Advances and Applications in Mathematical Sciences
Volume 18, Issue 11, September 2019, Pages 1383-1398

IMPACT OF DATA REDUCTION THROUGH SAMPLING TECHNIQUES IN DATA CLASSIFICATION

K. GANESHAMOORTHI[1], J. ARUNADEVI[2] and N. ILAKKIYA[3]

[1]Ph.D Scholar (Full-Time)
Department of Computer Science
Raja Doraisingam Govt. Arts College
Sivaganga, Affilliated to Alagappa University

[2]Assistant Professor
Department of Computer Science
Raja Doraisingam Govt. Arts College
Sivaganga, Affilliated to Alagappa University

[3]M.Phil. Scholar
Department of Computer Science
Raja Doraisingam Govt. Arts College
Sivaganga, Affiliated to Alagappa University

Abstract

Aim: The aim of this paper is to employ sampling techniques for the purpose of data reduction and to analyze the results obtained by the classifiers, when applied to the dataset.

Background: Data is abundant and it is to be processed for getting the information. Data preprocessing is important task to perform for getting things easy. Data reduction is the technique through which we can employ a small, subset of the original dataset to obtain approximately the same results. Sampling is the technique used for avoiding the time consumption and to reduce cost of processing the entire dataset.

Methodology: For this study the researchers employed three sampling techniques. They are random sampling, stratified sampling and bootstrapping sampling methods. These methods are tested against three synthetic datasets which were used for the data mining tasks. After the sampling procedures applied to the datasets, classification task is carried out. The sample data set is tested against three types of classifiers. The parameters used for analysis is the accuracy, error rate, kappa index, weighted mean precision and weighted mean recall.

2010 Mathematics Subject Classification: 62H30.

Keywords: data reduction, random sampling, bootstrap sampling, stratified sampling, decision tree, Naïve base, *K*-nearest neighbor.

Received July 15, 2019; Accepted September 25, 2019

Contribution: This paper identifies sampling as the important data reduction strategy. It is evident through the results obtained. This awareness is vital for the analyst and scientist working with the data day to day. We can utilize this model for the data reduction to avoid the time consumption, reduce the cost and speedy decision making. This paper clearly depicts the impact of the data reduction strategies over the classification which is an important task in the data mining.

1. Introduction

Data are abundant today because of the improvement in the data procurement and the increased applications that generate data. The problem with the abundant data is the enlarged complexity of the data processing to information. So we have to concentrate on the data reduction methods which would reduce the complexity of the data processing to achieve meaningful information. The proper usage of data preprocessing reduces the complexity and increase the performance of any machine learning algorithm.

In this paper we propose to use sampling techniques for the data reduction process and test it for data classification algorithms. Sampling means taking a part of the given data and trying to analyze the data instead of taking the whole data for processing. The statistical inference could be obtained and used for further processing [1]. The sampling technique is used as the data reduction method and the effect of this data reduction is tested using the data classification algorithms.

2. Background Study

The background study of this paper concentrates on the following topics such as data reduction, Data sampling and classifiers.

2.1. Data Reduction

Data reduction is the preprocessing step in which the amount of data to be processed will be reduced. The data reduction can be done using dimensionality reduction, feature selection, data compression, numerosity reduction etc [2].

2.2. Data Sampling

Data sampling is the technique which is used to choose some representative data from the whole dataset. This sampling can be divided

into probability sampling and non probability sampling. Sampling is one of the promising methods for the data reduction [3]. Many sampling techniques have emerged and this could be utilized for the data science research [4].

2.3. Data Classification

Classification is the supervised machine learning algorithm, which is used to group the data items based on the label provided for the instance. This could be carried out in two phases. The first phase is the training phase and the second phase is the testing phase. In the training phase the algorithm is trained by the data instances which is grouped based on the label. In the testing phase the algorithm is tested to identify the data instances on the label [5].

3. Problem Statement

The problem discussed in this paper is the data reduction. Due to the enormous amount of data generated every day, we are in the need to employ techniques which would reduce the data processing to avoid the growing complexity. So we have to reduce the data that should be processed. This data reduction could be employed for the improvement of the machine learning algorithms.

4. Proposed Methodology

The proposed methodology in the paper for the data reduction is the application of the sampling techniques. By the application of the sampling techniques we could reduce the amount of data that should be processed, thus the complexity of the problem could be reduced. But the point to be considered is by the reduction what happens to the quality of the data?. This is the open problem. So we try to apply this concept by testing the proposed approach in the machine learning algorithms. Here we have chosen classifiers to apply the proposed concept.

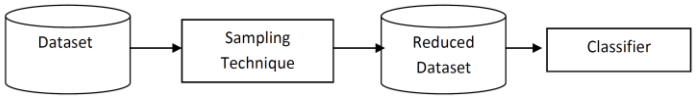


Figure 1. Workflow of the proposed methodology.

5. Experimental Setup

The experiment is carried out with three datasets, three sampling techniques are applied to this datasets, and three classifiers are tested for this research.

Table 1. Dataset description.

Name of the Dataset	No. of attributes	No. of examples
Deals	4	1000
Ripley Set	3	250
Sonar	61	208

Table 2. Details of the experiments conducted.

Dataset used	Sampling method used	Classification algorithm used
Deals, Ripley Set & Sonar	Sampling	Decision Tree
		K Nearest Neighbors
		Naive Bayes
Deals, Ripley Set & Sonar	Stratified sampling	Decision Tree
		K Nearest Neighbors
		Naive Bayes
Deals, Ripley Set & Sonar	Bootstrapping Sampling	Decision Tree
		K Nearest Neighbors
		Naive Bayes

5.1. Performance Metrics used

The performance metrics used for the experiment is given below

5.1.1. Accuracy:

Accuracy is how close a measured value is to the true value. It expresses the correctness of a measurement and determined by absolute and comparative way.

$$\text{Accuracy} = \frac{\text{Sum of true positives} + \text{Sum of true negatives}}{\text{Total population}}.$$

5.1.2. Classification Error

Relative number of misclassified examples or in other words percentage of incorrect predictions.

$$\text{Classification Error} = \frac{\text{Sum of false positives} + \text{Sum of false negatives}}{\text{Total population}}.$$

5.1.3. Kappa

The Kappa statistic (or value) is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance).

$$\text{Kappa} = \frac{\text{Accuracy} - \text{Random Accuracy}}{1 - \text{Random Accuracy}}.$$

Where accuracy is simply the sum of true positive and true negatives, divided by the total number of items

$$\text{Accuracy} = \frac{\text{Sum of true positives} + \text{Sum of true negatives}}{\text{Total population}}.$$

Random Accuracy is defined as the sum of the products of reference likelihood and result likelihood for each class. That is

$$\text{Random Accuracy} = \frac{a + b + c + d}{(\text{Total population})^2}.$$

Where a = sum of true negatives + sum of false positive

b = sum of true negatives + sum of false negative

c = sum of false negatives + sum of true positive

d = sum of false positive + sum of true positive.

5.1.4. Weighted Mean Recall

The weighted mean of all per class recall measurements. It is calculated through class recalls for individual classes.

$$\text{Recall} = \frac{\text{Sum of true positives}}{\text{Sum of true positives} + \text{Sum of false negatives}}.$$

5.1.5. Weighted Mean Precision

The weighted mean of all per class precision measurements. It is calculated through class precisions for individual classes

$$\text{Precision} = \frac{\text{Sum of true positives}}{\text{Sum of true positives} + \text{Sum of false positives}}.$$

6. Experimental Results

The following tables gives the details of the consolidated results obtained by the experiments conducted, which is discussed above.

Table 3. Results obtained when classification without sampling applied on Deals dataset.

	Accuracy	Error	Kappa	WM recall	WM precision
DT	99.6	0.4	0.992	99.62	99.59
KNN	97.3	27	0.946	97.4	97.33
Naïve	92.6	7.4	0.852	92.64	92.64

Table 4. Results obtained when classification without sampling applied on Ripley set dataset.

	Accuracy	Error	Kappa	WM recall	WM precision
DT	83.6	16.4	0.671	83.46	84.42
KNN	86.8	13.2	0.735	86.7	87.22
Naïve	84.4	15.6	0.688	84.39	85.08

Table 5. Results obtained when classification without sampling applied on Sonar dataset.

	Accuracy	Error	Kappa	WM recall	WM precision
DT	62.12	37.88	0.231	61.45	62.61
KNN	66.9	33.1	0.348	67.76	69.79
Naïve	82.14	17.86	0.639	81.68	84.21

Table 6. Results obtained when classification with sampling applied on Deals dataset.

	Accuracy	Error	Kappa	WM recall	WM precision
DT	98	2	0.96	98.06	97.96
KNN	80.9	19.1	0.618	81	80.91
Naïve	97.4	2.6	0.948	97.27	97.59

Table 7. Results obtained when classification with sampling applied on Deals dataset.

	Accuracy	Error	Kappa	WM recall	WM precision
DT	94.1	5.9	0.882	94.39	94.42
KNN	84.5	15.5	0.69	84.56	84.47
Naïve	97.4	2.6	0.948	97.25	97.65

Table 8. Results obtained when classification with Stratified sampling applied on Deals dataset.

	Accuracy	Error	Kappa	WM recall	WM precision
DT	99.7	0.3	0.994	99.72	99.68
KNN	98.7	1.3	0.974	98.77	98.66
Naïve	92.7	7.3	0.854	92.69	92.67

Table 9. Results obtained when classification with Bootstrap sampling applied on Deals dataset.

	Accuracy	Error	Kappa	WM recall	WM precision
DT	99.7	0.3	0.994	99.72	99.68
KNN	98.7	1.3	0.974	98.77	98.66
Naïve	92.7	7.3	0.854	92.69	92.67

Table 10. Results obtained when classification with sampling applied on Ripley set dataset.

	Accuracy	Error	Kappa	WM recall	WM precision
DT	88.4	11.6	0.768	88.4	88.96
KNN	91.2	8.8	0.824	91.2	91.58
Naïve	84.8	15.2	0.696	84.8	84.8

Table 11. Results obtained when classification with Stratified sampling applied on Ripley Set dataset.

	Accuracy	Error	Kappa	WM recall	WM precision
DT	87.6	12.4	0.752	87.6	87.8
KNN	84.4	15.6	0.688	84.4	84.58
Naïve	92	8	0.84	92	92.04

Table 12. Results obtained when classification with Boot strapping sampling applied on Ripley Set dataset.

	Accuracy	Error	Kappa	WM recall	WM precision
DT	92.8	7.2	0.856	92.8	92.81
KNN	94.4	5.6	0.888	94.4	94.69
Naïve	84.4	15.6	0.688	84.4	84.42

Table 13. Results obtained when classification with sampling applied on Sonar dataset.

	Accuracy	Error	Kappa	WM recall	WM precision
DT	77.88	22.12	0.551	77.26	78.54
KNN	90.87	9.13	0.816	90.66	91.02
Naïve	68.27	31.73	0.376	69.23	70.53

Table 14. Results obtained when classification with Stratified sampling applied on Sonar dataset.

	Accuracy	Error	Kappa	WM recall	WM precision
DT	67.79	32.21	0.34	66.7	68.89
KNN	88.94	11.06	0.776	88.47	89.77
Naïve	69.23	30.77	0.392	69.94	70.53

Table 15. Results obtained when classification with Boot strapping sampling applied on Sonar dataset.

	Accuracy	Error	Kappa	WM recall	WM precision
DT	79.81	20.19	0.34	79.07	81.03
KNN	95.67	4.33	0.776	95.56	95.77
Naïve	66.83	33.17	0.392	67.88	69.41

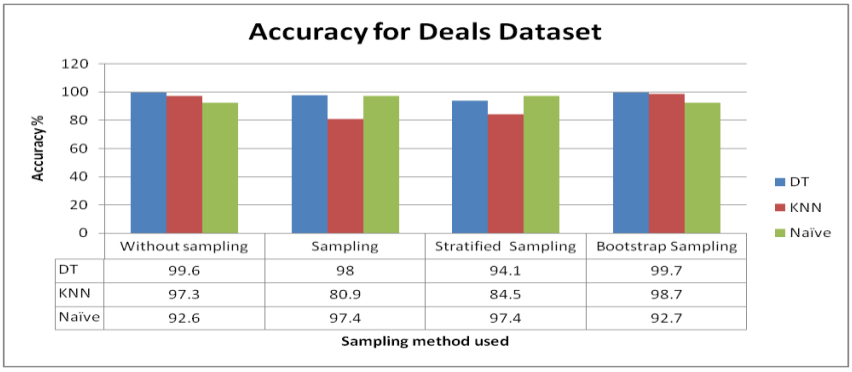
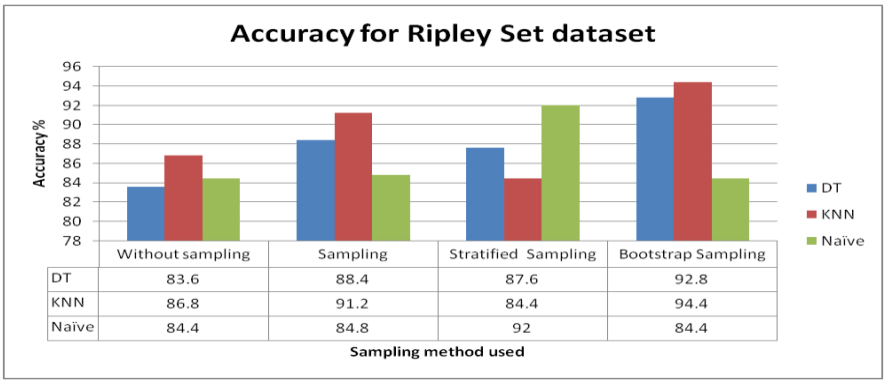


Figure 2. Comparison based on Accuracy for Deals dataset.

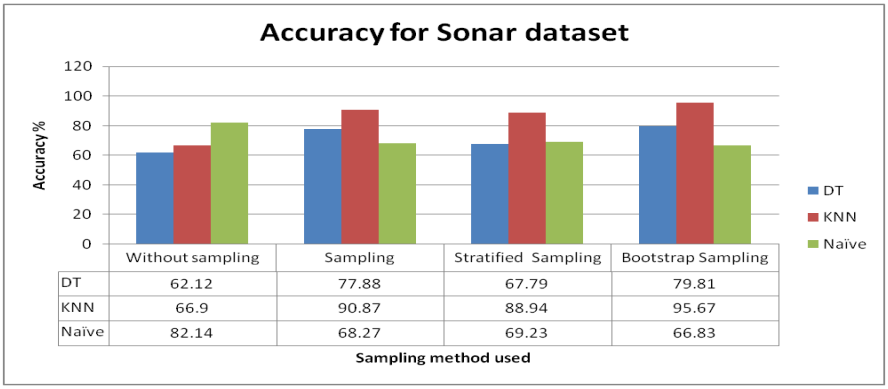
Figure 2 displays the comparison of various supervised learning methods when it is applied to the deals dataset. In terms of accuracy The Bootstrap sampling outperforms other methods when it is combined with the decision tree classifier.



	Without sampling	Sampling	Stratified Sampling	Bootstrap Sampling
DT	83.6	88.4	87.6	92.8
KNN	86.8	91.2	84.4	94.4
Naïve	84.4	84.8	92	84.4

Figure 3. Comparison based on Accuracy for Ripley set dataset.

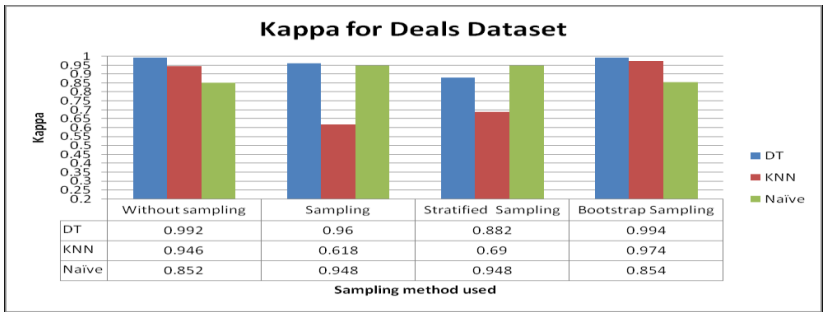
Figure 3 displays the comparison of various supervised learning methods when it is applied to the Ripley Set dataset. In terms of accuracy The Bootstrap sampling outperforms other methods when it is combined with the KNN classifier.



	Without sampling	Sampling	Stratified Sampling	Bootstrap Sampling
DT	62.12	77.88	67.79	79.81
KNN	66.9	90.87	88.94	95.67
Naïve	82.14	68.27	69.23	66.83

Figure 4. Comparison based on Accuracy for Sonar dataset.

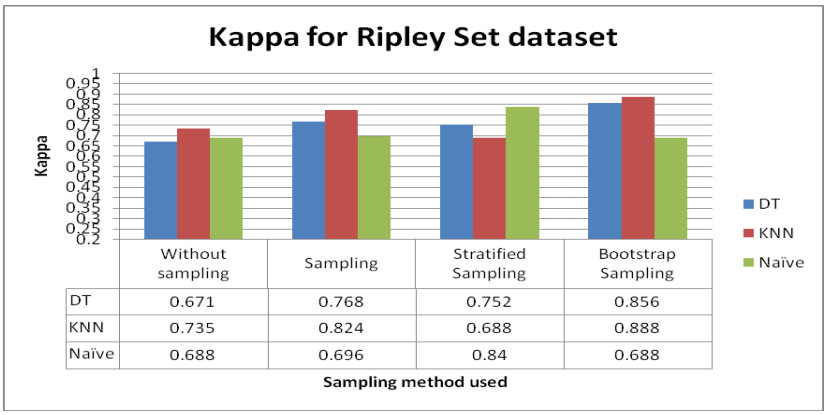
Figure 4 displays the comparison of various supervised learning methods when it is applied to the Sonar dataset. In terms of accuracy The Bootstrap sampling outperforms other methods when it is combined with the KNN classifier.



	Without sampling	Sampling	Stratified Sampling	Bootstrap Sampling
DT	0.992	0.96	0.882	0.994
KNN	0.946	0.618	0.69	0.974
Naïve	0.852	0.948	0.948	0.854

Figure 5. Comparison based on Kappa Statistics for Deals dataset.

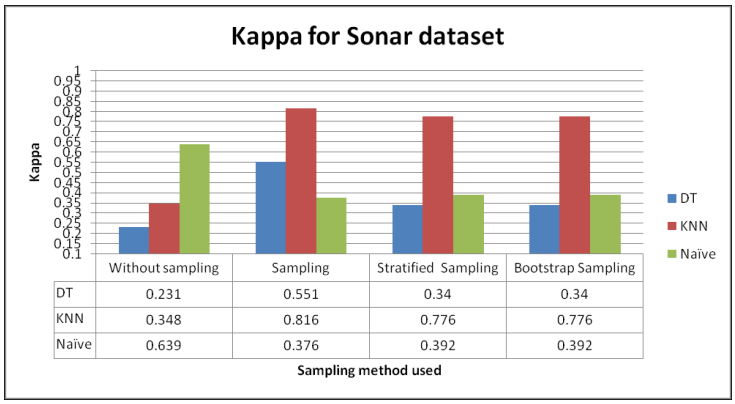
Figure 5 displays the comparison of various supervised learning methods when it is applied to the Deals dataset. In terms of Kappa without sampling with decision tree outperforms other all methods.



	Without sampling	Sampling	Stratified Sampling	Bootstrap Sampling
DT	0.671	0.768	0.752	0.856
KNN	0.735	0.824	0.688	0.888
Naïve	0.688	0.696	0.84	0.688

Figure 6. Comparison based on Kappa Statistics for Ripley set dataset.

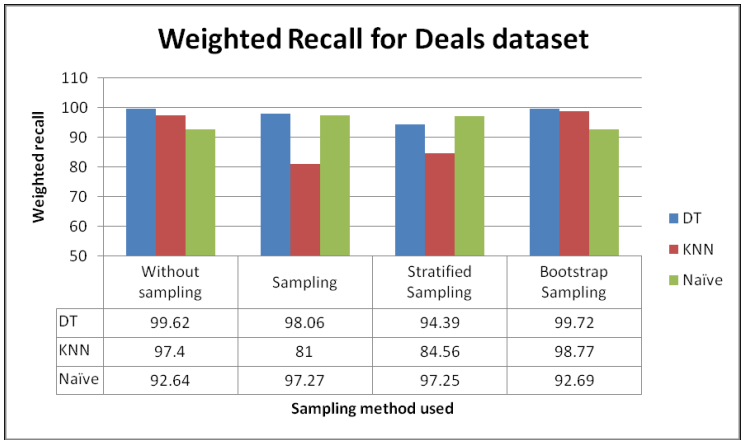
Figure 6 displays the comparison of various supervised learning methods when it is applied to the Ripley set dataset. In terms of Kappa Bootstrap sampling with KNN outperforms other all methods.



	Without sampling	Sampling	Stratified Sampling	Bootstrap Sampling
DT	0.231	0.551	0.34	0.34
KNN	0.348	0.816	0.776	0.776
Naïve	0.639	0.376	0.392	0.392

Figure 7. Comparison based on Kappa Statistics for Sonar dataset.

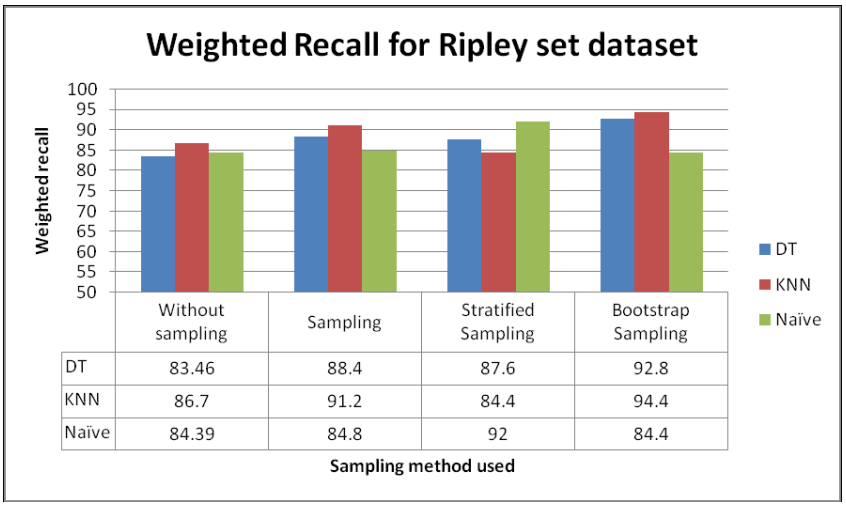
Figure 7 displays the comparison of various supervised learning methods when it is applied to the sonar dataset. In terms of Kappa sampling with KNN outperforms other all methods.



	Without sampling	Sampling	Stratified Sampling	Bootstrap Sampling
DT	99.62	98.06	94.39	99.72
KNN	97.4	81	84.56	98.77
Naïve	92.64	97.27	97.25	92.69

Figure 8. Comparison based on Weighted Recall for Deals dataset.

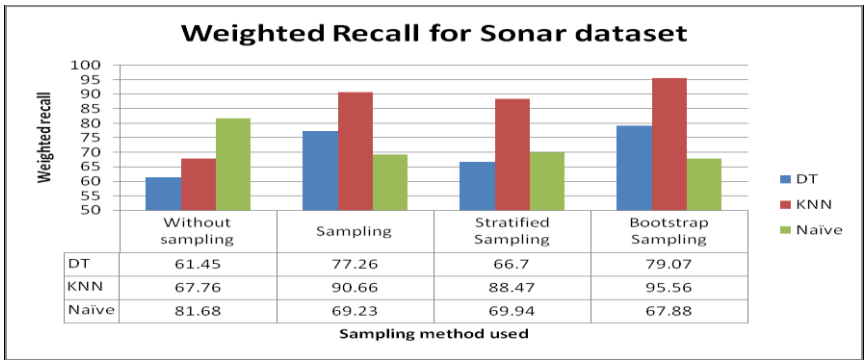
Figure 8 displays the comparison of various supervised learning methods when it is applied to the Deals dataset. In terms of Weighted Recall Bootstrap sampling with Decision Tree classifier outperforms other all methods.



	Without sampling	Sampling	Stratified Sampling	Bootstrap Sampling
DT	83.46	88.4	87.6	92.8
KNN	86.7	91.2	84.4	94.4
Naïve	84.39	84.8	92	84.4

Figure 9. Comparison based on Weighted Recall for Ripley Set dataset.

Figure 9 displays the comparison of various supervised learning methods when it is applied to the Ripley set dataset. In terms of Weighted Recall Bootstrap sampling with KNN classifier outperforms other all methods.



	Without sampling	Sampling	Stratified Sampling	Bootstrap Sampling
DT	61.45	77.26	66.7	79.07
KNN	67.76	90.66	88.47	95.56
Naïve	81.68	69.23	69.94	67.88

Figure 10. Comparison based on Weighted Recall for Sonar dataset.

Figure 10 displays the comparison of various supervised learning methods when it is applied to the Sonar dataset. In terms of Weighted Recall Bootstrap sampling with KNN classifier outperforms other all methods.

Figure 11 displays the comparison of various supervised learning methods when it is applied to the Deals dataset. In terms of Weighted Precision Bootstrap sampling with Decision Tree classifier outperforms other all methods.

Figure 12 displays the comparison of various supervised learning methods when it is applied to the Ripley set dataset. In terms of Weighted Precision Bootstrap sampling with KNN classifier outperforms other all methods.

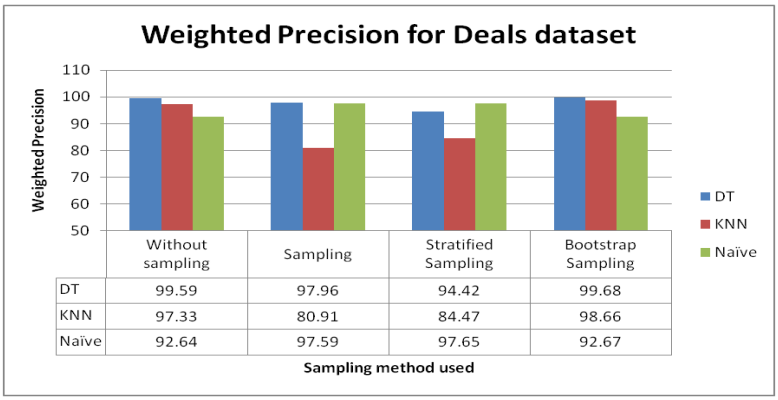


Figure 11. Comparison based on Weighted Precision for Deals dataset.

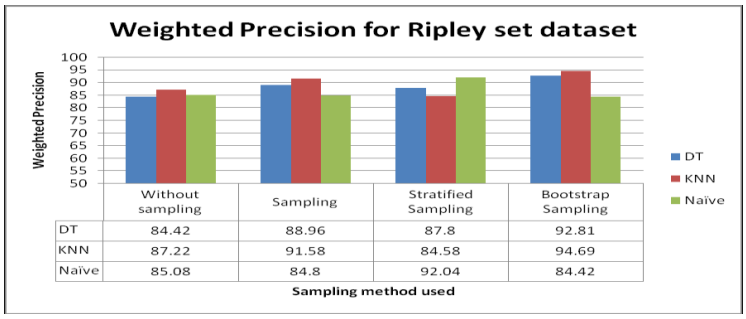


Figure 12. Comparison based on Weighted Precision for Ripley Set dataset.

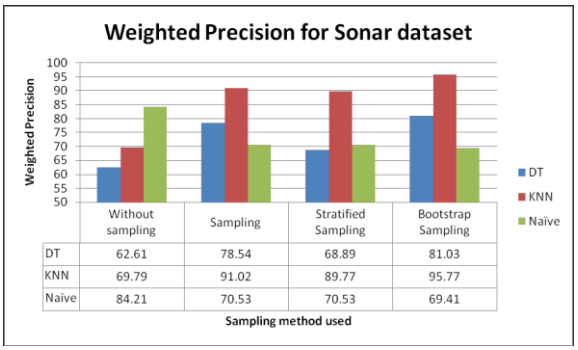


Figure 13. Comparison based on Weighted Precision for Sonar dataset.

Figure 13 displays the comparison of various supervised learning methods when it is applied to the Sonar dataset. In terms of Weighted Precision Bootstrap sampling with KNN classifier outperforms other all methods.

6.1. Comparison with Benchmark

The results obtained are discussed and the best results based on the various performance measures obtained from the classifiers are discussed and compared with the sampling facilitated classifiers are compared. The results are tabulated below. The analysis of the results shows that the application of sampling for the classification algorithms improves the classification performance measures.

Table 16. Comparison of the sampling facilitated classifier experiment with the bench marks results.

Dataset	Classifier	Accuracy	Error	Kappa	WM Recall	WM Precision
Deals	DT	99.6	0.4	0.992	99.62	99.59
	DT + Boot	99.7	0.3	0.994	99.72	99.68
Ripley Set	KNN	86.8	13.2	0.735	86.7	87.22
	KNN+ Boot	94.4	5.6	0.888	94.4	94.69
Sonar	Naïve	82.14	17.86	0.639	81.68	84.21
	KNN+ Boot	95.67	4.33	0.776	95.56	95.77

7. Conclusion

This paper analyses the results obtained from the simulation for the three classifiers algorithms over three sampling methods with three datasets. The results are discussed with the graphical tabulation of results obtained. The results are also compared with the benchmark results obtained from classification methods. The Bootstrapping sampling applied with the KNN classifier outperforms all other methods and classifiers for Ripley set and Sonar dataset. The Bootstrapping with Decision tree classifier applied for Deals dataset. This conclusion is limited to this study. The future work could

be extended by employing more sampling methods and testing in the real world datasets.

References

[1] Taherdoost, Hamed, Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research, International Journal of Academic Research in Management 5(2), 2016.

[2] Jiawei Han, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann Publishers, 2011.

[3] Noemí De Castro-García, Ángel Luis Muñoz Castañeda, David Escudero García, and Miguel V. Carriegos, Effect of the Sampling of a Dataset in the Hyperparameter Optimization Phase over the Efficiency of a Machine Learning Algorithm, Complexity, vol. 2019, pp. 1-16, 2019.

[4] S. L. Lohr, Sampling: Design and Analysis, Cengage. Learning, Boston, MA, USA, 2nd edition, 2009.

[5] Shekhar Pandey, M. Supriya and Abhilash Shrivastava, Data Classification Using Machine Learning Approach, The International Symposium on Intelligent Systems Technologies and Applications, 2018.