in predicting forest fire area, likely due to varying weather and fuel conditions. The DMC and temp variables, both related to fire behaviour, strengthen the model’s connection to real-world fire dynamics. Although the model is more complex than the one selected via forward selection, its slightly lower MSE suggests that the additional features offer marginal predictive benefit.

Forward-Backward Selection



The above graph illustrates the performance of a linear regression model as features are added and optionally removed in a stepwise selection process. The curve begins with a noticeable decline in MSE during the early iterations, showing that the model improves as meaningful features are added. The model reaches its first local minimum at iteration 6, where the MSE drops to 4017.74, this point is highlighted with a black dashed vertical line.

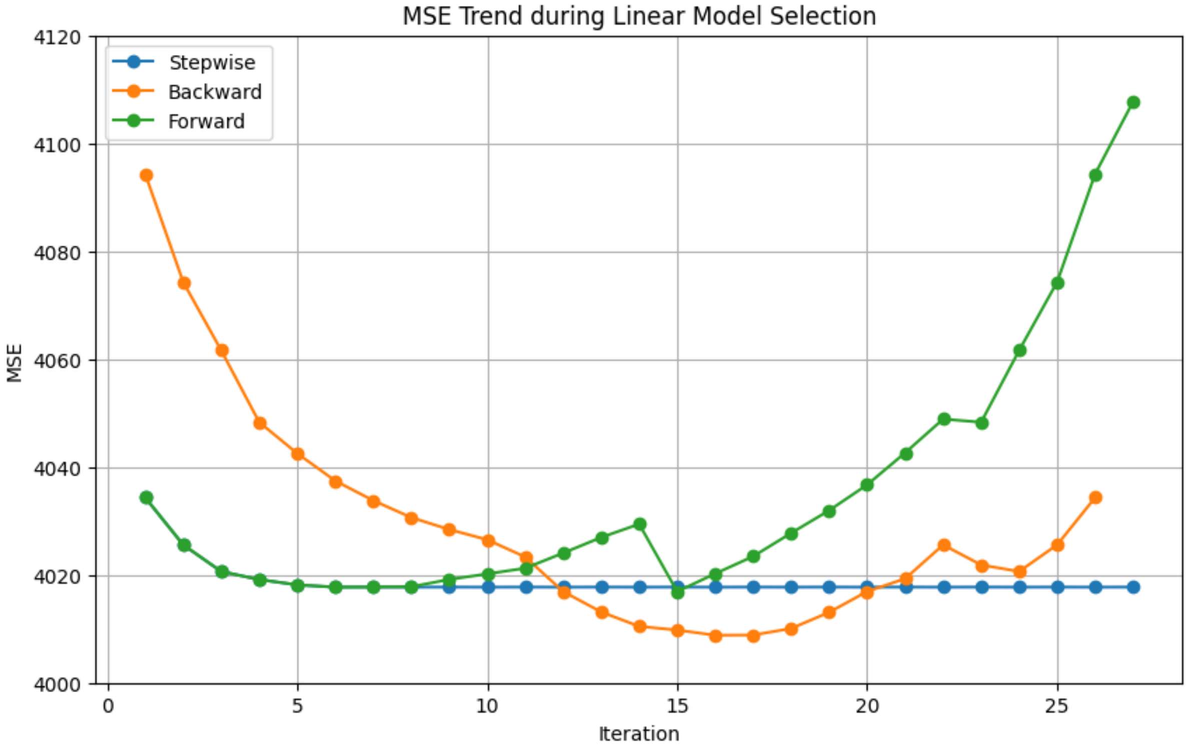
Beyond this iteration, although the algorithm continues to make structural adjustments, the MSE plateaus and no further significant gains are observed. This indicates that subsequent feature changes neither improve nor worsen model performance in a meaningful way. The global minimum, marked with a red dashed line, coincides with the local minimum, reinforcing this iteration as the most optimal model configuration in terms of predictive accuracy.

The model selected at this stage is:

$$area = \beta_0 + \beta_1\, temp + \beta_2\, X + \beta_3 month\_dec + \beta_4 month\_jun + \beta_5 month\_oct \\ + \beta_6 month\_feb + \beta_7 month\_nov + \varepsilon$$

This model integrates a mix of continuous variables (temp, X) and seasonal indicators, capturing the influence of specific months such as December, June, October, February, and November. The result reflects the seasonal nature of forest fire activity, with temperature and spatial location (X) playing central roles in fire area prediction.

Comparing the three selections



The above chart compares the performance of three feature selection techniques based on their cross-validated MSE over each iteration.

- Forward Selection (green) starts with no features and gradually adds them. It initially improves performance but leads to overfitting as more features are added, shown by the rising MSE after iteration 15.
- Backward Elimination (orange) begins with all features and removes the least useful. It steadily reduces MSE until a global minimum around iteration 16, after which performance degrades slightly.
- Stepwise Regression (blue) balances both approaches and achieves the most stable performance, reaching a low MSE early and maintaining it across iterations.

Overall, stepwise regression yields the most consistent and low-error model, while forward and backward methods show more fluctuation and sensitivity to model complexity.

Final model

The final model selected through backward elimination is:
$area = \beta_0 + \beta_1\, X + \ \beta_2\, DMC + \ \beta_3\, temp + \ \beta_4 month\_dec + \beta_4 month\_dec + \beta_5 month\_jun + \beta_6 month\_mar + \beta_7 month\_oct + \beta_8 month\_sep + \varepsilon.$

This model was chosen based on its ability to minimize the cross-validated MSE where the MSE for this model is only 4008.82 while maintaining interpretability and avoiding overfitting. It consists of both continuous and categorical predictors that are logically tied to forest fire.

- The variable X represents the east-west spatial location within the park. Its inclusion suggests that fires tend to affect different areas of the park differently, possibly due to variation in vegetation, human activity, or accessibility for suppression efforts.
- DMC (Duff Moisture Code) is a well-known fire weather index representing mid-layer forest fuel dryness. A higher DMC indicates drier conditions, which increases the likelihood and spread of fire.
- temp (temperature) positively influences fire area, which aligns with the understanding that fires spread more rapidly in hot, dry conditions. Higher temperatures lower the moisture content in vegetation, making ignition and spread easier.
- The inclusion of specific months (December, June, March, October, September) highlights the seasonal effects on forest fire occurrence and intensity. These months may coincide with regional dry seasons, wind patterns, or temperature peaks that contribute to higher fire risk.

Although wind speed was initially considered, it was not retained in the final model likely due to its effect being already captured indirectly through other variables like temperature or seasonal indicators.

In conclusion, the selected model effectively captures the key spatial, meteorological, and seasonal factors influencing forest fire area. It balances simplicity with predictive power and can serve as a practical tool for understanding fire behaviour and informing fire management strategies in similar forested environments.