INSTRUCTIONS

Assignment 1 – STQD6024 Machine Learning

Ahmad Hathim bin Ahmad Azman

Matrix No: P153146

Instructions:

1. Choose a suitable dataset from this repository: https://archive.ics.uci.edu/
2. Make sure the dataset that you choose have more than 10 variables, and the response variable should be quantitative.
3. By using the forward, backward, and forward-backward regressions, together with k-fold cross-validation, identify the best model for that dataset.

TABLE OF CONTENTS

Page

INSTRUCTIONS **1**

TABLE OF CONTENTS **2**

LIST OF TABLES **4**

LIST OF ILLUSTRATIONS **5**

LIST OF ABBREVIATIONS **6**

CHAPTER I INTRODUCTION

1.1 Dataset Selection 1

1.2 Variables 1

1.3 Assignment Objective 6

CHAPTER II METHODOLOGY

2.1 Libraries 7

2.1.1 Data Manipulation 7
2.1.2 Data Visualisation 7
2.1.3 Preprocessing and Modelling 7

2.2 Data Preprocessing 8

2.3 Data Modelling 8

2.3.1 Forward Selection 8
2.3.2 Backwards Elimination 8
2.3.3 Stepwise Regression 9

2.4 Evaluation Metrics 9

2.4.1 K-fold Cross Validation 9
2.4.2 Negative Mean Squared Error 10

CHAPTER III RESULTS AND DISCUSSION

3.1 Model Performance 11

3.1.1 Forward Selection 11
3.1.2 Backwards Elimintaion 12
3.1.3 Stepwise Regression 13

3.2 Comparison Between Models 14

3.3 Selecting Best Model 15

3.4 Evaluating Best Model 16

3.4.1 Features Selected 16
3.4.2 Final Model 17

CHAPTER IV CONCLUSION AND FUTURE WORKS

REFERENCES 20

LIST OF TABLES

Table No.		**Page**
Table 1	Table of Available Variables in Dataset	2
Table 2	Comparison of Model Performance	15
Table 3	Table of Selected Features	16
Table 4	Model Coefficients	17

LIST OF ILLUSTRATIONS

Figure No.		Page
Figure 1	Graph of Forward Feature Selection with Cross Validation	11
Figure 2	Graph of Backwards Feature Selection with Cross Validation	12
Figure 3	Graph of Stepwise Feature Selection with Cross Validation	13
Figure 4	Comparison of Feature Selection with Cross Validation	14

LIST OF ABBREVIATIONS

CV	Cross Validation
NMSE	Negative Mean Squared Error
RSS	Residual Sum of Squares
AIC	Akaike Information Criterion

CHAPTER I

INTRODUCTION

1.1 DATASET SELECTION

The Student Performance dataset from the UCI Machine Learning Repository was used for this assignment (Cortez 2008)The repository's dataset consists of two related datasets: one for student performance in mathematics and the other for student performance in Portuguese. The mathematics dataset contains 396 rows of data, while the Portuguese dataset has 649 data. The Portuguese dataset is selected for this project as it is a larger dataset that is more suitable for the project's objective.

1.2 VARIABLES

Variables in this dataset are clearly described in the source of the dataset. The target variable for this dataset is the G3 variable, which is the final grade of the students. The included variables as part of this study are as follows in Table 1.

Table 1 Table of Available Variables in Dataset

Variable	Role	Type	Demographic	Description
school	Feature	Categorical		student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
sex	Feature	Binary	Sex	student's sex (binary: 'F' - female or 'M' - male)
age	Feature	Integer	Age	student's age (numeric: from 15 to 22)
address	Feature	Categorical		student's home address type (binary: 'U' - urban or 'R' - rural)
famsize	Feature	Categorical	Other	family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Pstatus	Feature	Categorical	Other	parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
Medu	Feature	Integer	Education Level	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)

… to be continued

… continuation

Fedu	Feature	Integer	Education Level	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Mjob	Feature	Categorical	Occupation	mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
Fjob	Feature	Categorical	Occupation	father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
reason	Feature	Categorical		reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
guardian	Feature	Categorical		student's guardian (nominal: 'mother', 'father' or 'other')
traveltime	Feature	Integer		home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

… to be continued

… continuation

studytime	Feature	Integer	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	Feature	Integer	number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup	Feature	Binary	extra educational support (binary: yes or no)
famsup	Feature	Binary	family educational support (binary: yes or no)
paid	Feature	Binary	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities	Feature	Binary	extra-curricular activities (binary: yes or no)
nursery	Feature	Binary	attended nursery school (binary: yes or no)
higher	Feature	Binary	wants to take higher education (binary: yes or no)
internet	Feature	Binary	Internet access at home (binary: yes or no)

… to be continued

… continuation

romantic	Feature	Binary	with a romantic relationship (binary: yes or no)
famrel	Feature	Integer	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	Feature	Integer	free time after school (numeric: from 1 - very low to 5 - very high)
goout	Feature	Integer	going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	Feature	Integer	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	Feature	Integer	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	Feature	Integer	current health status (numeric: from 1 - very bad to 5 - very good)
absences	Feature	Integer	number of school absences (numeric: from 0 to 93)
G3	Target	Integer	final grade (numeric: from 0 to 20, output target)

1.3 ASSIGNMENT OBJECTIVE

The primary objective of this assignment is to identify the best linear regression model for this dataset using various feature selection methods and the best model to predict student performance in Portuguese. Multiple techniques, including forward selection, backwards elimination, and stepwise regression, are applied in this dataset to identify the best subset of features. These methods are coupled with K-fold cross-validation to ensure the consistency and robustness of results.

CHAPTER II

METHODOLOGY

2.1 LIBRARIES

This project utilises multiple open-source libraries, mainly for data manipulation, data visualisation, and data modelling.

2.1.1 Data Manipulation

Two main libraries were used to perform data manipulation: Pandas and NumPy. Pandas is a powerful Python library for manipulating and analysing structured data, such as DataFrames (Mckinney 2010). NumPy, on the other hand, is a Python library that supports large, multi-dimensional arrays and matrices and allows for efficient operation of mathematical functions on these arrays (Harris et al. 2020).

2.1.2 Data Visualisation

The main library used for data visualization is matplotlib. It is a comprehensive 2D graphics package for creating publication-quality images and interactive scripting in Python (Hunter 2007).

2.1.3 Preprocessing and Modelling

Data preprocessing involves cleaning, scaling, and one-hot-encoding. These are all mostly done with the help of the Scikit-learn package. Scikit-learn is a machine learning toolkit that provides simple and efficient algorithms for handling medium-scale supervised and unsupervised tasks (Pedregosa et al. 2011).

Data modelling is a tedious task if done manually. Open-source Python libraries are often pre-made to handle specific problems with much more efficient resource management, such as utilising parallel processing for large dimensional datasets. MLxtend (Machine Learning Extensions) is a Python library that offers these services, such as sequential feature selection algorithm , model evaluation and other ensemble methods (Raschka 2018).

2.2 DATA PREPROCESSING

As mentioned before, the first step to model building is data preprocessing. Once data has been uploaded into the notebook, the first step is to remove duplicate values, address missing values, and assign the correct data type to each corresponding feature. Once the data has been thoroughly cleaned, it is split between numerical and categorical features. The numerical data is scaled with a standard scaler, and the categorical data is one-hot-encoded to ensure cohesive data.

2.3 DATA MODELLING

Aligning with the objective of this project, three methods of feature selection will be used to identify the best regression model.

2.3.1 Forward Selection

Forward selection is a variable selection method in regression analysis that begins with an empty model and iteratively adds features, selecting only those that most significantly improve the model fit according to a set criterion, as done by Hamaker (1962), where a variable which produces the highest reduction in the residual sum of squares is taken. However, Austin (2008) criticised that coefficients from regression models are biased and demonstrated that methods such as bootstrapping help mitigate this issue, and have superior confidence interval coverage.

2.3.2 Backwards Elimination

Backwards regression, also known as backwards elimination, starts with a complete model containing all possible features and sequentially removes the least contributing features based on a set of statistical criteria. Backwards regression may be computationally more expensive, but forward selection has a tendency to miss out on a pair of essential variables that are not significant marginally but are jointly significant (Rao & Rao 2012).

2.3.3 Stepwise Regression

A more general algorithm compared to backwards elimination and forward selection is to consider the possibility of both adding and deleting a variable at each step. This is also known as the forward-backwards regression or the stepwise regression. However, this model demands far more computational power than the previous two. Stepwise approach implements a stopping criterion typically based on p-values or F-test statistics (Bendel & Afifi 1977; Finos et al. 2010).

2.4 EVALUATION METRICS

In obtaining the best model for our project, multiple methods can be used to assess the model. One of the methods presented in this project is the k-fold cross-validation to make the assessment more robust, and the metric used to evaluate the model is the Negative Mean Squared Error (NMSE)

2.4.1 K-fold Cross Validation

K-fold cross-validation is a widely used resampling technique in machine learning. It partitions the dataset into k equally sized folds, iteratively using one of the folds for validation and the remaining folds for training. This minimises overfitting and provides a more reliable estimate of model generalisation by averaging results across multiple iterations. K is usually set to 5 or 10 (Uzer 2025). In this study, 5-fold cross-validation is applied. Thus, the dataset was divided into five parts for each model. To evaluate the model, the average Negative Mean Squared Error was calculated by averaging the results across all iterations, thus providing more generalisation, preventing overfitting, and improving performance reliability.

2.4.2 Negative Mean Squared Error

Negative Mean Squared Error is a performance metric used in regression analysis. It transforms the conventional mean squared error (MSE) by taking its negative, thus allowing the metric to be maximised instead of minimised. If an MSE is 5, then the NMSE would be -5. This matches the goal of cross-validation in maximising the score (Verma & Yadav 2024).

CHAPTER III

RESULTS AND DISCUSSION

3.1 MODEL PERFORMANCE

Based on the abovementioned methods, this study will further detail each model's performance based on the NMSE and observe the best possible model from these three methods. The features contributing to each model will be further discussed into the paper.

3.1.1 Forward Selection

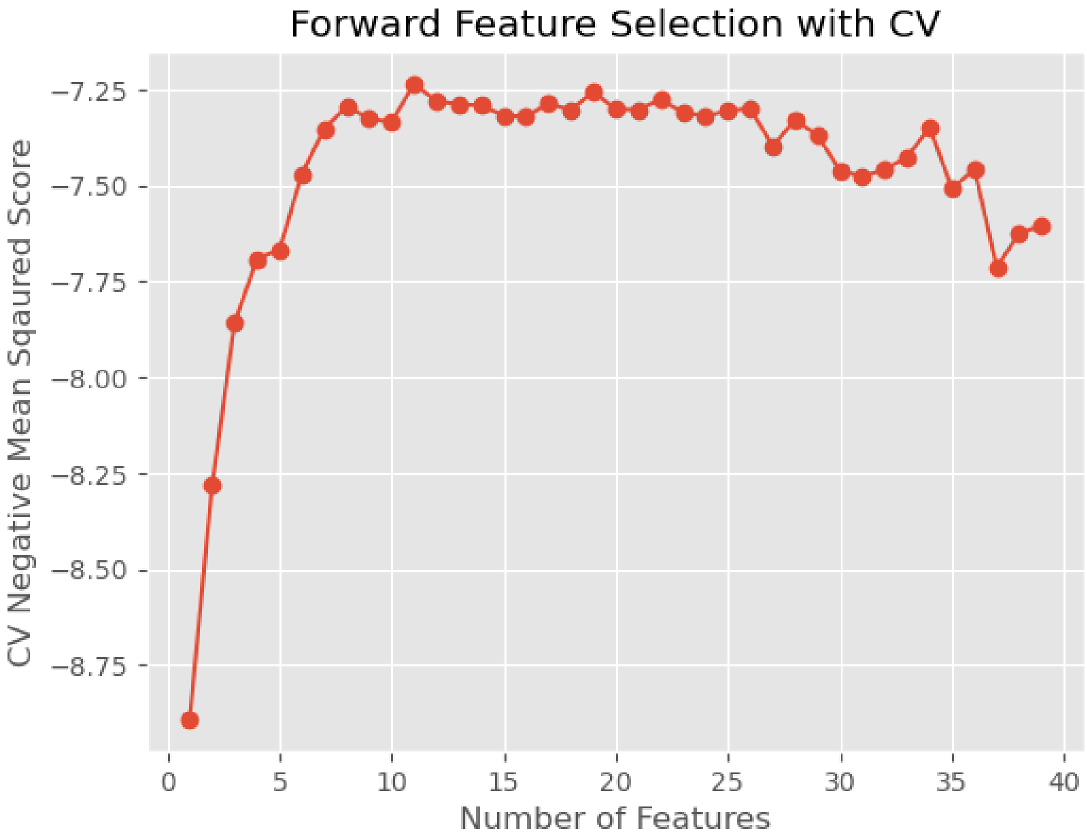

Figure 1 Graph of Forward Feature Selection with Cross Validation

Based on Figure 1, the best number of features for the forward selection regression model selection is 11, and the negative mean squared error is highest at -7.2349. The features obtained from this model are studytime, failures, Dalc, Walc, health, school_MS, sex_M, Fjob_teacher, reason_reputation, schoolsup_yes, and higher_yes.

3.1.2 Backwards Elimintaion

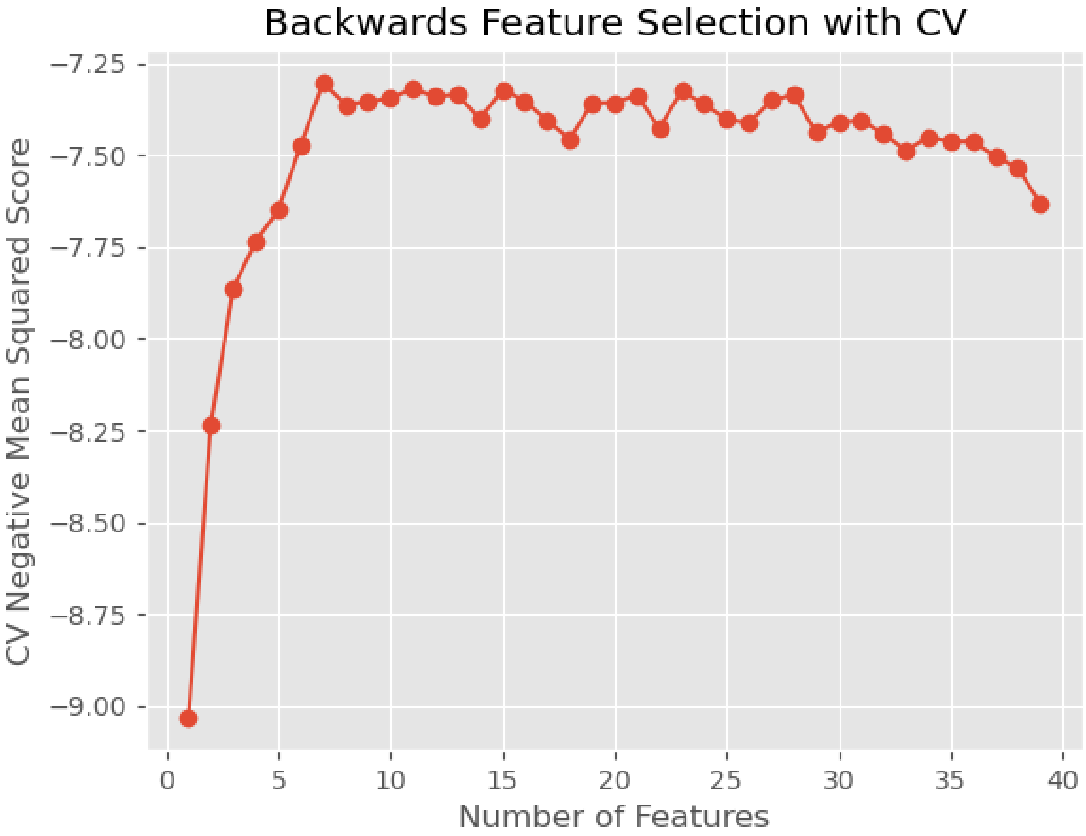

Figure 2 Graph of Backwards Feature Selection with Cross Validation

On the other hand, backwards elimination regression selected seven features for its best model, and the negative mean squared error is highest at -7.3033. The features obtained from this model are Fedu, studytime, failures, Walc, school_MS, schoolsup_yes, higher_yes.

3.1.3 Stepwise Regression

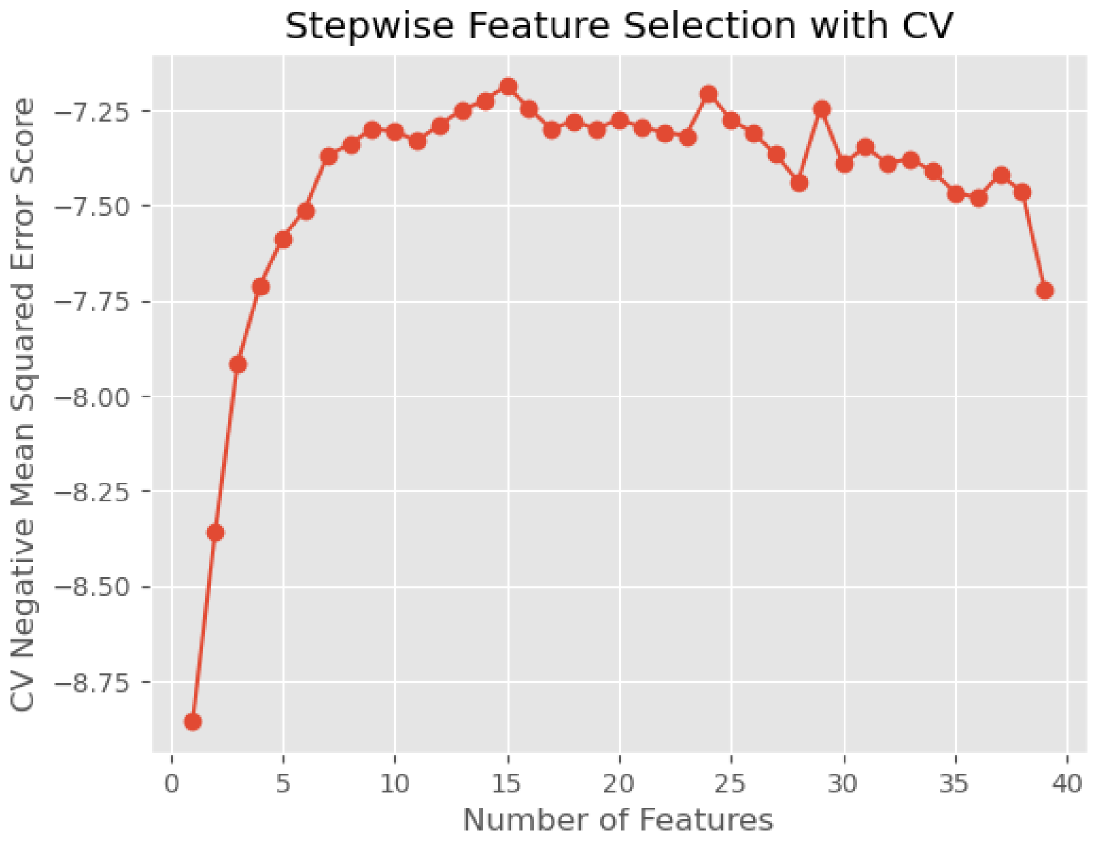

Figure 3 Graph of Stepwise Feature Selection with Cross Validation

Stepwise regression selected 15 features with a negative mean squared error highest at -7.1850 as in Figure 3. The features obtained from this model are Medu, traveltime, studytime, failures, freetime, Dalc, Walc, health, school_MS, sex_M, Fjob_teacher, schoolsup_yes, activities_yes, higher_yes, romantic_yes.

3.2 COMPARISON BETWEEN MODELS

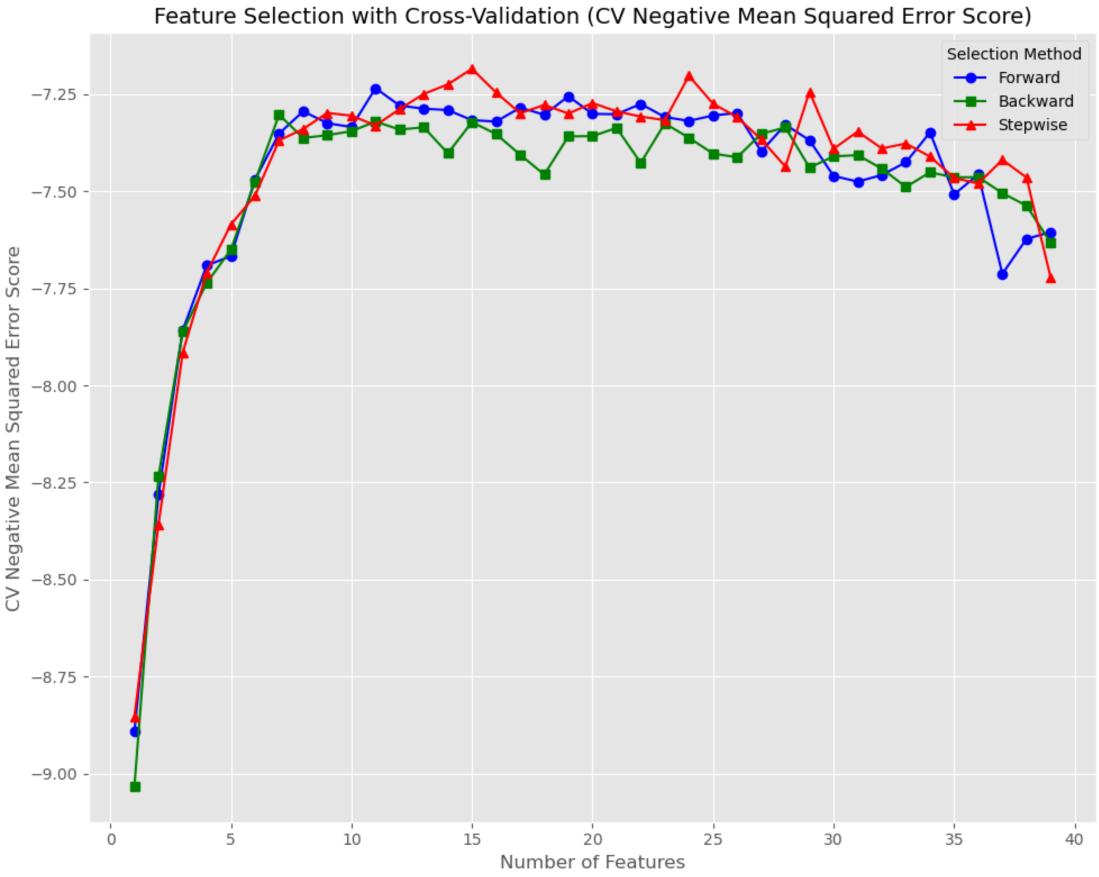

Figure 4 Comparison of Feature Selection with Cross Validation

Figure 4 shows the comparison between Forward, Backwards, and Stepwise feature selection methods based on the respective cross-validated negative mean squared error (CV-NMSE) across different numbers of selected features. All three methods show a rapid improvement in performance between 1 and 10 features. Thereafter, performance plateaus and declines slightly, which may indicate overfitting or diminishing returns from additional features. Stepwise regression shows a generally better CV-NMSE across a broader range of feature counts, peaking at 12-15 features. Forward and Backwards selection perform similarly, but Forward selection shows more stability in a higher-dimensional feature set.

3.3 SELECTING BEST MODEL

Table 2 Comparison of Model Performance

Model	Number of Features	RSS	Adjusted R^2	F-statistic	AIC Score
Forward Selection	11	3320.72	0.891	339.3	2567
Backwards Elimination	7	3216.53	0.880	447.9	2606
Stepwise Regression	15	3266.55	0.896	261.5	2551

In evaluating the best predictive model of three feature selection techniques; Forward Selection, Backwards Elimination, and Stepwise Regression, key performance metrics such as Residual Sum of Squares (RSS), Adjusted R^2, F-Statistic, and Akaike Information Criterion (AIC) is used. A summary of the evaluation can be seen as in Table 2. Stepwise Regression emerged the most optimal model, achieving the highest Adjusted ^{2} value of 0.896, indicating best explain of variance in the response variable after accounting for number of predictors. It also has the lowest AIC score of 2551. While Backward Elimination had the lowest RSS of 3216.53, and the highest F-statistic of 447.9. These came at the cost of a lower Adjusted R^2 and a higher AIC score, suggesting potential overfitting or loss of valuable predictors. Forward Selection performed moderately across all metrics but did not outperform the other approaches. Therefore, Stepwise Regression is justified as the best model for its overall balance between fit and simplicity.

3.4 EVALUATING BEST MODEL

3.4.1 Features Selected

Table 3 Table of Selected Features

X	Forward Selection	Backwards Elimination	Stepwise Regression
X_1	studytime	Fedu	Medu
X_2	Failures	studytime	traveltime
X_3	Dalc	failures	studytime
X_4	Walc	Walc	failures
X_5	health	school_MS	freetime
X_6	school_MS	schoolsup_yes	Dalc
X_7	sex_M	higher_yes	Walc
X_8	Fjob_teacher	-	health
X_9	reason_reputation	-	school_MS
X_{10}	schoolsup_yes	-	sex_M
X_{11}	higher_yes	-	Fjob_teacher
X_{12}	-	-	schoosup_ye
X_{13}	-	-	activities_yes
X_{14}	-	-	higher_yes
X_{15}	-	-	romantic_yes

Stepwise regression selected a comprehensive set of 15 features that most significantly contributed to predicting student academic performance. The chosen features span multiple domains, including academic behaviour (weekly study time and number of past class failures), demographic and institutional factors (student gender and school), and family background (mother's education level and father’s occupation as a teacher). Lifestyle and well-being indicators also played a role with weekday and weekend alcohol consumption, free time after school, and overall health status, which were also included in the modelling. Additionally, motivational factors such as students’ desire to pursue higher education and the presence of extra educational support were identified as significant predictors. Involvement in extracurricular activities and romantic relationships also appeared to influence performance. This diverse selection of features suggests that student achievement is multifactorial, influenced by academic input and personal, familial, and social contexts. Stepwise regression thus provided a balanced and interpretable model by iteratively including variables that improved predictive performance while excluding those with limited or redundant contribution.

3.4.2 Final Model

Table 4 Model Coefficients

X	Variables	Coefficients
X_1	Medu	-3.8050
X_2	traveltime	-0.3369
X_3	studytime	0.1408
X_4	failures	-0.4426
X_5	freetime	-0.1034
X_6	Dalc	-0.3400
X_7	Walc	-0.2617
X_8	health	-0.4539
X_9	school_MS	0.3903
X_{10}	sex_M	2.0521
X_{11}	Fjob_teacher	0.6690
X_{12}	schoosup_ye	-0.5815
X_{13}	activities_yes	1.2961
X_{14}	higher_yes	10.0512
X_{15}	romantic_yes	1.5649

$$\begin{aligned}\hat{y} = -x_1 3.8050 - x_2 0.3369 + x_3 0.1408 - x_4 0.4426 - x_5 0.1034 \\ - x_6 0.34 - x_7 0.2617 - x_8 0.4539 + x_9 0.3903 \\ + x_{10} 2.0521 + x_{11} 0.6690 - x_{12} 0.5815 + x_{13} 1.2961 \\ + x_{14} 10.0512 + x_{15} 1.5649\end{aligned} \quad \ldots(1.1)$$

Deducing from the selected model and features, a final model can be obtained from the list of variables and coefficients in Table 4. The final equation is as in Equation 1.1. The final stepwise regression model includes 15 variables with their corresponding coefficients, indicating the direction and strength of their impact on student academic performance. Positive coefficients suggest a direct relationship with performance, while negative coefficients indicate an inverse relationship.

Notably, the strongest positive predictors were aspiration for higher education (higher_yes, +10.0512), being in a romantic relationship (romantic_yes, +1.5649), involvement in extracurricular activities (activities_yes, +1.2961), and being male (sex_M, +2.0521). These suggest that motivation, social engagement, and gender may positively influence academic outcomes.

On the other hand, negative coefficients were observed for several factors, including school support (schoolsup_yes, –0.5815), health status (health, –0.4539), past failures (failures, –0.4426), and weekday alcohol use (Dalc, –0.34000). These imply that students needing extra academic help, those in poorer health, with prior academic failures, or who consume alcohol during the week are more likely to perform worse.

Interestingly, parental factors such as mother's education (Medu, –0.3805) and father's occupation as a teacher (Fjob_teacher, +0.6690) showed mixed effects, suggesting that while certain familial backgrounds may support performance, the effect is nuanced.

Overall, the final model highlights the multifactorial nature of academic success, encompassing academic habits, personal lifestyle, social dynamics, and family context.

CHAPTER IV

CONCLUSION

This project systematically evaluated three feature selection strategies using a combination of key statistical metrics and cross-validation techniques to identify the most robust predictive model. While each method offered unique advantages, Stepwise Regression consistently outperformed the others, achieving the highest adjusted R^2, lowest AIC, and best CV-NMSE, while preserving the highest number of features in the model. These findings highlight the importance of balancing model complexity with predictive accuracy. By carefully selecting an optimal subset of features, Stepwise Regression not only improved model generalizability but also mitigated the risk of overfitting. However, there is a risk of being computationally expensive on large-dimensional datasets. Ultimately, this reinforces the value of thoughtful feature selection as a crucial step in building efficient and interpretable machine learning models.

REFERENCES

Austin, P. C. 2008. Using the bootstrap to improve estimation and confidence intervals for regression coefficients selected using backwards variable elimination. *Statistics in medicine* 27(17): 3286-3300.

Bendel, R. B. & Afifi, A. A. 1977. Comparison of stopping rules in forward “stepwise” regression. *Journal of the American Statistical association* 72(357): 46-53.

Cortez, P. 2008. Student Performance. UCI Machine Learning Repository.

Finos, L., Brombin, C. & Salmaso, L. 2010. Adjusting stepwise p-values in generalized linear models. *Communications in Statistics—Theory and Methods* 39(10): 1832-1846.

Hamaker, H. 1962. On multiple regression analysis. *Statistica Neerlandica* 16(1): 31-56.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., Van Kerkwijk, M. H., Brett, M., Haldane, A., Del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. & Oliphant, T. E. 2020. Array programming with NumPy. *Nature* 585(7825): 357-362.

Hunter, J. D. 2007. Matplotlib: A 2D Graphics Environment. 90-95.

Mckinney, W. 2010. Data structures for statistical computing in Python. *SciPy* 445(1): 51-56.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. & Dubourg, V. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12: 2825-2830.

Rao, C. R. & Rao, T. S. 2012. *Handbook of statistics.* Elsevier Science & Technology.

Raschka, S. 2018. MLxtend: Providing machine learning and data science utilities and extensions to Python’s scientific computing stack. *Journal of open source software* 3(24): 638.

Uzer, M. S. 2025. Deep Learning-Based Classification Consisting of Pre-Trained Models and Proposed Model Using K-Fold Cross-Validation for Pistachio Species. *Applied Sciences* 15(8): 4516.

Verma, B. K. & Yadav, A. K. 2024. Advancing Software Vulnerability Scoring: A Statistical Approach with Machine Learning Techniques and GridSearchCV Parameter Tuning. *SN Computer Science* 5(5): 595.