UNIVERSITÉ PARIS 1
PANTHÉON SORBONNE

Statistical learning
vs
Machine learning

2023 - 2024
Programming in SAS

Authors :
Salma BENMOUSSA
Cecilia DONG
Flora ZHENG

Directed by Philippe De Peretti

Master 1 Econometrics, Statistics

Abstract

This paper compares the efficiency and performances of popular variable selection methods in Statistical Learning and Machine Learning. In a world where the data is increasing tremendously, it is important to keep relevant information only. Our goal is to provide a comprehensive analysis of variable selection methods in Statistical learning and Machine learning to aid in the development of relevant models. Thus, we developed four data generating processes with different characteristics in order to test each algorithm in various scenarios.

Our tests conducted to the following results : In general, Machine learning algorithms have better performance than Statistical learning algorithms, especially the Elastic Net method, which offers outstanding performance. However, these algorithms can be sensitive to changes such as the presence of multicollinearity or outliers.

Keywords: Forward / backward / Stepwise selection, Statistical learning, Machine learning, LARS / LASSO / Ridge / Elastic Net, k-fold, cross-validation

Summary

1 Introduction

The rapid evolution of technology and the increasing availability of big data have transformed our approach to statistical analysis and predictive models in econometrics. In this era of dynamic changes, two emerging concepts have gained popularity: Statistical Learning and Machine Learning. While traditional econometrics has long relied on rigorous statistical models, new methodologies of Machine learning offer innovative perspectives.

In econometrics, working with voluminous data containing numerous explanatory variables is common. However, some variables may be unnecessary for explaining the studied phenomenon or can make the information overly complex. Therefore, to establish a clear and effective description of the studied relationship, it is necessary to build a parsimonious model—a model that achieves the desired level of explanation or prediction with as few predictor variables as possible. In this paper, we will explore two variable selection methods. On one hand, Statistical Learning is an inference-based procedure that relies on the formalization of the relationship between variables in the form of mathematical equations. On the other hand, Machine Learning is an inference-free procedure that can learn from data without being explicitly programmed.

The purpose of this paper is to explore and test the effectiveness of both Statistical Learning and Machine Learning methods. In Section 2, we conduct a theoretical exploration of each algorithm, elucidating their underlying mechanisms. Section 3 focuses on data simulation using SAS, making different assumptions in each data generating process. Moving forward to Section 4, we employ the algorithms on the simulated data to discern the optimal fit for variable selection. Furthermore, in Section 5, we conduct an empirical test on diabetes data to validate the performance of the most promising algorithms identified in the previous section. This paper ends with a discussion of the overall results.

2 Models

2.1 Statistical learning

Statistical learning involves a collection of methods aimed at estimating a function f. There are two primary motivations for estimating f. The first motivation is prediction accuracy. While least squares estimates typically have low bias, may also be associated with high variance. To enhance prediction accuracy, it is sometimes beneficial to employ techniques such as shrinking or setting certain coefficients to zero. The second motivation is interpretation. When dealing with a multitude of predictors, there is often a desire to identify a more manageable subset that brings out the most pronounced effects. In the pursuit of understanding the big picture, there is a willingness to trade off some of the more subtle details.

2.1.1 Criteria for selecting variables

Several model selection criteria are available to choose the best among various options. These criteria play a key role in decision-making when evaluating models in multiple linear regression. In our study, we do not use the F-Test, likelihood Ratio Test, R^2 and R^2_{adj}.

1. **Fisher Test (F-Test)**

The F-statistic serves as the test statistic in the analysis of variance approach to assess the significance of either the entire model or its individual components.

$$F = \frac{RSS_c - RSS_{\text{nc}}}{q} \times \frac{n-(p+1)}{RSS_{\text{nc}}} \sim \mathcal{F}(q, n-(p+1))$$

where RSS_{nc} is the Residual Sum of Square of the full model, RSS_c is the Residual Sum of Square of the reduced model (only the constant), p is the number of explanatory variables, q is the number of constraints tested and n is the number of observations.

The F-test is used to determine whether adding an additional set of explanatory variables to the full model significantly improves the model fit. If the F-value is significantly large, we can conclude that the full model is statistically better than the reduced model.

2. Likelihood Ratio Test (LRT)

The likelihood ratio test statistic λ is calculated as the ratio of likelihoods under a null hypothesis (H0) and its alternative hypothesis (H1). It is expressed as:

$$\lambda = -2\log\left(\frac{L(\theta_0)}{L(\theta_1)}\right)$$

where λ is the likelihood ratio test statistic, $L(\theta_0)$ is the likelihood under the null hypothesis and $L(\theta_1)$ is the likelihood under the alternative hypothesis.

This ratio is then compared to the chi-square distribution with degrees of freedom equal to the difference in the number of parameters between two models. If the calculated ratio is significantly different, it provides evidence to reject the null hypothesis. In comparison the best model keeps being the one with the highest likelihood.

3. R^2

The R^2 criterion serves as a measure to evaluate the model's fit. When comparing two models with an equal number of explanatory variables, the obtained R^2 values are compared, and the model with the highest R^2 is chosen. Therefore, the formula is given by:

$$R^2 = 1 - \frac{SSE}{SST}$$

where SSE represents the error sum of squares and SST the total sum of squares.

However, when comparing two models with a different number of variables, the Fisher test is used. Indeed, R^2 can only be used to compare two models of the same level.

Furthermore, R^2 increases monotonically with the introduction of new variables, even if these variables are weakly correlated with the explained variable Y. Therefore, it is recommended to turn to the use of other alternatives such as adjusted R^2, AIC, AICc and BIC. These alternatives provide a more nuanced approach to variable selection, taking into account factors such as model fit and complexity.

4. R^2_{adj}

The adjusted R^2, denoted as R^2_{adj}, does not necessarily increase with the introduction of additional variables into the model. Therefore, it is possible to compare two models with different numbers of variables using this statistical measure and choose the model for which the R^2_{adj} is the highest.

The R^2_{adj} is given by the following formula:

$$R^2_{adj} = R^2 - \frac{k(1-R^2)}{n-k-1} = 1 - \frac{(1-R^2)(n-1)}{n-p}$$

However, it is essential to note that the R^2_{adj} is not necessarily a perfect square and can even take negative values.

5. Akaike Information Criterion (AIC)

The AIC criterion is commonly used to evaluate various statistical analyses, including multiple regressions (potentially reinforcing an adjusted R^2), time series forecasts, and logistic regressions. Its usage is widespread, and it is valued for its versatility. This criterion is specifically applicable to models estimated through the maximum likelihood method.

The AIC criterion is defined as:

$$AIC = n\log(\frac{SSE}{n}) + 2p + n + 2$$

This criterion represents a trade-off between bias, decreasing with the number of free parameters, and parsimony, the desire to describe the data with the fewest possible parameters. The best model is the one with the lowest AIC. However, it is crucial not to blindly trust AIC when calculated on a small dataset.

Accordingly, there are various versions of the corrected AIC criterion, particularly tailored for adjustments in the case of small sample sizes. One of these versions can be expressed as follows:

$$AICc = n\log(\frac{SSE}{n}) + \frac{n(n+p)}{n-p-2}$$

6. Bayesian Information Criterion (BIC)

The BIC criterion is more parsimonious than the AIC as it imposes a stricter penalty based on the number of variables present in the model. According to Ripley in 2003, AIC was introduced to keep relevant variables during predictions, whereas the BIC criterion aims at selecting statistically significant variables in the model. The best model is the one with the lowest BIC.

The Sawa Bayesian Information Criterion (BIC) is defined by the expression:

$$BIC = n\log\left(\frac{SSE}{n}\right) + 2(p+2)q - 2q^2 \qquad where: q = \frac{n\hat{\sigma}^2}{SCR}$$

$\hat{\sigma}^2$ is the estimation of the pure error variance following the fitting of the full model.

In addition, we have the Schwarz Bayesian information Criteron (SBC) defined by:

$$SBC = n\log\left(\frac{SSE}{n}\right) + p\log(n)$$

The SBC is a simplified alternative to the BIC and may be more suitable in certain situations, particularly when the sample sizes are small.

2.1.2 Forward stepwise selection

When dealing with a large number of variables, selecting the best subset may face statistical challenges due to the increased complexity and potential for overfitting. For these reasons, stepwise methods, which explore a more limited set of models, emerge as appealing alternatives to best subset selection. Stepwise methods encompass three main techniques: forward stepwise selection, backward stepwise selection, and stepwise selection.

The forward selection procedure begins with an "empty" regression equation and gradually incorporates one variable at a time until all significant variables are included or a predefined stopping criterion is met. The most significant variable can be chosen based on the smallest p-value, the highest increase in R^2, or the greatest decrease in the model's Residual Sum of Squares (RSS) compared to other considered predictors. The stopping criterion is satisfied when all remaining variables to be considered have a p-value higher than a specified threshold. This threshold can be a fixed value or determined by criteria like AIC or BIC.

The forward selection method is the most economical as it avoids working with more variables than necessary and improves the equation at each step. However, the major inconvenience of the forward selection method is that once a variable is introduced into the model, it cannot be eliminated. This may lead to including non-significant variables in the final model. This issue is addressed in Section 2.1.4 by the stepwise procedure.

2.1.3 Backward stepwise selection

Similarly, backward stepwise selection offers an efficient alternative to best subset selection. However, unlike forward stepwise selection, it initiates with the full least squares model containing all variables and then iteratively removes the variables contributing the least to the model's performance, one at a time.

2.1.4 Stepwise selection

Stepwise regression is a procedure that combines elements from both forward and backward techniques: it follows a forward-oriented approach by incorporating aspects of the backward idea. The procedure initiates like forward selection, including one variable at a time. However, after each selection step, an additional elimination step is introduced. This allows for the exclusion of variables added previously if they prove to be less influential in the current model. The procedure concludes when no further inclusions or exclusions of explanatory variables are necessary or possible, resulting in the final model.

It is an improvement over the forward method because at each step, we reevaluate all variables previously introduced into the model. Indeed, a variable considered the most significant at one stage of the algorithm may become non-significant at a later stage. This phenomenon is due to its correlations with other variables introduced later into the model.

The stepwise procedure appears to be the most effective variable selection method among those studied previously.

2.2 Machine learning

Machine learning is a branch of artificial intelligence that empowers computers to learn from data, akin to human learning. It enables software applications to refine their performance continuously through a diverse range of techniques. Machine learning algorithms, honed to identify patterns and connections buried within data, utilize historical information to make forecasts, categorize information, cluster data points, and even generate novel content. This capability is being harnessed in groundbreaking applications like ChatGPT, Dall-E 2, and GitHub Copilot.

In the realm of Machine learning, penalized regressions emerge as a potent family of supervised learning algorithms that leverage a penalty term to enhance the predictive performance and generalizability of models. These algorithms, which extend from the conventional Ordinary Least Squares (OLS) approach, introduce a trade-off between minimizing the Residual Sum of Squares (RSS) and regulating the magnitude of the regression coefficients. While this penalty introduces bias to the coefficient estimates, it concurrently lowers the variance, leading to improved overall prediction accuracy and a reduced risk of overfitting.

Amongst the notable penalized regressions are Least Angle Regression (LARS), Least Absolute Shrinkage and Selection Operator (LASSO), Ridge Regression and Elastic Net.

2.2.1 Least Angle Regression (LARS)

Introduced by Efron and al. (2004), LARS, works similarly to the forward stepwise regression and is related to LASSO regression (cf. Section 2.2.2). Instead of including one variable per step, LARS involves retaining the variables selected in previous iterations.

Algorithm:

1. Standardize all the predictors to have a null mean and a variance of 1. Start with the residuals $r_i = y - \hat{y}$ with $\beta = (\beta_1^i, \beta_2^i, ..., \beta_p^i) = 0$

2. Find the predictor x_j that is the most correlated with r_i.

3. Gradually move β_j towards their OLS estimators with the current residuals r, until some other competitor β_k are as much correlated with the current residual as does x_j. Increase the information set.

4. Gradually move (β_j, β_k) towards their OLS estimators with the residuals, until a variable β_l presents a stronger correlation with the current residuals. Increase the information set.

5. Continue in this way until all p predictors are entered in the model.

The LARS algorithm is extremely efficient, needing the same order of computation as that of a single least squares fit using the p predictors. LARS always require

p steps to get to the full least squares estimates. The LASSO path can have more than p steps, although the two can often be relatively similar.

Furthermore, it is important to note that LARS has two major drawbacks: i) It is very sensitive to noise. ii) It is based on the correlations between the predictors and the residuals. Hence, it does not behave well when the model represents multicollinearity.

2.2.2 Least Absolute Shrinkage and Selection Operator (LASSO)

Introduced by Tibshirani in 1996, LASSO is a penalized regression method aimed at adding a constraint on the coefficients using the L1 norm as a penalty function. The LASSO method is widely used, particularly in large dimensions. It aims to shrink specific coefficients while emphasizing those with the most impact.

Let's suppose that we have $X = (x_1, x_2, ..., x_p)$, the matrix of covariates and y the response. The x_i's are assumed to be standardized with unit standard deviation and a null mean, y also has mean zero. Then considering $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)$, the LASSO estimate $\hat{\beta}_{LASSO}$ is defined as:

$$\hat{\beta}_{LASSO} = \arg\min_{\beta} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

When λ is close to zero, the results tipically converge towards the OLS solution. Conversely, a larger λ causes the process to yield a parsimonious model, wherein certain coefficients are set to zero. At last, the final model is selected using various criteria such as AIC, Cp, PRESS, etc.

2.2.3 Ridge regression

Ridge regression is a method that uses the L2 norm as the penalty function and estimates the coefficients of multiple-regression models in situations where the explanatory variables are strongly correlated.

The Ridge regression estimates are obtained by minimizing the following expression:

$$\arg\min_{\beta} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Similarly to LASSO, when $\lambda = 0$, the coefficients are estimated without penalty, converging to the OLS estimates. The purpose of Ridge regression is to homogenize the value of the model's variables. The Ridge method does not remove variables from a model like the LASSO method does.

2.2.4 Elastic Net

Elastic Net is a regularized regression developped by Zou and Hastie in 2005, that combines the penalties of the LASSO and Ridge methods. The Elastic Net approach overcomes the limitations of the LASSO technique which uses the L1 norm. Using this penalty function has several limitations. For example, if our data presents "large p, small n", the LASSO selects at most n variables before saturating. Moreover, in presence of multicollinearity, the LASSO model tends to select one variable from a correlated group and disregard the others. This issue can be overcomed by using the Elastic Net which adds a quadratic term to the penalty. When used alone, we get back to the Ridge regression. The estimates from the Elastic Net approach are defined by solving:

$$\arg\min_{\beta} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 + \lambda \times (\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2)$$

Here, α and λ are tuning parameters. When $\alpha = 0$, the Elastic Net reduces to the Ridge regression, and when $\alpha = 1$, it becomes the LASSO regression. Theoretically, Elastic Net performs well in the presence of multicollinearity due to the L2 norm limiting the shrinkage of the L1 norm. However, its performance is highly dependent on the appropriate setting of the α and λ parameters and may lead to poor performance.

2.3 Stopping criteria

Besides the criteria mentioned in Section 2.1.1, we also have Cp, CV and PRESS that perform well as stopping criteria.

1. **The Mallows' Cp**

The Mallows' Cp is a criterion that compares the accuracy and bias of the complete model to models with a subset of predictors.

$$C_p = \frac{SSE}{\hat{\sigma}_c^2} + 2p - n$$

where SSE is the sum of squared residuals, $\hat{\sigma}_c^2$ is the estimate of the model variance, n is the number of observations, and p is the number of parameters in the model.

It is customary to seek a model that minimizes C_p while yielding a value that is lower and close to p.

2. **K-fold Cross Validation (K-fold CV)**

The original sample is randomly divided into k equal-sized sub-samples in k-fold cross-validation, often referred to as "folds". A single sub-sample is retained among the k sub-samples, to use as validation data and to test the model. The remaining

$k-1$ sub-samples are used as training data. The CV procedure is then repeated k times, with each of the k sub-samples used exactly once as the validation data. The k results can then be averaged to make a single estimation. This technique offers an advantage compared to repeated random sub-sampling as it allows all observations to be used in both training and validation phases.

3. Leave-One-Out Cross-Validation (LOOCV)

LOOCV is an extreme case of the k-fold CV. Indeed, when k equals to the number of observations N, the model is estimated N times and validated on each of the N observations in the sample. LOOCV has the maximum computational cost and is not appropriate for larger datasets.

4. PRESS Statistic

The Predicted Residual Sum of Squares Statistic (PRESS) is a criterion that can be used to compare regression models. For a dataset of size n, we calculate the regression equation $\hat{y}_{(i)}$ by using the $n-1$ observations. The formula for PRESS is thus given by:

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2$$

In general, the predictive ability of the model is better when the PRESS value is smaller.

3 Method

To evaluate the performance of the algorithms presented in the previous sections, we first generated the data. We simulated four Data Generating Processes (DGPs) with variations in structure, including differences in correlation and the presence of outliers. This way, we can determine under which conditions each algorithm performs best. Therefore, each DGP consists of 1000 databases, each containing 100 observations. The columns consist of one dependent variable (Y) and 50 explanatory variables (X1 to X50), resulting in a dataset of dimension (100, 51).

3.1 First Data Generating Process (DGP1)

To begin with, we simulated our first DGP in which each dataset follows the basic assumptions of econometrics, such as residual errors of the model should be independent and identically distributed random variables, following a normal distribution. There are no correlations between the explanatory variables, so we use the identity matrix as the variance-covariance matrix:

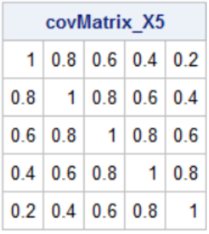
$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

After generating the dataset using the multivariate normal distribution, we can build the following equation: $Y = 1.5X_1 + 0.9X_2 + X_3 + 1.8X_4 - 0.5X_5$. This will serve as the 'good' model with the variables we aim to select.

3.2 Second Data Generating Process (DGP2)

For our second DGP, we replicate the previously outlined procedure, with the introduction of multicollinearity. This can be done by using the Toeplitz matrix as the variance-covariance matrix so that the correlation among the first five variables is positive and symmetric, but decreases as we approach X5. Since there is no correlation between Y and the variables from the sixth to the last, we employ the identity matrix.

The variance-covariance matrix of the variables X1 to X5 has the following structure :



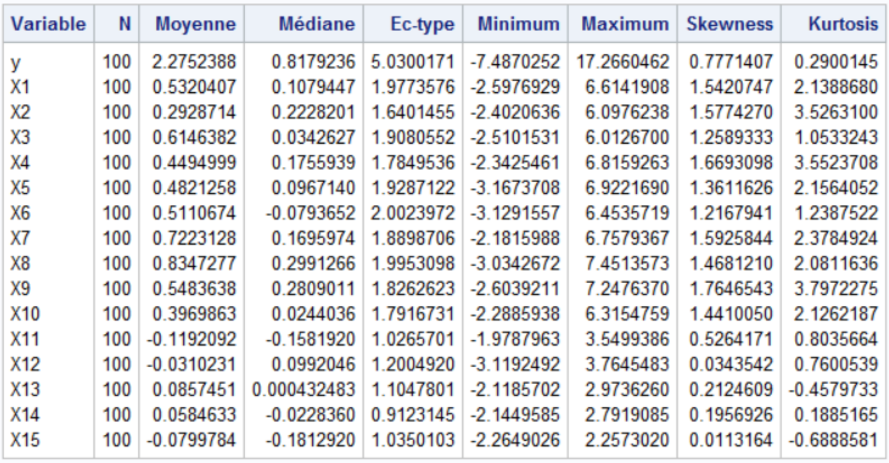
covMatrix_X5				
1	0.8	0.6	0.4	0.2
0.8	1	0.8	0.6	0.4
0.6	0.8	1	0.8	0.6
0.4	0.6	0.8	1	0.8
0.2	0.4	0.6	0.8	1

Figure 1: Correlation between the first 5 variables

We first generated a dataset for X1 to X5 with the multivariate normal distribution using this matrix. Then we generated a second dataset for X6 to X50 using the identity matrix. Finally, we assembled the two datasets into one, forming a matrix X, that has the same dimension as the datasets in DGP1. The values of Y are also generated with the same equation as in DGP1.

3.3 Third Data Generating Process (DGP3)

The third DGP adds outliers to the datasets. We simulated a dataset using the multivariate normal distribution. We have no correlation between the explanatory variables, but we have extreme values in the first ten explanatory variables. Indeed, in the table below, we can see that the skewness of the first 10 variables is positive, meaning that the tail is more pronounced on the right side. Whereas, the five next variables have a skewness close to zero, indicating a normal distribution.



Variable	N	Moyenne	Médiane	Ec-type	Minimum	Maximum	Skewness	Kurtosis
y	100	2.2752388	0.8179236	5.0300171	-7.4870252	17.2660462	0.7771407	0.2900145
X1	100	0.5320407	0.1079447	1.9773576	-2.5976929	6.6141908	1.5420747	2.1388680
X2	100	0.2928714	0.2228201	1.6401455	-2.4020636	6.0976238	1.5774270	3.5263100
X3	100	0.6146382	0.0342627	1.9080552	-2.5101531	6.0126700	1.2589333	1.0533243
X4	100	0.4494999	0.1755939	1.7849536	-2.3425461	6.8159263	1.6693098	3.5523708
X5	100	0.4821258	0.0967140	1.9287122	-3.1673708	6.9221690	1.3611626	2.1564052
X6	100	0.5110674	-0.0793652	2.0023972	-3.1291557	6.4535719	1.2167941	1.2387522
X7	100	0.7223128	0.1695974	1.8898706	-2.1815988	6.7579367	1.5925844	2.3784924
X8	100	0.8347277	0.2991266	1.9953098	-3.0342672	7.4513573	1.4681210	2.0811636
X9	100	0.5483638	0.2809011	1.8262623	-2.6039211	7.2476370	1.7646543	3.7972275
X10	100	0.3969863	0.0244036	1.7916731	-2.2885938	6.3154759	1.4410050	2.1262187
X11	100	-0.1192092	-0.1581920	1.0265701	-1.9787963	3.5499386	0.5264171	0.8035664
X12	100	-0.0310231	0.0992046	1.2004920	-3.1192492	3.7645483	0.0343542	0.7600539
X13	100	0.0857451	0.000432483	1.1047801	-2.1185702	2.9736260	0.2124609	-0.4579733
X14	100	0.0584633	-0.0228360	0.9123145	-2.1449585	2.7919085	0.1956926	0.1885165
X15	100	-0.0799784	-0.1812920	1.0350103	-2.2649026	2.2573020	0.0113164	-0.6888581

Figure 2: Descriptive table

The values of Y are generated with the same equation as in DGP1. And we assembled everything into one table for better manipulation as before.

3.4 Fourth Data Generating Process (DGP4)

The last DGP is a combination of DGP2 and DGP3, including multicollinearity and outliers at the same time. Indeed, we observe correlations among the first five variables and encounter extreme values within the first ten variables. This approach aims to create a diverse set of scenarios to thoroughly test and evaluate algorithms. The values of Y are then generated with the same equation as in DGP1. And we assembled everything into one table for better manipulation as before.

4 Results

In this section, we will use the glmselect procedure in SAS for model selection. This procedure supports a variety of model selection methods, both from Statistical learning and Machine learning. For the Statistical learning methods, we will test forward, backward and stepwise regressions. For the Machine learning part, LAR, LASSO and Elastic Net will be tested. In addition, to select the most efficient algorithm, we will apply the following criteria: AIC, AICc, SBC, CP, CV and PRESS. The selection process can be customized with the CHOOSE= option, which allows us to specify the criteria for picking a model from the sequence of models obtained by the selection process. Combined with the STOP= option that allows us to specify criteria for terminating the selection process.

To compare the selection methods, we elaborated four measures:

1. **Success**: The algorithm successfully selected the first 5 variables.
2. **Overfitting**: The algorithm selected the first 5 variables and at least one other variable.
3. **Underfitting**: The algorithm selected fewer than the first 5 variables.
4. **Fail**: The algorithm failed to select all 5 variables and included at least one other variable.

By executing the algorithm over 1000 databases, we will be able to construct percentages. The best selection procedure will be the one with the highest percentage of success, but the lowest percentage of fail, overfitting and underfitting.

4.1 First Data Generating Process (DGP1)

For DGP1 with linear independance, the models performed better with SBC as the stopping criterion.

4.1.1 Statistical learning

We tested Forward, Backward and Stepwise methods for each choosing criterion.

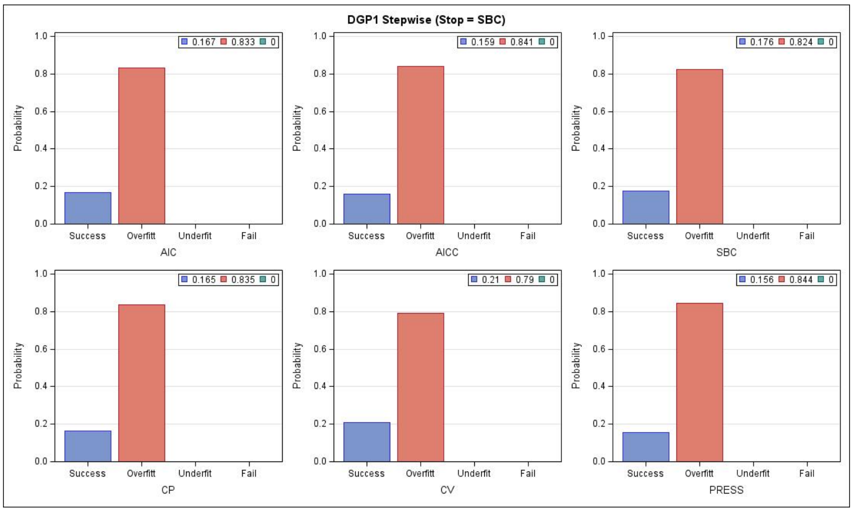


Figure 3: Stepwise

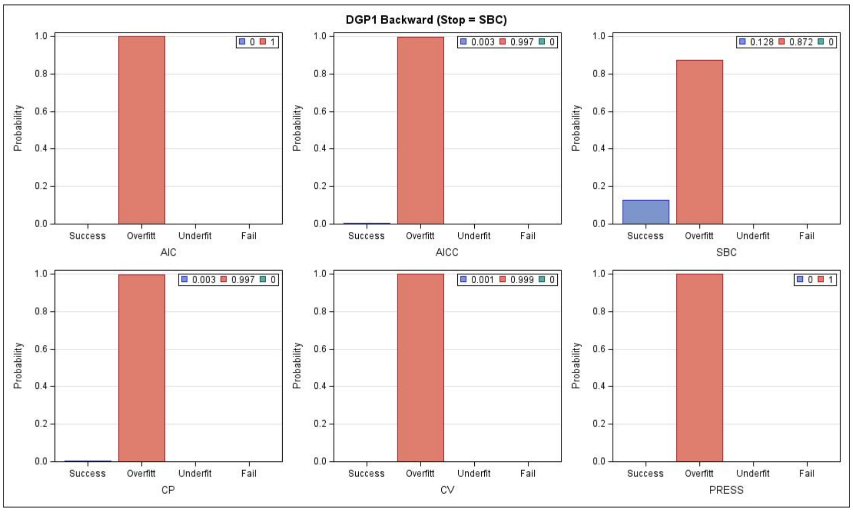


Figure 4: Backward

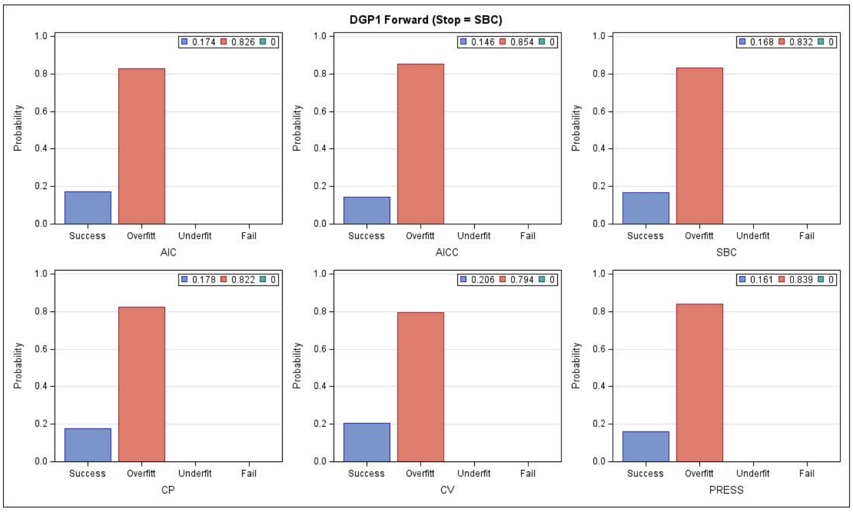


Figure 5: Forward

The optimal Statistical learning models in this context are evidently the stepwise and forward models, utilizing the CV criterion. We achieve a 21% satisfactory fit and encounter a notable 79% overfit. These values fall short of being considered satisfactory. Additionally, the success rate in the backward model is null for the majority of criteria. Consequently, we decide to focus solely on testing the stepwise and forward methods. Let us investigate whether the Machine learning models demonstrate superior performance.

4.1.2 Machine learning

We tested LAR, LASSO and Elastic Net methods for each choosing criterion.

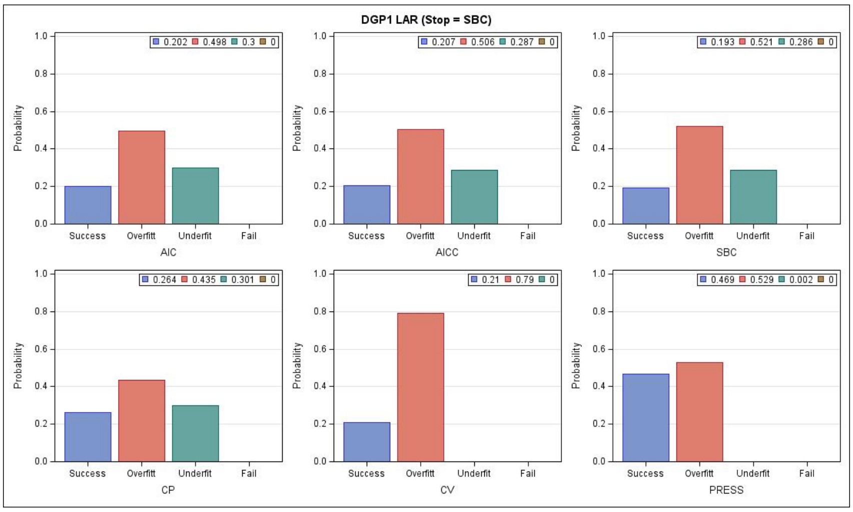


Figure 6: LAR

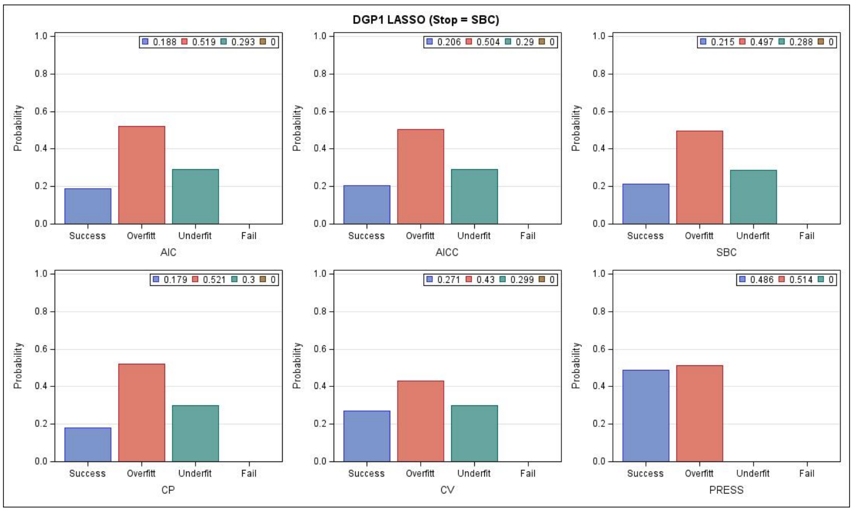


Figure 7: LASSO

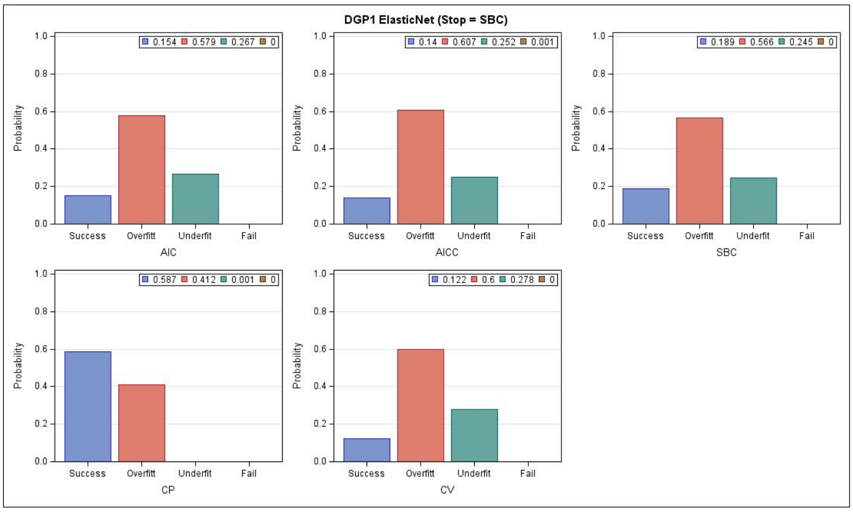


Figure 8: Elastic Net

All Machine learning models outperform the stepwise model in at least one criterion. We observe commendable scores for LASSO and LAR, particularly with the PRESS criterion yielding a success rate of 46.9% for LAR and 48.6% for LASSO. Regarding the Elastic Net, using the CP criterion results in a favorable fit of 58.7%, surpassing the performance of any statistical models.

The preeminent model in Machine learning is the Elastic Net model with the CP criterion combined with the SBC criterion, exhibiting a commendable fit rate of 58.7% and a mere 41.2% overfit.

In summary, for a dataset devoid of outliers or variable correlations, the Elastic Net model emerges as the superior choice. Additionally, in our scenario, Machine learning models outperform statistical models. The latter exhibit, on average, excessively high overfitting values and insufficiently low values for satisfactory fitting.

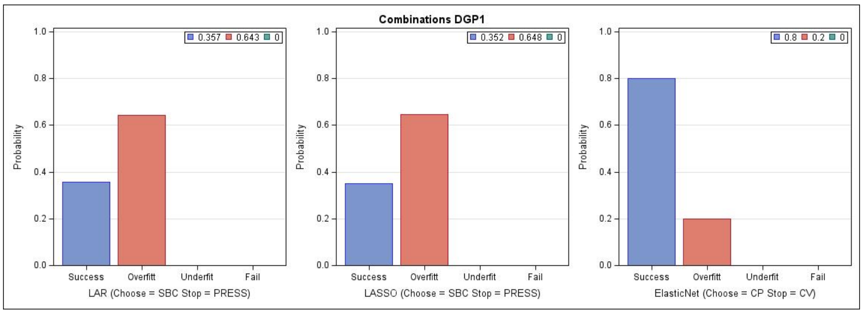


Figure 9: Combinations DGP1 with N = 100

To create combinations of criteria for each Statistical learning and Machine learning model to assess their performance, we evaluated all possible combinations of criteria for each model, and we applied the same process to all the DGP configurations. However, only those combinations that yielded the best results were retained.

In this DGP, the Elastic Net model stands out for its exceptional performance, particularly when incorporating both CP and CV criteria. In contrast to earlier method that relied on SBC criterion, the Elastic Net model achieves a remarkable 80% good fit and only 20% overfit.

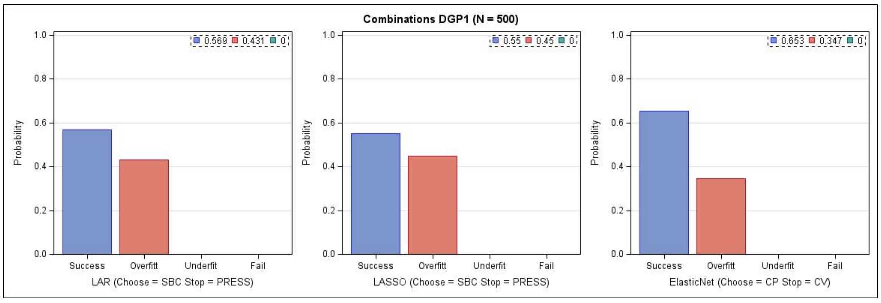


Figure 10: Combinations DGP1 with N = 500

As we increase the number of observations, we observe enhancements in the performance of the LAR and LASSO models with our previous criteria combinations. However, the effectiveness of the elastic net model with CV and CP criteria diminishes.

4.2 Second Data Generating Process (DGP2)

In our second DGP containing multicollinearity, we used the same criterion for the CHOOSE= and STOP= options.

4.2.1 Statistical learning

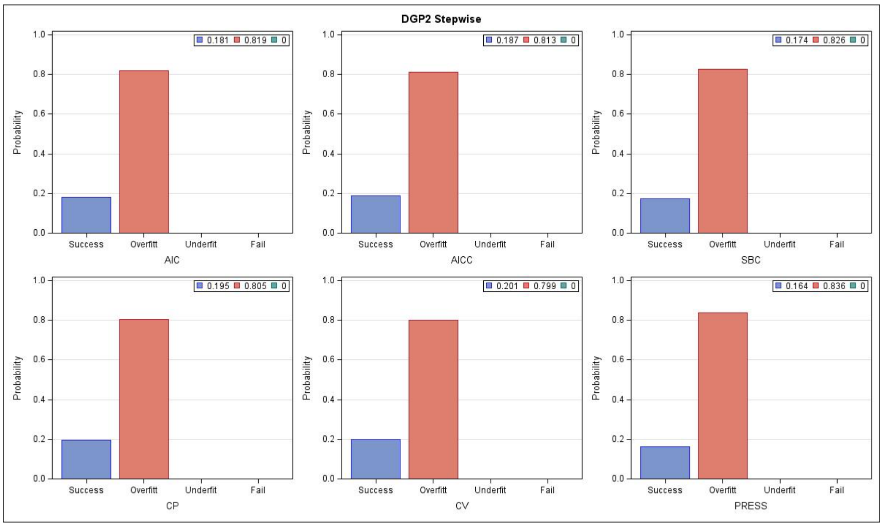


Figure 11: Stepwise

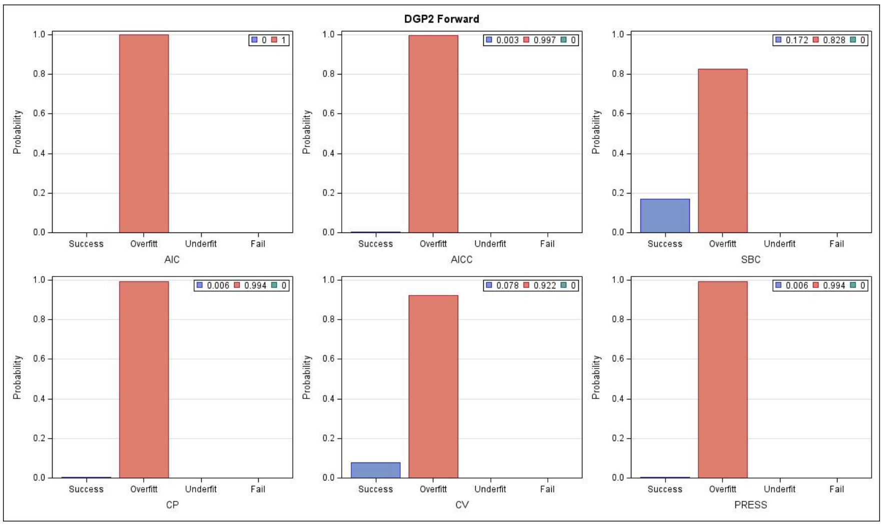


Figure 12: Forward

The correlation among X1 to X5 appears to have a significant impact on the forward model's performance. It demonstrates poor performance on this dataset, in contrast to DGP1. The optimal Statistical learning model is the stepwise model with the CV criterion, achieving a 20.1% satisfactory fit, 79.9% overfitting and underfitting.

4.2.2 Machine learning

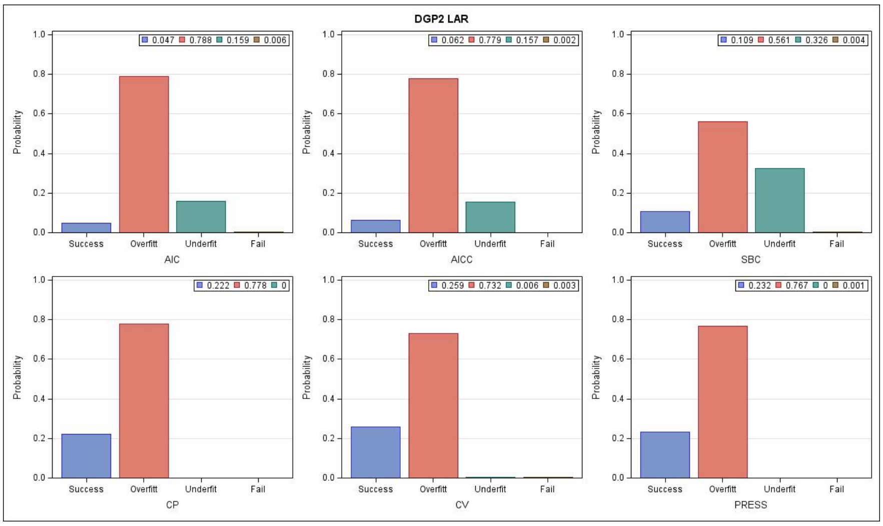


Figure 13: LAR

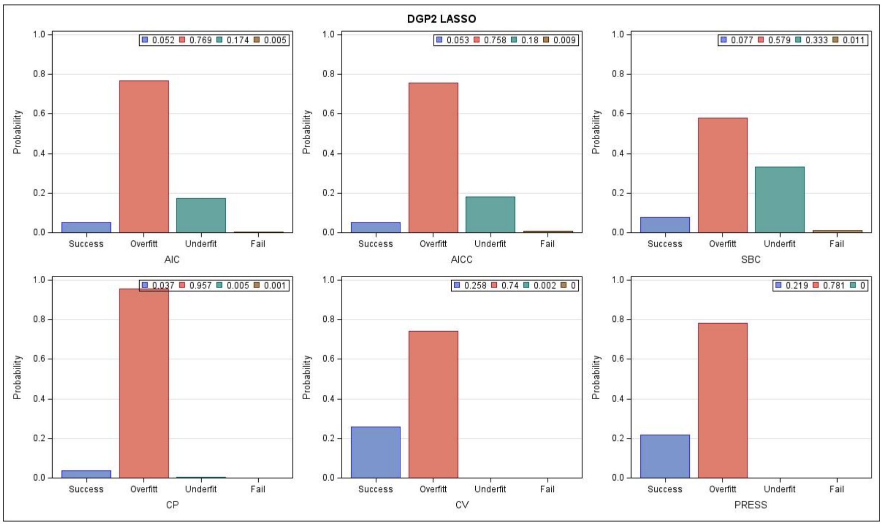


Figure 14: LASSO

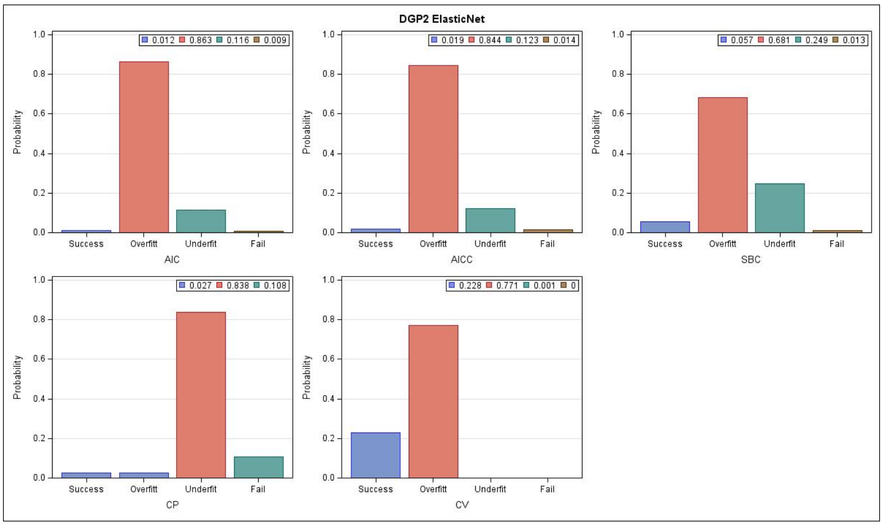


Figure 15: Elastic Net

In the realm of Machine learning models, there are nearly identical good fit scores with the CV criterion: approximately 26% for both LAR and LASSO models and around 23% for Elastic Net models. However, these models did not perform as well as those of DGP1, probably due to the presence of multicollinearity.

In summary, Statistical learning with stepwise model and Machine learning exhibit almost similar results with the CV criterion in this case. However, the LAR and LASSO models have slightly better performance. As for the Elastic Net model, despite the expectation of good theoretical results, its performance did not meet our initial expectations in our case. Let's test different combinations to see if there is an improvement.

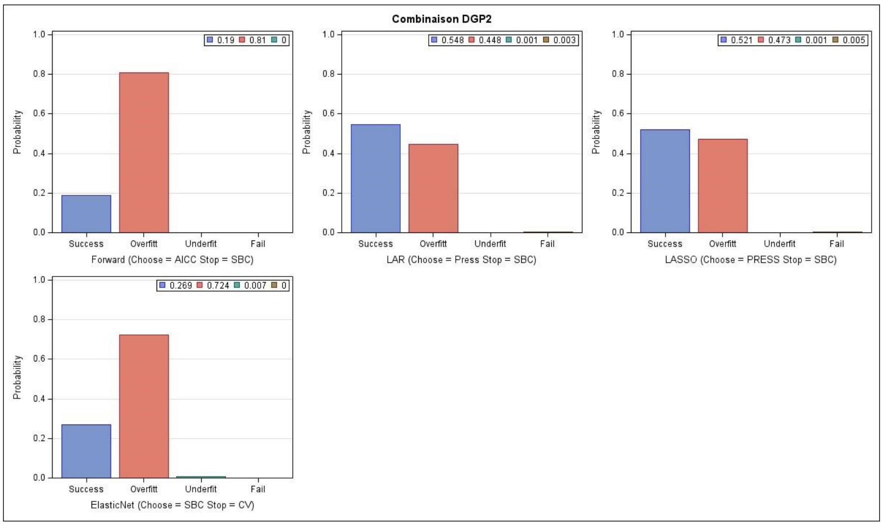


Figure 16: Combinations DGP2 with N = 100

Combining PRESS and SBC criteria in the LAR or LASSO models yields much better performance results than previously, with an approximate 54.8% good fit and an approximate 44.8% overfit. Consequently, within this DGP, combining criteria once more proved its effectiveness in enhancing model performance.

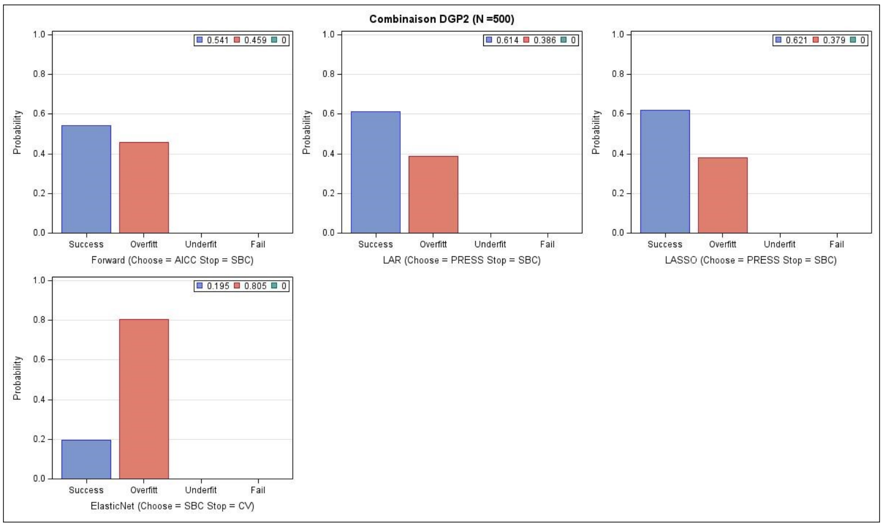


Figure 17: Combinations DGP2 with N = 500

Yet, with 500 observations, we notice unexpected improvements in the forward model's performance when combining AICC and SBC criteria, as well as in the LAR and LASSO models' performance when combining SBC and PRESS criteria. However, the Elastic Net model's performance decreases in larger dataset.

4.3 Third Data Generating Process (DGP3)

In our second DGP containing outliers, we used the same criterion for the CHOOSE= and STOP= options.

4.3.1 Statistical learning

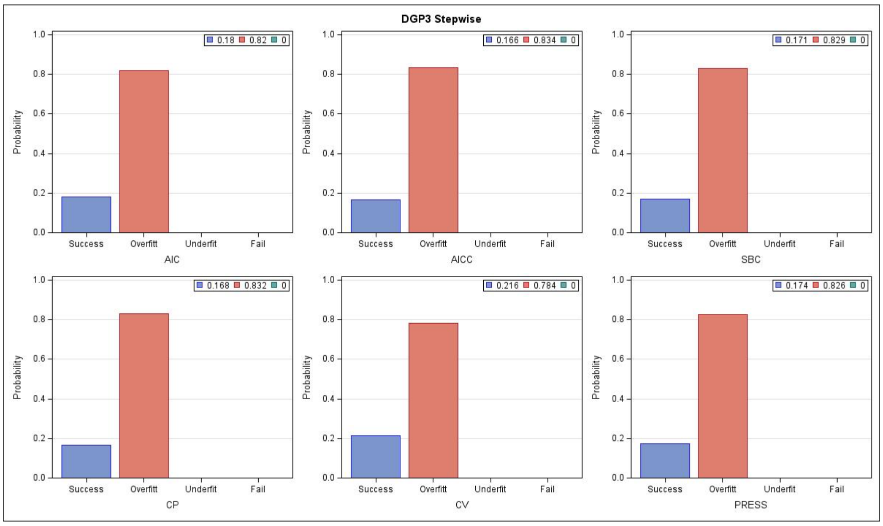


Figure 18: Stepwise

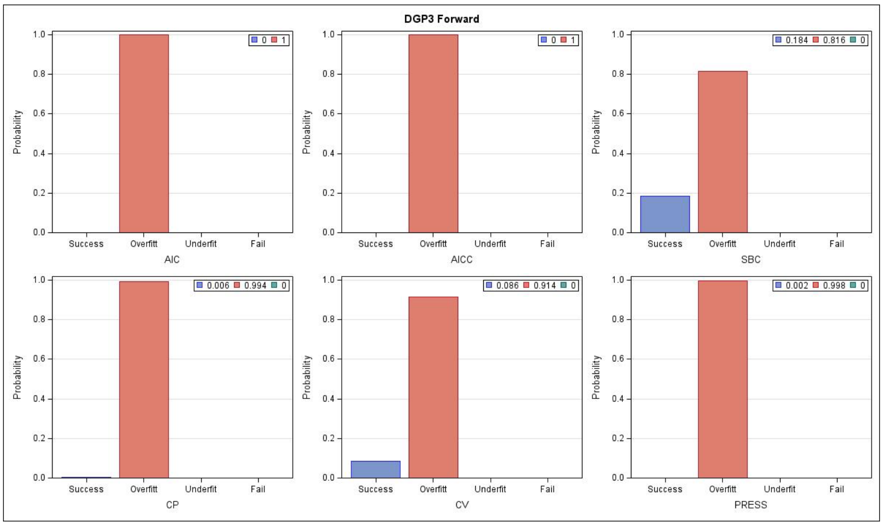


Figure 19: Forward

In Statistical learning, stepwise model with the CV criterion is the best model: it has 21.6% good fit and 78.4% overfit. Conversely, the forward model exhibits dismal good fit scores for all criteria except for SBC, where it reaches nearly 18.4%. Its good fit rates generally hover around 0%, accompanied by overfitting values exceeding 91% across most criteria.

4.3.2 Machine learning

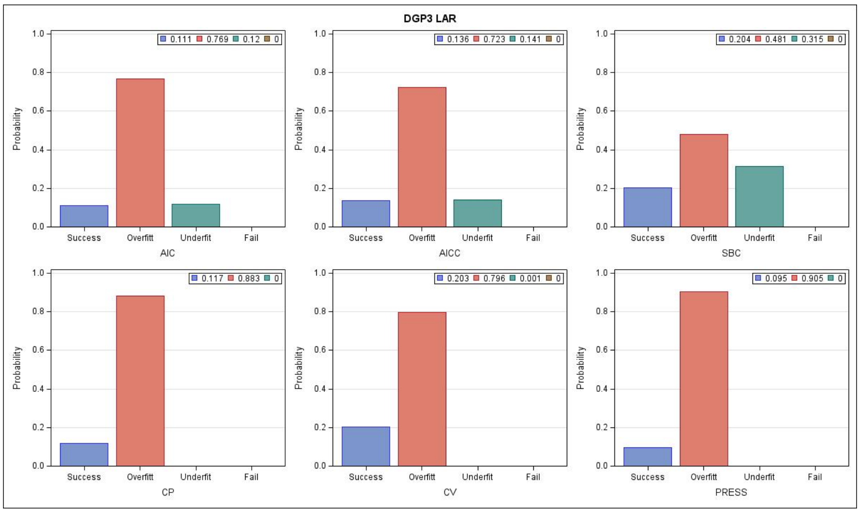


Figure 20: LAR

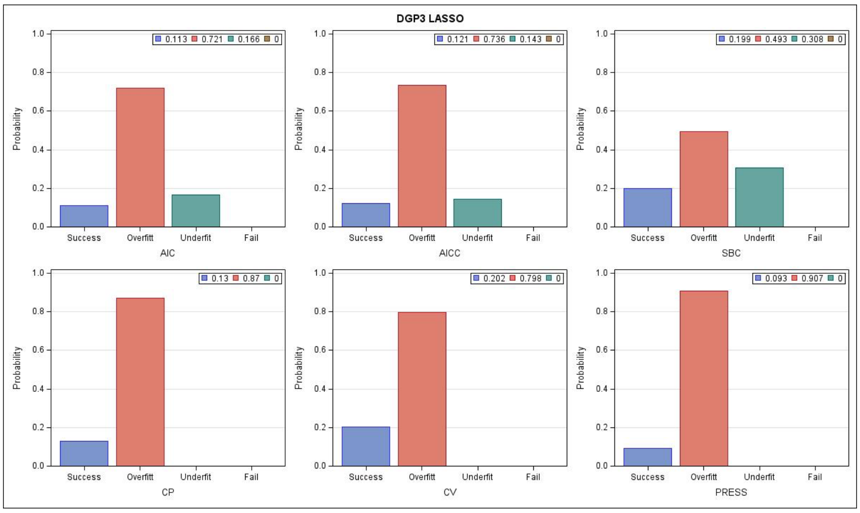


Figure 21: LASSO

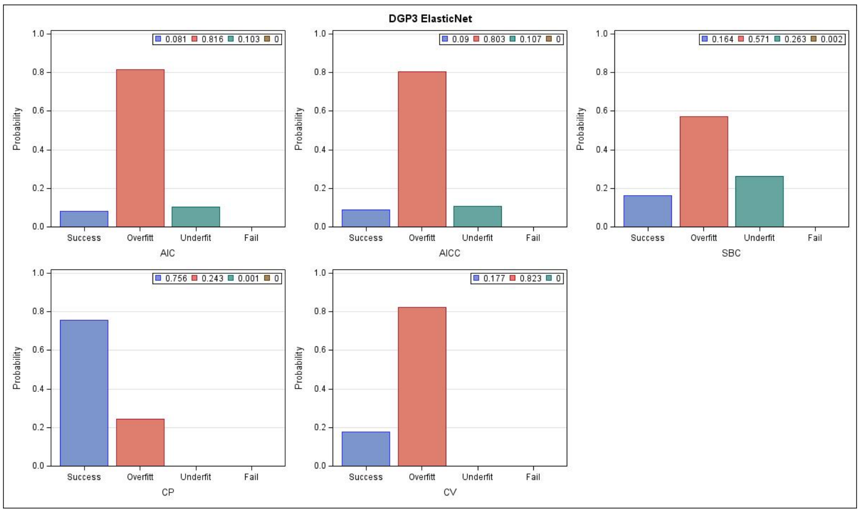


Figure 22: Elastic Net

Turning to Machine learning, the Elastic Net model, employing the CP criteria, demonstrates superior performance with an exceptional good fit of 75.6% and minimal overfitting and underfitting at 24.3% and 0.1%, respectively, marking highly satisfactory results.

In summary, when dealing with a dataset containing outliers, the Elastic Net model with the CP criterion prove to be the most effective in discerning accurate predictions.

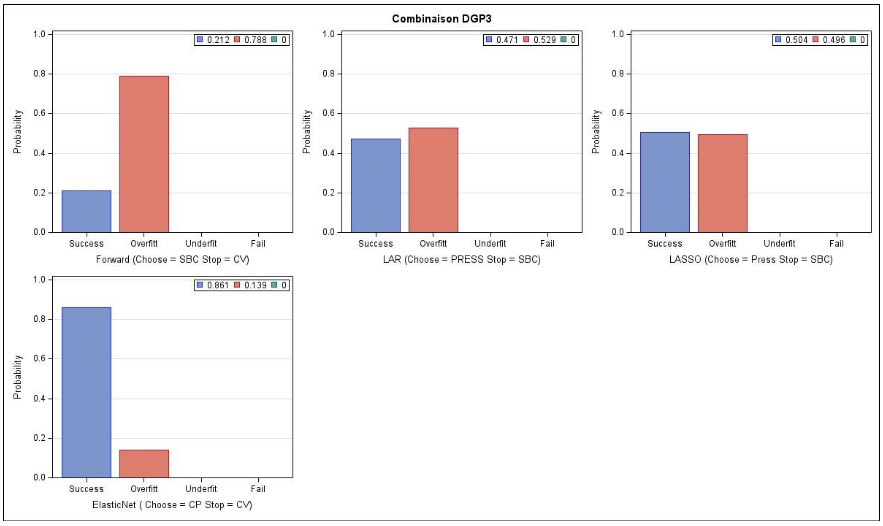


Figure 23: Combinations DGP3 with N = 100

Within this Data Generating Process, just like DGP1, the Elastic Net model, incorporating both CP and CV criteria, demonstrates notably superior performance compared to previous methods utilizing individual criteria. It achieves an impressive 86.1% good fit and a mere 13.9% overfit, establishing it as the most effective model for this DGP.

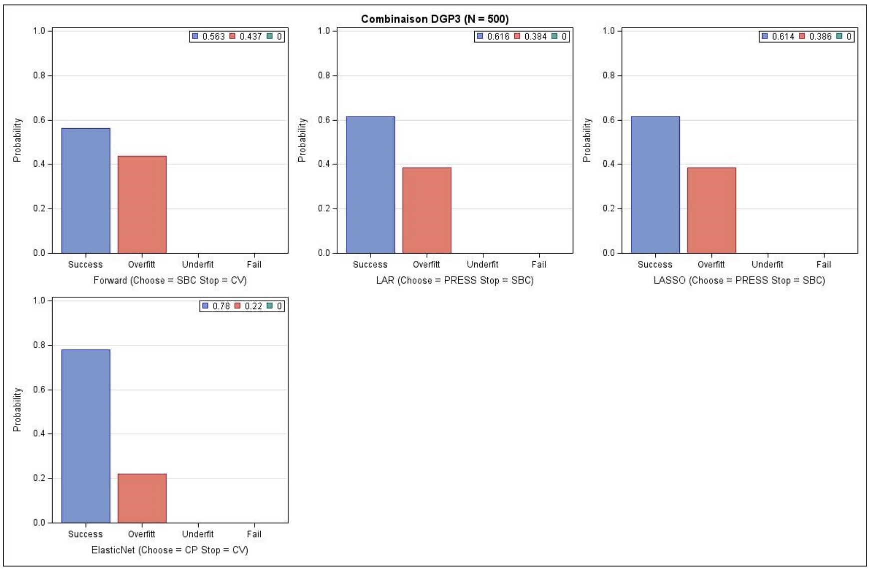


Figure 24: Combinations DGP3 with N = 500

Similar to DGP2, increasing the number of observations results in improved performance for the forward model with combined CV and SBC criteria, as well as for the LAR and LASSO models with combined SBC and PRESS criteria. However, the Elastic Net model's performance appears to slightly decrease with larger dataset.

4.4 Fourth Data Generating Process (DGP4)

At last, in this fourth DGP with the presence of outliers and multicolinearity, the models performed better with SBC as the stopping criterion.

4.4.1 Statistical learning

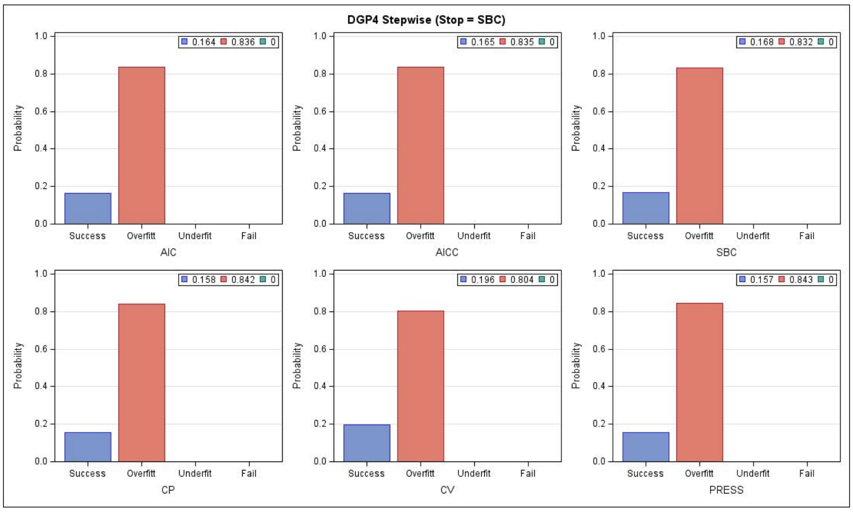


Figure 25: Stepwise

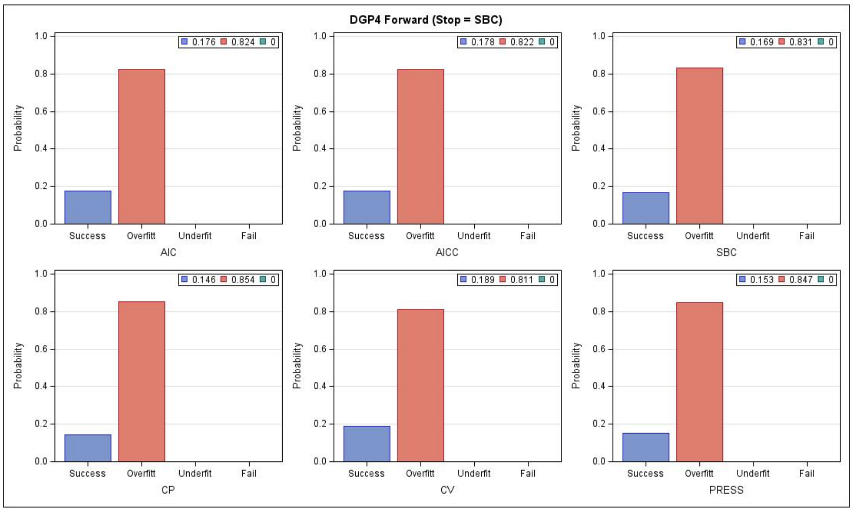


Figure 26: Forward

In Statistical learning, the best model involves employing the stepwise model with the CV criterion, resulting in a 19.6% good fit and 80.4% overfit.

4.4.2 Machine learning

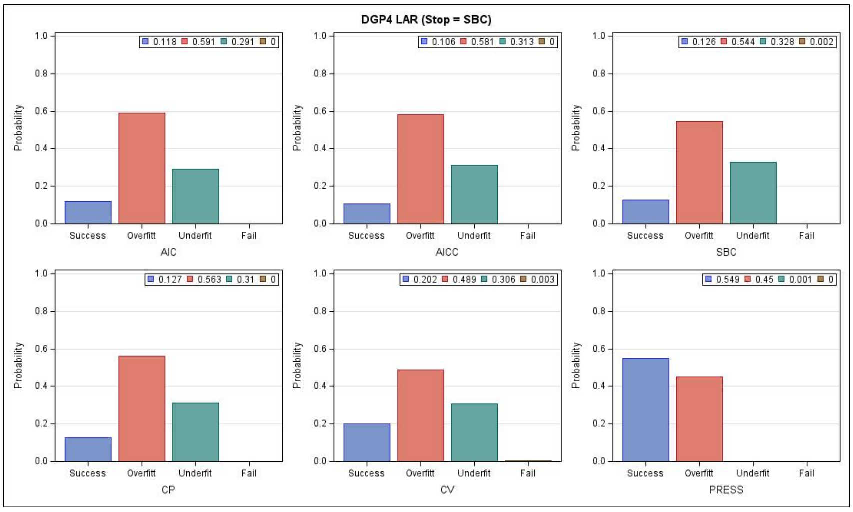


Figure 27: LAR

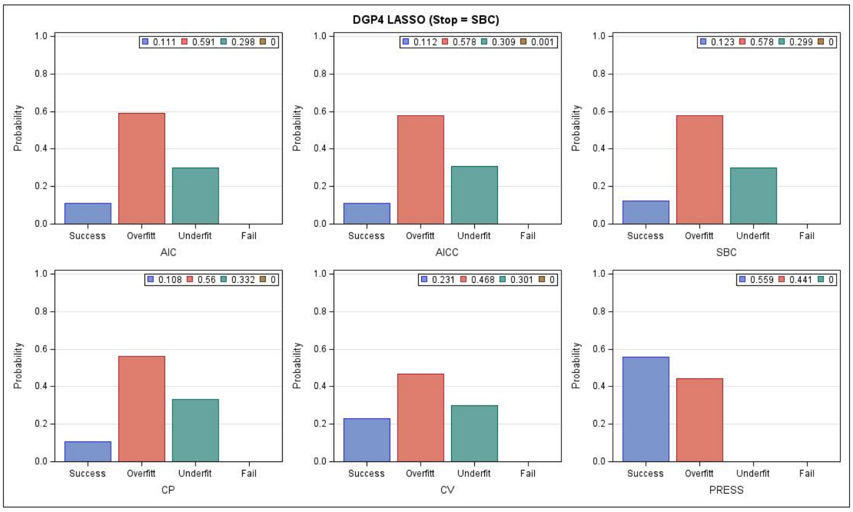


Figure 28: LASSO

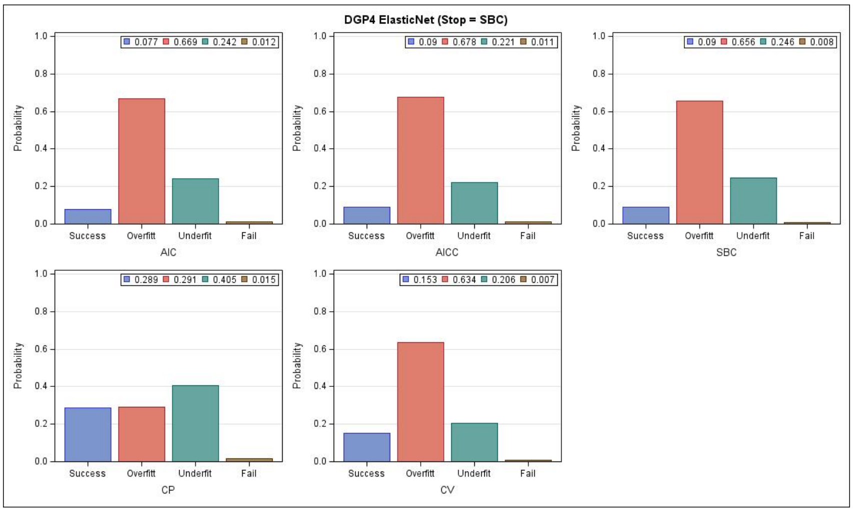


Figure 29: Elastic Net

In the context of Machine learning, the LAR and LASSO models, when assessed using the PRESS criterion, proves to be the most efficient, showcasing an approximate 55% good fit and an approximate 44% overfit.

In conclusion, when confronted with a dataset containing outliers and correlated variables, opting for LAR or LASS models with the PRESS criterion proves to be the most effective approach for achieving accurate predictions.

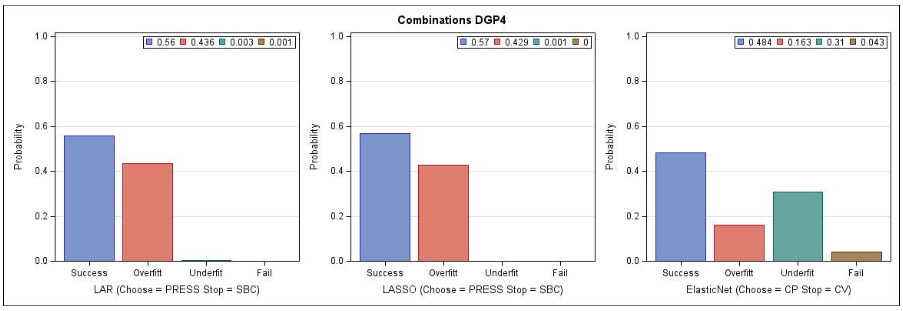


Figure 30: Combinations DGP4 with N = 100

In this DGP, the LAR and LASSO models demonstrated equivalent performance when using a combination of PRESS and SBC criteria compared to assessing them individually with PRESS Criteria. However, the performance of the Elastic Net model notably improved when employing a combination of CV and CP criteria, achieving a 48.8% good fit rate with minimal overfit at 16.3%.

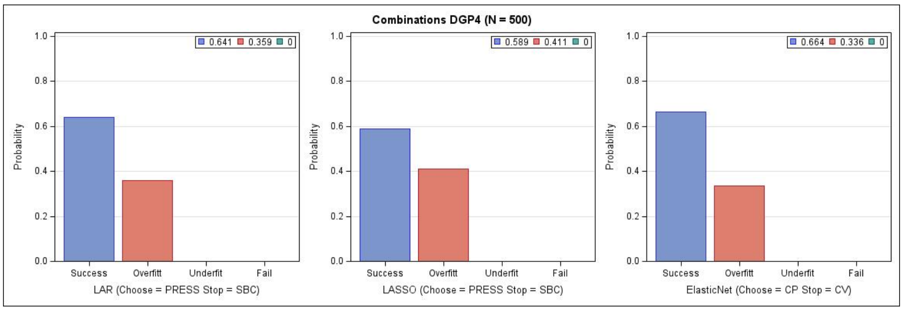
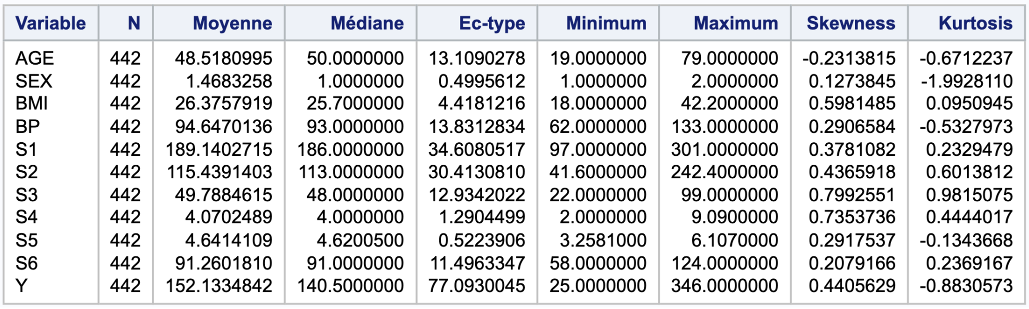


Figure 31: Combinations DGP4 with N = 500

As the number of observations increases, there is an enhancement in the performance of the Machine learning models. Overall, the three models give fairly similar results, with LASSO performing least well.

5 Empirical Analysis : Diabete Database

After conducting a theoretical comparison between Machine learning and Machine learning methodologies, it would be valuable to apply our most effective approach to real-world data. We've opted to utilize a database related to diabetes. This database comprises 442 observations, with a single response variable (Y) and 10 predictor variables, encompassing factors such as age, gender, body mass index (BMI), average blood pressure (BP), and six blood serum measurements (S1 to S6).

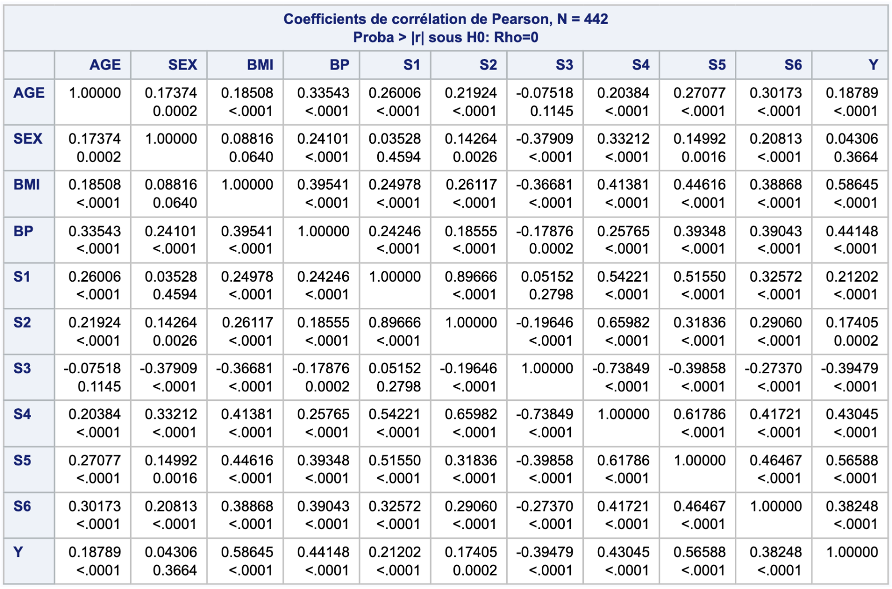


Variable	N	Moyenne	Médiane	Ec-type	Minimum	Maximum	Skewness	Kurtosis
AGE	442	48.5180995	50.0000000	13.1090278	19.0000000	79.0000000	-0.2313815	-0.6712237
SEX	442	1.4683258	1.0000000	0.4995612	1.0000000	2.0000000	0.1273845	-1.9928110
BMI	442	26.3757919	25.7000000	4.4181216	18.0000000	42.2000000	0.5981485	0.0950945
BP	442	94.6470136	93.0000000	13.8312834	62.0000000	133.0000000	0.2906584	-0.5327973
S1	442	189.1402715	186.0000000	34.6080517	97.0000000	301.0000000	0.3781082	0.2329479
S2	442	115.4391403	113.0000000	30.4130810	41.6000000	242.4000000	0.4365918	0.6013812
S3	442	49.7884615	48.0000000	12.9342022	22.0000000	99.0000000	0.7992551	0.9815075
S4	442	4.0702489	4.0000000	1.2904499	2.0000000	9.0900000	0.7353736	0.4444017
S5	442	4.6414109	4.6200500	0.5223906	3.2581000	6.1070000	0.2917537	-0.1343668
S6	442	91.2601810	91.0000000	11.4963347	58.0000000	124.0000000	0.2079166	0.2369167
Y	442	152.1334842	140.5000000	77.0930045	25.0000000	346.0000000	0.4405629	-0.8830573

Figure 32: Descriptive table

Initially, we examined our diabetes database for outliers. Assessing skewness coefficients provides insight into the distribution's symmetry. A positive skewness indicates a right-skewed distribution and the presence of outliers, observed for variables S3 and S4 with coefficients of 0.79 and 0.73 respectively. Conversely, negative skewness suggests a left-skewed distribution, also with outliers.

Additionally, examining the kurtosis coefficient, or the flattening coefficient, sheds light on the distribution's tails. Higher kurtosis indicates observations farther from the mean. Some variables, such as S2 or S3, exhibit positive kurtosis, suggesting heavier tails compared to a Gaussian distribution. Conversely, variables like SEX and AGE show negative kurtosis, implying fewer observations in the tails.



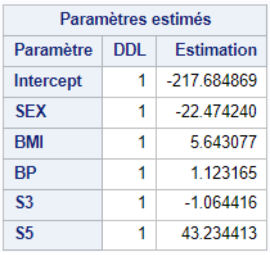
Coefficients de corrélation de Pearson, N = 442
Proba > |r| sous H0: Rho=0

	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
AGE	1.00000	0.17374 0.0002	0.18508 <.0001	0.33543 <.0001	0.26006 <.0001	0.21924 <.0001	-0.07518 0.1145	0.20384 <.0001	0.27077 <.0001	0.30173 <.0001	0.18789 <.0001
SEX	0.17374 0.0002	1.00000	0.08816 0.0640	0.24101 <.0001	0.03528 0.4594	0.14264 0.0026	-0.37909 <.0001	0.33212 <.0001	0.14992 0.0016	0.20813 <.0001	0.04306 0.3664
BMI	0.18508 <.0001	0.08816 0.0640	1.00000	0.39541 <.0001	0.24978 <.0001	0.26117 <.0001	-0.36681 <.0001	0.41381 <.0001	0.44616 <.0001	0.38868 <.0001	0.58645 <.0001
BP	0.33543 <.0001	0.24101 <.0001	0.39541 <.0001	1.00000	0.24246 <.0001	0.18555 <.0001	-0.17876 0.0002	0.25765 <.0001	0.39348 <.0001	0.39043 <.0001	0.44148 <.0001
S1	0.26006 <.0001	0.03528 0.4594	0.24978 <.0001	0.24246 <.0001	1.00000	0.89666 <.0001	0.05152 0.2798	0.54221 <.0001	0.51550 <.0001	0.32572 <.0001	0.21202 <.0001
S2	0.21924 <.0001	0.14264 0.0026	0.26117 <.0001	0.18555 <.0001	0.89666 <.0001	1.00000	-0.19646 <.0001	0.65982 <.0001	0.31836 <.0001	0.29060 <.0001	0.17405 0.0002
S3	-0.07518 0.1145	-0.37909 <.0001	-0.36681 <.0001	-0.17876 0.0002	0.05152 0.2798	-0.19646 <.0001	1.00000	-0.73849 <.0001	-0.39858 <.0001	-0.27370 <.0001	-0.39479 <.0001
S4	0.20384 <.0001	0.33212 <.0001	0.41381 <.0001	0.25765 <.0001	0.54221 <.0001	0.65982 <.0001	-0.73849 <.0001	1.00000	0.61786 <.0001	0.41721 <.0001	0.43045 <.0001
S5	0.27077 <.0001	0.14992 0.0016	0.44616 <.0001	0.39348 <.0001	0.51550 <.0001	0.31836 <.0001	-0.39858 <.0001	0.61786 <.0001	1.00000	0.46467 <.0001	0.56588 <.0001
S6	0.30173 <.0001	0.20813 <.0001	0.38868 <.0001	0.39043 <.0001	0.32572 <.0001	0.29060 <.0001	-0.27370 <.0001	0.41721 <.0001	0.46467 <.0001	1.00000	0.38248 <.0001
Y	0.18789 <.0001	0.04306 0.3664	0.58645 <.0001	0.44148 <.0001	0.21202 <.0001	0.17405 0.0002	-0.39479 <.0001	0.43045 <.0001	0.56588 <.0001	0.38248 <.0001	1.00000

Figure 33: Correlation matrix

Further analysis involved examining the correlation between variables. The correlation matrix, depicted in figure 33, reveals notable findings. For instance, variables like S1 and S2 exhibit a strong positive correlation (0.89), while S3 and S4 are negatively correlated (-0.73).

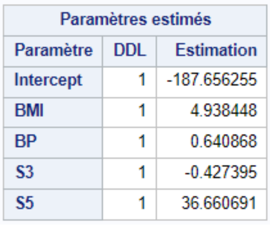
Finally, following a thorough examination of our dataset alongside the theoretical groundwork outlined in this paper, we determined that employing the LAR or ElasticNet Machine learning models with the combinations used figure 31 is pertinent for variable selection in our context. This scenario, characterized by the presence of correlated variables and outliers, aligns with our DGP4, as well as the number of observations that is close to 500.



Paramètres estimés

Paramètre	DDL	Estimation
Intercept	1	-217.684869
SEX	1	-22.474240
BMI	1	5.643077
BP	1	1.123165
S3	1	-1.064416
S5	1	43.234413

Figure 34: Selected variables by LAR (Choose = PRESS Stop = SBC)



Paramètres estimés

Paramètre	DDL	Estimation
Intercept	1	-187.656255
BMI	1	4.938448
BP	1	0.640868
S3	1	-0.427395
S5	1	36.660691

Figure 35: Selected variables by ElasticNet (Choose = CP Stop = CV)

We can see that the LAR method with PRESS and SBC criteria selected SEX, BMI, BP, S3 and S5. Whereas Elastic Net with the CP and CV criteria selected the same variables but failed to include SEX. When considering the adjusted R^2, the model selected by LAR demonstrates an R^2_{adj} of 0.5 whereas the model chosen by Elastic Net has an R^2_{adj} 0.46. Hence, we can conclude that the optimal model is the one chosen by LAR, given its higher adjusted R^2. Both selection algorithms give satisfactory results overall.

6 Discussion

Upon delving into an extensive exploration and testing of Statistical learning and Machine learning methods, this paper aims to elucidate their effectiveness. However, it is essential to acknowledge the limitations inherent in our study. One such limitation pertains to the selection criteria utilized to determine the model with the best performance. While our study primarily focuses on criteria such as SBC and PRESS, alternative criteria like bootstrap resampling could offer valuable insights into model selection. Bootstrap resampling, for instance, involves repeatedly sampling observations with replacement from the dataset to estimate the sampling distribution of a statistic. It can be used to calculate confidence intervals for model parameters and to assess the stability of model estimates. Models that produce consistent estimates across bootstrap samples are generally preferred. Nevertheless, in situations where the original dataset is relatively small or contains limited variability, bootstrap resampling may produce bootstrap samples that closely resemble the original dataset. As a result, the variability captured in the resampled datasets may not fully reflect the variability present in the population. This can lead to overly optimistic estimates of model performance and suboptimal model selection.

In conclusion, while this study contributes significant insights into the effectiveness of Statistical learning and Machine learning methods, it is imperative to recognize and address the inherent limitations. By acknowledging these limitations and suggesting avenues for future research, this paper aims to foster continued exploration and improvement in the field of predictive modeling and data analysis.

7 Bibliography

1. Michael Mitchell. *The discipline of Machine learning.* Carnegie Mellon University, School of Computer Science, Machine Learning, 2006.

2. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Machine Society: Series B (Methodological)*, 58(1), 267-288.

3. Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani. "Least angle regression." *Ann. Statist.* 32 (2) 407–499, April 2004.

4. Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 111-147.

5. Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of Statistical learning* (2nd ed.). New York: Springer.

6. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to Statistical learning with applications in R* (2nd ed.). New York: Springer.

7. Hoerl, A. E., & Kennard, R. W. (1970). *Ridge regression: An attempt at improving the performance of the ordinary least squares estimates.* Technometrics, 12(1), 69-82.

8. Information Criteria and PRESS" in STAT 462: Statistical Computing and Programming by Penn State University (2023)

9. Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2005.

10. McLachlan, Geoffrey J.; Do, Kim-Anh; Ambroise, Christophe (2004). *Analyzing microarray gene expression data.* Wiley.

11. *Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Edition*

12. Vanwinckelen, Gitte (2 October 2019). "On Estimating Model Accuracy with Repeated Cross-Validation."

13. Robert A. Cohen, SAS Institute Inc. Cary, NC. *Introducing the GLMSELECT PROCEDURE for Model Selection*, Paper 207-31.