Journal of Science Research

# Predictive Performance of Cross-Validation Techniques in Classification Models

**Adedeji, O. F.*** **and Olubusoye, O. E.**
Department of Statistics, University of Ibadan, Nigeria
*Corresponding author: Email: defasol@yahoo.com; +234 7067614339

**Abstract**
Machine learning algorithms have proven to be breakthroughs in scientific research and other dynamic research areas. One attribute of machine learning process is data splitting to measure the generalization ability of the learning algorithm. However, due to the nature of sample size attribute (often small sample size) in many clinical and biological studies, data splitting may suffer relevance due to limited samples. This study investigates the learning generalization potential on small sample sizes and ascertains the most appropriate split ratio. Consequently, the study considers a family of Cross Validation (CV) techniques, namely K-fold, Nested and Repeated CV, given different split ratios and sample sizes. The study considers Naïve Bayes and Logistic Regression algorithms over simulation experiments and finds that when the sample size is small, the best split ratio is either 50:50 or 60:40.

**Keywords:** Cross Validation; Classification Model; Machine Learning; K-fold; Overfitting.

## Introduction

The importance of Modeling and Model Section in scientific research cannot be over-emphasized. Modeling is a central foundation of all scientific research and the backbone of all statistical research. While modeling deals with the transformation of an observation or an idea into models that produce an explanation of the data and possibly predictions of the future data, it remains a fact that several statistical algorithms can be used for solving a given statistical problem, so the need to select a statistical “best model” from a set of candidate models given data.

As a result, many criteria have been proposed for predictions of the future data and many are still in the pipeline of research. There are several competing and innovative methods of model selection, however, such criteria can be data-driven. In essence, a criteria could be better with data A and not with data B. In this paper various criteria that can determine the suitability or otherwise of a model in solving the problem at hand are being evaluated.

### *The Essence of Model Selection*

As said earlier, model selection became a concern for scientists since several models can be built to the explanation of a single statistical problem. How do we save time and cost by choosing a single “best model” or subset of a set of competing models given data? This and many more questions necessitated the need for choosing model. it is quite important to address the question of why?" At heart we think that the reasons are pragmatic, having to do with saving computer time and analyst attention. Viewed this way, however, there is no particular reason to choose a single best model according to some criterion. Rather it makes more sense to deselect" models that are obviously poor, maintaining a subset for further consideration. Sometimes this subset might consist of a single model, but sometimes perhaps not. Furthermore, if it is indeed the case that model choice is driven by consideration of costs, perhaps these can be included explicitly to the process via utility functions as suggested by Winkler [1]. Hence we think there are good reasons to challenge the traditional formulations of this problem.

Model selection is an important part of any statistical analysis, and indeed is central to the pursuit of science in general. It is an independent step in the process of developing a functional model or a model for understanding the data-generating mechanism. According to Guyon [2], model selection designates an ensemble of technique used to select a model that best explains some data or phenomena, or best predicts future data, observations or the consequences of actions.

Model selection is what happens after something is observed, and multiple distinct attempts have been made to understand what the observation means. Generally, data are obtained while models are fitted to the data. Each of these models produces an explanation of the data and possibly predictions of future data as well. The task of model selection is to decide which of the several models produces the best and most consistent estimate.

*Model Selection Criteria*

Researchers may decide or have to use a multivariate model to estimate say exposure of effects or use several variables to determine or study another variable. This task is most likely to incur a bewildering variety of potential models. Take for example, a situation where we wish to estimate the effect of one exposure factor and one has data containing seven potential control variables with possible confounding or interactive effects.

The question is how many different models could one construct for the process that generated these data? Basically, there would be numerous models from which one can choose. Thus, there is a high likelihood that there would be a model form corresponding to every way in which one can express the dependent variables as a mathematical function of the focus and auxiliary variables. To select a model form, one will have to choose a subset model from the 2k possible subsets of the explanatory variables to be used in the model. More so, comparing and selecting a good model may attract a variety of challenges ranging from; the use of appropriate priors cum non-informative priors, non-nested models, and uses of small sample sizes etc. Both Frequentists and Bayesian Schools have weighed in on the matter with methods such as F-tests for nested models, AIC, Malow Cp, Exhaustive Search, Cross Validation, Bayes Factor of various flavors, BIC, Bayesian Model Averaging and so on.

*Cross-Validation Technique*

Cross-validation is a model evaluation method that is better than simply looking at the residuals. Residual evaluation does not indicate how well a model can make new predictions on cases it has not already seen. Cross-validation techniques tend to focus on not using the entire data set when building a model. Some cases are removed before the data is modeled; these removed cases are often called the *testing set*. Once the model has been built using the cases left (often called the *training set*), the cases which were removed (testing set) can be used to *test* the performance of the model on the "unseen" data (i.e. the testing set).

Cross-validation allows models to be tested using the full training set using repeated resampling; thus, maximizing the total number of points used for testing and potentially, helping to protect against overfitting. Improvements in computational power, recent reductions in the (computational) cost of classification algorithms, and the development of closed-form solutions (for performing cross validation in certain classes of learning algorithms) make it possible to test thousands or millions of variants of learning models on the data.

Thus, it is now possible to calculate cross-validation performance on a much larger number of tuned models than would have been possible otherwise. However, we empirically show how under such a large number of models the risk for overfitting increases and the performance estimated by cross validation is no longer an effective estimate of generalization.

As mentioned before, both nested cross-validation and repeated cross-validation are designed for model selection. Nested cross-validation utilizes multi-layer cross- validation to tune more parameters, and repeated cross-validation repeats the procedure of generating K-folds to alleviate the randomness of fold generation. In this work, we intend to compare the strength of Nested and Repeated Cross validation techniques in relation to the k-fold Cross Validation method in data classification. We implement the proposed approach in the simulated data assessing its predictive performance and present the comparison with different embedded variable selection methods with respect to predictive performance and selection accuracy. To the best of our knowledge, although the idea of using nested/repeated cross-validation has been mentioned elsewhere, for instance, Stone [3], addressed the idea of cross-validation in research, no existing literature has proposed or assessed a systematic framework to utilize nested/repeated cross validation at computational level.

**Methodology**

In this section, we introduce the proposed method of feature selection and model selection using nested and repeated cross-validation. When building the predictive model, the most critical part for the model is to identify the optimal values of the tuning parameters to achieve the minimum test set error.

One of the widely-used techniques for model selection is *K*-fold cross-validation, for which the final model is chosen when the minimum cross-validation error is achieved [4] In the *K*-fold cross-validation, the original training dataset is randomly divided into *K* subsets of equal size then the following step repeats *K* times: *K* - 1 of the subsets are combined to build the model, and the remaining one subset is used to compute the prediction errors. The *K* sets of predication errors are averaged to produce the cross-validation error. To estimate the optimal value of tuning parameters, a grid of *m* candidate values of tuning parameters are created, and *m* models are built, indexed by different value of tuning parameters. The cross-validation error of each of the *m* models is computed, and the final model is then determined by the model with minimum cross-validation error.

Furthermore, the feature subset also can be determined by the model using some criteria, such as coefficients shrinkage. As mentioned in the introduction section, the commonly used single cross validation is not efficient in dealing with the overfitting of the data in general [5]. Repeated cross-validation is an improved method by generating multiple sets of *K* folds. Also, the cross-validation error is calculated as the average across the repeated partitions. On the other hand, we sometimes want to select features, and use the selected features to build a predictive model. In this case, nested cross-validation can be very useful. To achieve the above goals, we propose a systematic framework of combining nested and repeated cross-validation to build the final model. In the proposed method, the cross-validation is

carried out in two different layers: the inner loop and the outer loop. In the inner loop, the subset of features is selected as candidate features. In the outer loop, only the candidate features selected in the inner loop are carried forward to build the final model.

The performance of nested and repeated cross-validation has not been extensively explored and discussed in the past mainly because of the computational costs. In this work, we show that the nested and repeated cross-validation can improve the predictive performance and selection accuracy over the traditional single cross- validation method.

*Experimental Procedure*

To achieve the objectives of the study, the following procedures are relevant.

1. We stimulate the input features following the Gaussian distribution;

with $x_i \sim N(0, 10)$ with $\mu = (0)$ and $\sigma = (10)$, under $n =$ (10, 20, 30, 40, 50, 100, 200, 500, 1000, 5000, 10000, 50000)

2. We stimulate $y$ from Binomial distribution with $y \sim \beta_1\ (n, 0.5)$ with $p = q = 0.5$

$$f(x) = \begin{cases} y = 0, Bad \\ y = 1, Good \end{cases}$$

3. We use the following training; test split ratio:

***Scenario* 1**: [60 : 40]
***Scenario* 2**: [70 : 30]
***Scenario* 3**: [80 : 20]
***Scenario* 4**: [90 : 10]

4. We formulate the table of performance using the following learning algorithms.
- Naïve Bayes
- Logistics Regression

5. Finally, we consider three (3) cross validation techniques; $k - fold$ and the two subsets of the $k - fold\ CV$ namely repeated $Cross\ Validation$ and $Nested\ Cross\ ValidationV$.

We shall identify the strength and weaknesses of each of the CV approaches using the split ratio, data sample size and learning algorithm as criteria; which will fulfill the research objectives. The following research questions shall be evaluated based on the yardsticks;

1. Is there any significant difference in the performances of the Cross Validation (CV) methods across various training samples?
2. If there are any observable differences in the CV performances across the training samples, then: what n-size does *k-fold CV* perform better than *repeated CV* and *nested CV*? This shall be evaluated on *repeated CV* versus nested CV and nested CV versus repeated CV and *k-fold.*
3. At what split ratio does *k-fold CV* perform better than *repeated CV* and *nested CV*? This shall be evaluated in *repeated CV* and *nested CV*, and *nested CV* versus *repeated CV* and *k-folds CV*.
4. In summary, which of the subset of the *k-fold* method performs best?

And if there is or if there is no significant difference in the *CV methods'* performances across the training samples,

5. What is the general observation of the *CVs* performance across the training samples?

**Results**

Each table on the right is plotted into the bar chats on the left, to explain how the CV techniques perform with different data split ratios and different learning algorithms while we vary the training sizes.

**Report of findings: Case 1 → *Small Training sizes***

Result 1**:** Performance of the CVs when the sample size, n = 20.

**Naïve Bayes Result at n = 20**

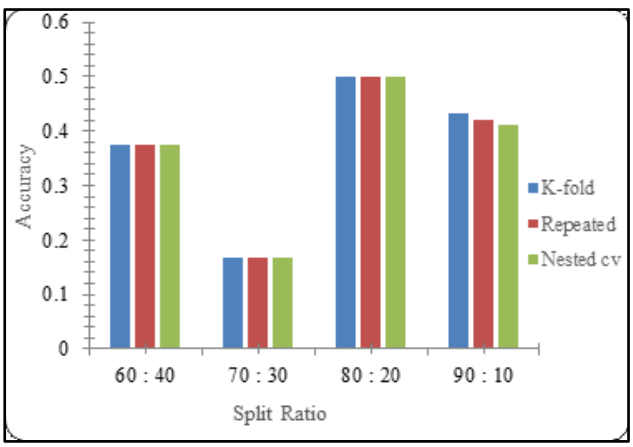


Figure 1: CVs performance for Naïve Bayes when n = 20.

Table 1: CVs performance for Naïve Bayes when n = 20

| n = 20 | K-fold | Repeated | Nested |
|---|---|---|---|
| 60 : 40 | 0.375 | 0.375 | 0.375 |
| 70 : 30 | 0.168 | 0.168 | 0.168 |
| 80 : 20 | 0.500 | 0.500 | 0.500 |
| 90 : 10 | 0.433 | 0.420 | 0.412 |

**Logistic Result at n = 20**

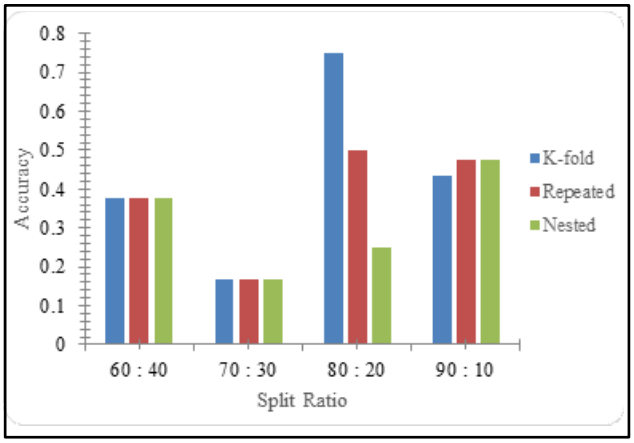


Figure 2: CVs performance for Logistic when n = 20.

Table 2: CVs performance for Logistic when n = 20.

| n = 20 | K-fold | Repeated | Nested |
|---|---|---|---|
| 60 : 40 | 0.375 | 0.375 | 0.375 |
| 70 : 30 | 0.168 | 0.168 | 0.168 |
| 80 : 20 | 0.750 | 0.500 | 0.250 |
| 90 : 10 | 0.433 | 0.473 | 0.475 |

**Naïve Bayes Result at n = 30**

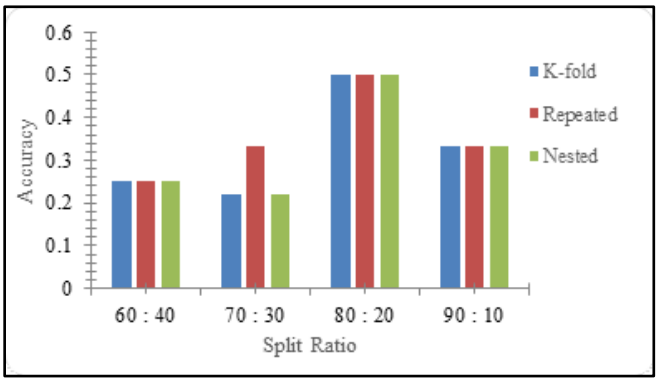


Figure 3: CVs performance for Naïve Bayes when n = 30.

Table 3: CVs performance for Naïve Bayes when n = 30.

| n = 30 | K-fold | Repeated | Nested |
|---|---|---|---|
| 60 : 40 | 0.250 | 0.250 | 0.250 |
| 70 : 30 | 0.222 | 0.333 | 0.222 |
| 80 : 20 | 0.500 | 0.500 | 0.500 |
| 90 : 10 | 0.333 | 0.333 | 0.333 |

**Logistic Result at n = 30**

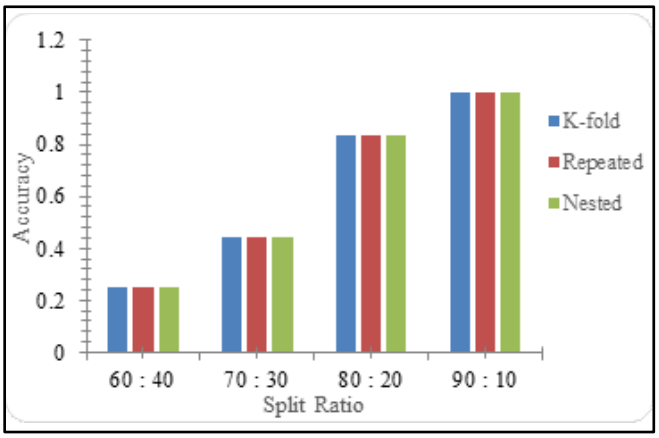


Figure 4: CVs performance for Logistic when n = 30.

Table 4: CVs performance for Logistic when n = 30.

| n = 30 | K-fold | Repeated | Nested |
|---|---|---|---|
| 60 : 40 | 0.250 | 0.250 | 0.250 |
| 70 : 30 | 0.444 | 0.444 | 0.444 |
| 80 : 20 | 0.833 | 0.833 | 0.833 |
| 90 : 10 | 1 | 1 | 1 |

The logistic learning result shows that the performance of the CVs increases as the split ratio increases (At split ratio 60:40 the performance was 0.250. It was 0.444 at 70:30, 0.833 at 80:20 and 1.000 at 90:10. All the CVs have equal performances across the split ratio and at optimum in [90:10] splits. See the pictorial illustration of Figure 4 based on Table 4 above.

**When n = 40**, the CVs performance remains the same (0.500) at split ratios [60:40] and [90:10]. At [70:30] and [80:20], both *k-fold* and *nested CV* outperform the *repeated CV* under the Naïve learning algorithm.

The result of the logistic learning algorithm shows that *k-fold CV* outperforms both *repeated* and *nested CV* in [60:40], while k-fold performance was 0.563 both Nested and Repeated CV was 0.5000 each. Both *k-fold* and *repeated CV* outperform the *nested CV*. The performance remains the same (0.500for all CVs) at [80:20] and [90:10] split, respectively ( See Figures 5 and Table 5; Figure 6 and Table 6).

**When n = 50**, the CVs performance remains the same at all the split ratios (0.600 at 60:40,0.533 at 70:30, 6.000 at 70:30 ) and optimum, 8.000 at [90:10]. Under the Naïve learning algorithm. We obtain a similar summary result in Logistic learning output. See Figures 7and 8; Table 7 and 8 for details.

**When n = 100**, the CVs performance remains the same (0.475) at [60:40] split ratio, 0.450 at [80:20] split ratio, and 4.000 at [90:10] split ratio but varies at [70:30], under the Naïve learning algorithm. The *repeated CV* outperforms both the *k-fold* and *nested CV* at [70:30] where its performance was 0.567 and other are 0.500 each. The CVs performed equally under the Logistic learning output, and optimal at [60:40] and [80:20]. See Figures 9 and 10; Tables 9 and 10 for details.

**When n = 200**, the CVs performance remains the same, 0.488 at [60:40], 0.467 at[70:30], and 0.5000 at [90:10] split ratio. But it varies at [80:20]. Repeated CV performance was 0.500 while others were 0.450 each. Under the Naïve learning algorithm. The *repeated CV* outperforms both the *k-fold* and *nested CV* at [80:20], while Repeated CV performance was 0.500, k-fold and nested remained 0.450 each. All CVs performances were optimal, 5.000 at [90:10] split ratio. The CVs performed equally under the Logistic learning output, and optimal at [70:30], where the performance is 0.5000. See Figures 11 and 12; Tables 11 and 12 for details.

**When n = 500**, the CVs performance remains the same at all the split ratio and optimal at [70:30] where the performance is 0.500. Under the Naïve learning algorithm. Both *repeated CV* and *nested CV* outperform the *k-fold* at [70:30] and *k-fold CV* outperforms both *repeated CV* and *nested CV* at [80:20]. All CVs performances were optimal in [90:10] under the Logistic learning output. See Figures 13 and 14; Tables 13 and 14 for details .

**Naïve Bayes Result at n = 40**

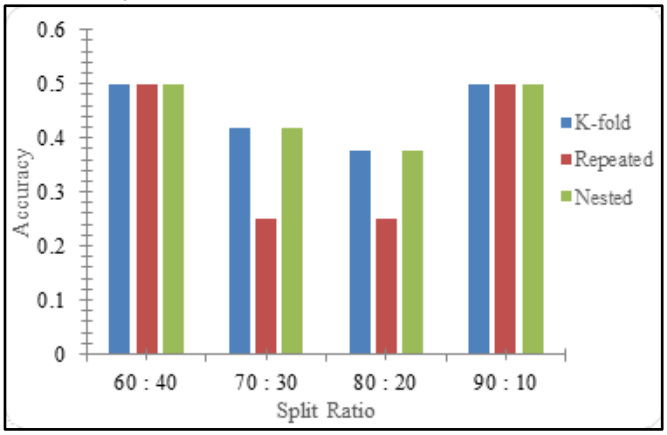


Figure 5: CVs performance for Naïve Bayes when n = 40.

Table 5: CVs performance for Naïve Bayes when n = 40.

| n = 40 | K-fold | Repeated | Nested |
|---|---|---|---|
| 60 : 40 | 0.500 | 0.500 | 0.500 |
| 70 : 30 | 0.417 | 0.250 | 0.417 |
| 80 : 20 | 0.375 | 0.250 | 0.375 |
| 90 : 10 | 0.500 | 0.500 | 0.500 |

**Logistic Result at n = 40**

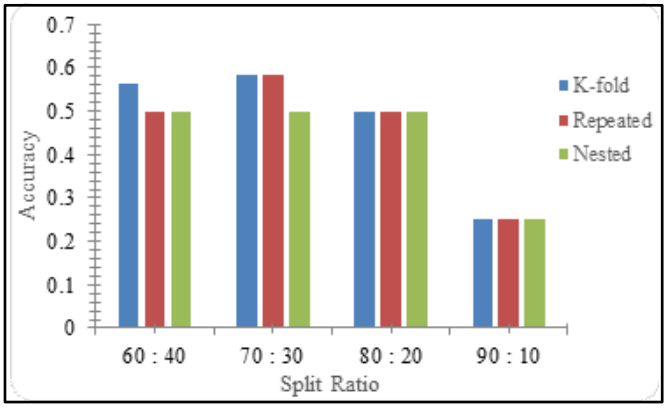


Figure 6: CVs performance for Logistic when n = 40.

Table 6: CVs performance for Logistic when n = 40.

| n = 40 | K-fold | Repeated | Nested |
|---|---|---|---|
| 60 : 40 | 0.563 | 0.500 | 0.500 |
| 70 : 30 | 0.583 | 0.583 | 0.500 |
| 80 : 20 | 0.500 | 0.500 | 0.500 |
| 90 : 10 | 0.250 | 0.250 | 0.250 |

**Naïve Bayes Result at n = 50**

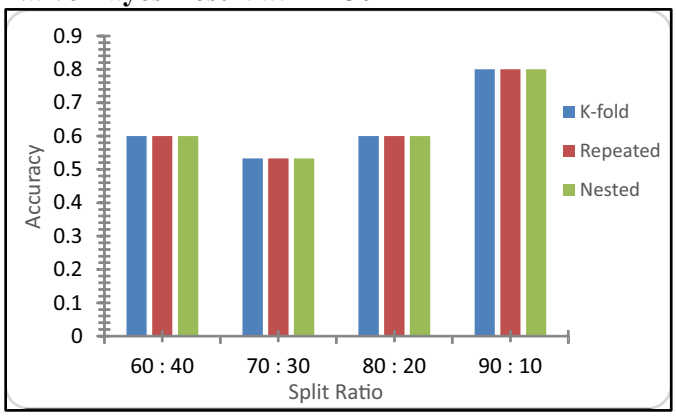


Figure 7: CVs performance for Naïve Bayes when n = 50.

Table 7: CVs performance for Naïve Bayes when n = 50.

| n = 50 | K-fold | Repeated | Nested |
|---|---|---|---|
| 60 : 40 | 0.600 | 0.600 | 0.600 |
| 70 : 30 | 0.533 | 0.533 | 0.533 |
| 80 : 20 | 0.600 | 0.600 | 0.600 |
| 90 : 10 | 0.800 | 0.800 | 0.800 |

**Logistic Result at n = 50**

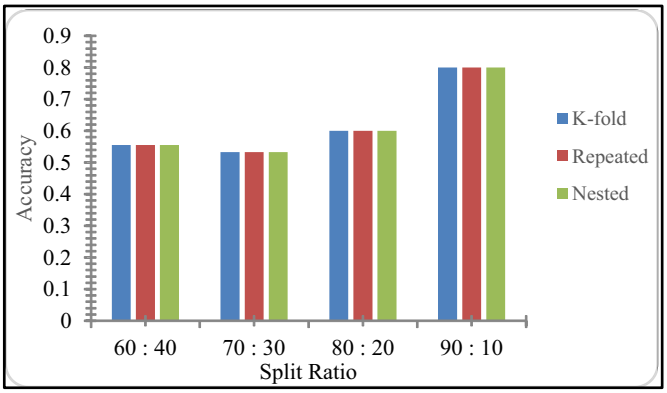


Figure 8: CVs performance for Logistic when n = 50.

Table 8: CVs performance for Logistic when n = 50.

| n = 50 | K-fold | Repeated | Nested |
|---|---|---|---|
| 60 : 40 | 0.555 | 0.555 | 0.555 |
| 70 : 30 | 0.533 | 0.533 | 0.533 |
| 80 : 20 | 0.600 | 0.600 | 0.600 |
| 90 : 10 | 0.800 | 0.800 | 0.800 |

**Naïve Bayes Result at n = 100**

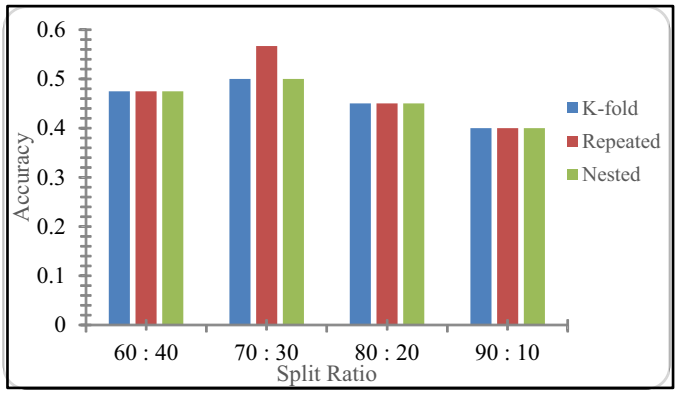


Figure 9: CVs performance for Naïve Bayes when n = 100.

Table 9: CVs performance for Naïve Bayes when n = 100.

| n = 100 | K-fold | Repeated | Nested |
|---|---|---|---|
| 60 : 40 | 0.475 | 0.475 | 0.475 |
| 70 : 30 | 0.500 | 0.567 | 0.500 |
| 80 : 20 | 0.450 | 0.450 | 0.450 |
| 90 : 10 | 0. 400 | 0. 400 | 0. 400 |

**Logistic Result at n = 100**

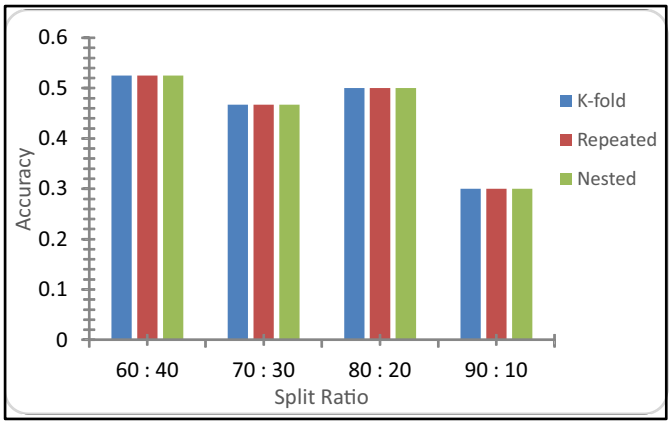


Figure 10: CVs performance for Logistic when n = 100.

Table 10: CVs performance for Logistic when n = 100.

| n = 100 | K-fold | Repeated | Nested |
|---|---|---|---|
| 60 : 40 | 0.525 | 0.525 | 0.525 |
| 70 : 30 | 0.467 | 0.467 | 0.467 |
| 80 : 20 | 0.500 | 0.500 | 0.500 |
| 90 : 10 | 0.300 | 0.300 | 0.300 |

**Naïve Bayes Result at n = 200**

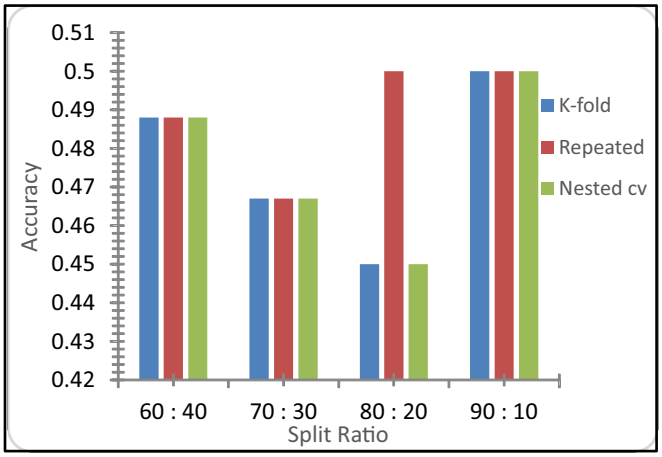


Figure 11: CVs performance for Naïve Bayes when n = 200.

Table 11: CVs performance for Naïve Bayes when n = 200.

| n = 200 | K-fold | Repeated | Nested |
|---|---|---|---|
| 60 : 40 | 0.488 | 0.488 | 0.488 |
| 70 : 30 | 0.467 | 0.467 | 0.467 |
| 80 : 20 | 0.450 | 0.500 | 0.450 |
| 90 : 10 | 0.500 | 0.500 | 0.500 |

**Logistic Result at n = 200**

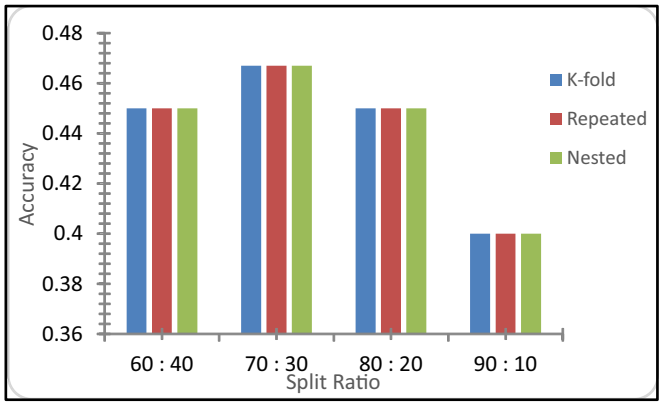


Figure 12: CVs performance for Logistic when n = 200.

Table 12: CVs performance for Logistic when n = 200.

| n = 200 | K-fold | Repeated | Nested |
|---|---|---|---|
| 60 : 40 | 0.450 | 0.450 | 0.450 |
| 70 : 30 | 0.467 | 0.467 | 0.467 |
| 80 : 20 | 0.450 | 0.450 | 0.450 |
| 90 : 10 | 0.400 | 0.400 | 0.400 |

**Naïve Bayes Result at n = 500**

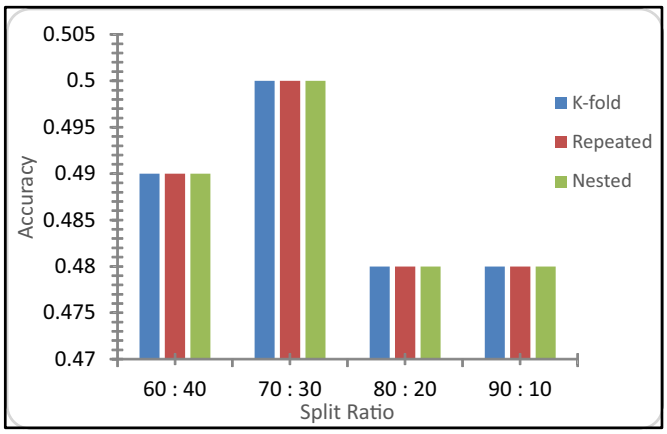


Figure 13: CVs performance for Naïve Bayes when n = 500.

Table 13: CVs performance for Naïve Bayes when n = 500.

| n = 500 | K-fold | Repeated | Nested |
|---|---|---|---|
| 60 : 40 | 0.490 | 0.490 | 0.490 |
| 70 : 30 | 0.500 | 0.500 | 0.500 |
| 80 : 20 | 0.480 | 0.480 | 0.480 |
| 90 : 10 | 0.480 | 0.480 | 0.480 |

**Logistic Result at n = 500**

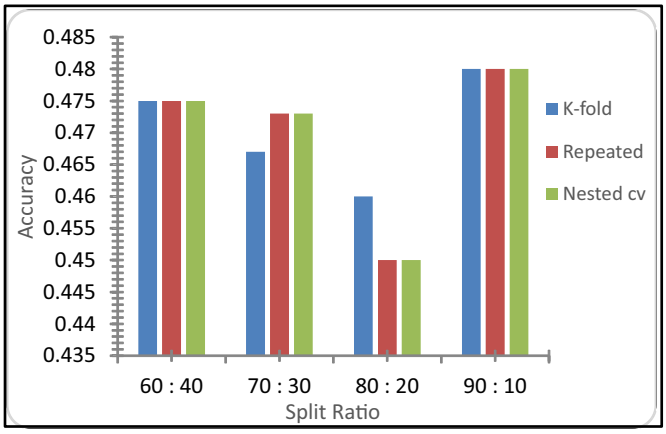


Figure14: CVs performance for Logistics when n = 500.

Table 14: CVs performance for Logistic when n = 500.

| n = 500 | K-fold | Repeated | Nested |
|---|---|---|---|
| 60 : 40 | 0.475 | 0.475 | 0.475 |
| 70 : 30 | 0.467 | 0.473 | 0.473 |
| 80 : 20 | 0.460 | 0.450 | 0.450 |
| 90 : 10 | 0.480 | 0.480 | 0.480 |

*Report of findings: Case 2 → Large Training sizes*

Having considered the small case scenario [training sizes 10, 20, 30, 40, 50, 100, 200, and 500] above, we extend the simulation to cover training sizes [1000, 2000, 5000, 10000, 50000]. Each of the scenarios in the split ratio are also considered.

**When n = 1000**, the CVs performance remains the same at [60:40], [70:30] and [80:20] split ratios and all the CVs are optimum at [60:40]. Meanwhile, the *repeated CV* outperforms both *k-fold CV* and *nested CV* at [90:10] split ratio, all these happen under the Naïve learning algorithm.

Under the Logistic learning output, all CVs perform equally across all the split ratios. Though performances vary based on split by-split ratio uniqueness. The point is, their individual performances at any given split ratio are the same. More so, all the CVs were optimal in [70:30]. See Figures 15 and 16; Tables 15 and 16 for details.

**When n = 2000**, the CVs performance is the same at [70:30 and [90:10] split ratios and varies at [60:40] and [80:20]. Meanwhile, the *repeated CV* outperforms both *k-fold CV* and *nested CV* at [80:20], all these happen under the Naïve learning algorithm.

Under the Logistic learning output, the CVs perform equally at [60:40] and [80:20] split ratios. Though performances vary across split by split ratio. *Nested CV* outperforms both *k-fold* and *repeated CV* at [70:30] split ratio and both *k-fold* and *nested CV* outperform the *repeated CV* at [90:10] split and all the CVs were optimal at [90:10]. See Table 10 for detailed information and pictorial results.

**When n = 5000**, the CVs performance varies at [60:40], [70:30] and [90:10] split ratios and is the same at [80:20]. Thus, the *k-fold CV* outperforms both the *repeated CV*and *nested CV* at [60:40]. Similarly, *repeated CV* outperforms *k-fold CV* and *k-fold CV* outperforms *nested CV,* in [70:30] split ratio. Both *repeated CV* and *k-fold CV* outperform the *nested CV* at [90:10] split ratio. All these happen under the Naïve learning algorithm.

## Performance of the CVS when n = 1000

### Naïve Bayes Result at n = 1000

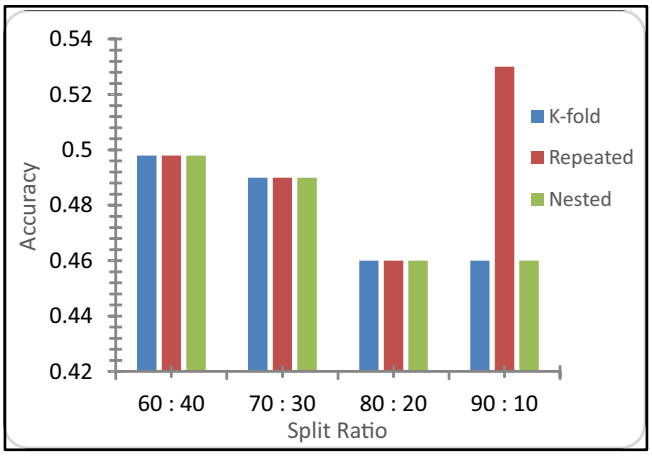


Figure 15: CVs performance for Naïve Ba yes when n = 1000.

Table 15: CVs performance for Naïve Bayes when n = 1000.

| n = 1000 | K-fold | Repeated | Nested |
|---|---|---|---|
| 60 : 40 | 0.498 | 0.498 | 0.498 |
| 70 : 30 | 0.490 | 0.490 | 0.490 |
| 80 : 20 | 0.460 | 0.460 | 0.460 |
| 90 : 10 | 0.460 | 0.530 | 0.460 |

### Logistic Result at n = 1000

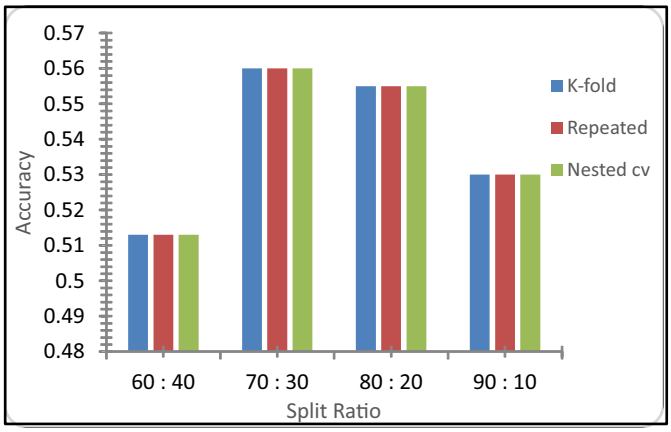


Figure 16: CVs performance for Logistic when n = 1000.

Table 16: CVs performance for Logistic when n = 1000.

| n = 1000 | K-fold | Repeated | Nested |
|---|---|---|---|
| 60 : 40 | 0.513 | 0.513 | 0.513 |
| 70 : 30 | 0.560 | 0.560 | 0.560 |
| 80 : 20 | 0.555 | 0.555 | 0.555 |
| 90 : 10 | 0.530 | 0.530 | 0.530 |

## Conclusion

So far, we have been able to address all the research questions raised in this study. First, there is no particular weakness among the CV methods considered with respect to the performances at various n-training samples and various learning algorithms considered. While, a particular CV approach may outperform at some points, the other may overtake it in another scenario. Essentially, there are no significant observable differences in the CVs performances across the training sizes. Some differences are observed as reported below.

The work observed that variation in the performances of learning algorithms may be subject to data. Consequently, any splitting ratio may be adopted as perceived by the researcher. More so, the training sizes (n) may be a source of motivation for choosing a splitting percentage for any particular research.

Our study further establishes that; various training sizes (n) may not significantly influence the performances of the CVs methods; particularly in the content of the comparing non-exhaustive CV subsets as in our study. CV methods show some promising variations. This prompted us to highlight the point of variations in respect of the research questions raised earlier.

1. Is there any significant difference in the performances of the *CVs* methods across various training samples?

**Answer:** Yes, at some points, however such differences or variations in performances are inconsistent across CV methods and training sizes If there are any observable differences in the *CVs* performances across the training samples, then:

2. At what n-size is *k-fold CV* performs better than *repeated CV* and *nested CV*? This shall be evaluated on *repeated CV* versus nested CV and nested CV versus repeated CV and *k-fold.*

3. At which split ratio is *k-fold CV* performing better than *repeated CV* and *nested CV*? This shall be evaluated in *repeated CV* and *nested CV*, and *nested CV* versus *repeated CV* and *k-folds CV*.

**Answer:**
*Under the Naïve Bayes Classifier*

a. At [60:40], we observe that *k-fold* at n = 5000 outperforms the *nested* and *repeated CV.* And *repeated CV* outperforms *k-fold and nested CV* at n = 10000.
b. At [70:30], *repeated CV* at *n* = 30 and 50 outperforms both the *k-fold* and *nested CV.* And *nested* and *k-fold* CV outperform *repeated CV* at n=40
c. At [80:20], *repeated CV* at n = 200 and 2000 outperform both the *k-fold* and *nested CV*. And nested and *k-fold CV* outperform *repeated* CV at n=40 and 10000
d. At [90:10], *repeated CV* at n=1000 and 10000 outperforms both the *k-fold CV* outperform *repeated CV* and *repeated CV* outperform *nested CV* at n=20 *k-fold CV* at n = 20.

Under the Logistics regressive learning

a. At [60:40], *k-fold CV* at n=40 outperform both the *repeated* and *nested CV*.
b. At [70:30], *repeated CV* at n=40 outperform both *k-fold* and *nested CV* at n-2000 and *k-fold* outperform the *nested* and *repeated CV* at n=500
c. At [80:20], that *k-fold* outperforms the *nested* and *repeated CV* at n=20 and 500
d. At [90:10], *nested* and *repeated CV* outperform *k-fold CV* at n=20. Both *k-fold* and *nested CV* outperform the *repeated CV* at n=2000 and 5000

1. In summary, which of the subset of the *k-fold* method performs best?

**Answer:** Repeated Learning Training CV
And if there is no significant difference in the *Cvs* performance across the training samples,

2. What is the general observation of the *CVs* performance across the training samples?

In conclusion, the performance of the CV methods varies. And have no particular trends or consistencies across samples and learning algorithms. Thus, we have established empirically that, the performance of *k-fold, repeated* and *nested CV* are not significantly different. And any of the methods can be adopted for any learning task. Essentially, if considering computational cost, *k-fold CV* would be the most favorable among other subsets considered.

## References

[1]. Winkler, W. E. 1999. The state of record linkage and current research problems. *In Statistics of income division, Internal Revenue publication* R99/04

[2]. Guyon, I.2015. An introduction of variable and feature selection. *Journal of Machine Learning Research* 3, 1157 - 1182.

[3]. Stone, M. 1974. Cross-validatory choice and assessment of statistical prediction. *Journal of the Royal Statistical Society: series B ( Methodological)* Vol. 36 No.2 pp 111-133

[4]. Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The elements of statistical learning: data mining, inference, and prediction.* 2nd edition. Springer, New York.

[5]. Varma, S. and Simon, R. 2006. Bias in error estimation when using cross-validation for model selection. *BMC bioinformatics*. 7(1), 91.

*Journal of Science Research* ISSN 1119 7333

**Citation:** Adedeji, O. F. and Olubusoye, O. E. Predictive Performance of Cross-Validation Techniques in Classification Models. Volume 20, 2023, pp 177-186