ISSN 2278-3091
Volume 9 No.2, March - April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse197922020.pdf
https://doi.org/10.30534/ijatcse/2020/197922020

Assessment on Tubalan Marine Sanctuary in Conserving Existing Coastal Habitat Using Data Mining Techniques

Orlando E. Ang[1*], Allemar Jhone P. Delima[2], Jan Carlo T. Arroyo[3]
[1, 2, 3]Professional Schools, University of Mindanao, Davao City, Davao del Sur, Philippines
[2]College of Engineering, Technology and Management, Cebu Technological University-Barili Campus, Philippines
[3]College of Computing Education, University of Mindanao, Davao City, Davao del Sur, Philippines
[1*]orlandoang@umindanao.edu.ph, [2]allemardelima@umindanao.edu.ph, [3]jancarlo_arroyo@umindanao.edu.ph

ABSTRACT

This paper predicts the accuracy of the marine coastal ecosystem health evaluation dataset based on the conducted assessment by the Department of Environment and Natural Resources (DENR). The study conducted is within a marine sanctuary in Malita, Davao Occidental, where fishing is the means of livelihood. The use of data mining techniques can help in the close monitoring of the reefs in the area. The marine coastal ecosystem health evaluation dataset used in this study consisted of seven variables, with 966 instances and were assessed using three data mining algorithms, namely the Naive Bayes, K-Nearest Neighbor (KNN), and C4.5 algorithms. The results of the study show that the Naïve Bayes, KNN, and C4.5 algorithms obtained 91.24%, 98.05%, and 96.86% prediction accuracies, respectively, where the identified optimal algorithm for prediction is the KNN algorithm. Finally, this study paved the way to determine indicators of a healthy marine ecosystem through mining targets or high economic value fishes.

Keywords: Data mining, Fish classification, Marine coastal ecosystem, Prediction

1. INTRODUCTION

The Tubalan marine sanctuary is located in the municipality of Malita, Davao Occidental. The residents' primary source of livelihood is agriculture and fisheries. It has been a challenge for the community to achieve a healthy marine ecosystem when faced with practicalities of harvesting species in multispecies marine communities.

In the Philippines, among other threats, widespread illegal and legal fishing is responsible for much of the population loss[1].Similarly, the market demand and poverty allow local fisher folks to make more catch without looking at the ecological problem, which may affect brought about by overfishing and eventually harm the environmental balance of the marine ecosystem[2].

Fish abundance in a given area is calculated in terms of total fish abundance or terms of abundance of key fish species or families. The number of fishes per unit area or when combined with size data is calculated in terms of fish biomass (total weight of fish per unit area). This fish biomass gives essential information about the overall trophic structure and reproduction of fishes in the reef, of which the Biodiversity Management Bureau conducted an assessment under the Department of Environment and Natural Resources through a Fish Visual Census (FVC).

The purpose of FVC is to quantify fish species' size and abundance in an area. It uses a long 50-meter transect line laid underwater down to the shallowest part possible. Fish observed within the estimated 5 meters "box" (belt-transect method) is counted according to abundance and classified according to the community composition of crucial fish species[3]. Data collected from this census can be processed using data mining techniques.

Data mining is a process for efficient extraction and classification of essential information within the dataset[4].This study used the famous data mining algorithms, namely Naïve Bayes, KNN, and C4.5 to evaluate and predict the accuracy of the dataset obtained by the DENR in one of the marine sanctuaries in Davao Occidental, Philippines. The result of this study will aid the department in decision and policy making undertakings in relation to preserving the coastal marine ecosystem in the region and would be added to the literature of data mining particularly the application of Naïve Bayes, KNN, and C4.5 algorithms in fish classification and prediction.

2. LITERATURE REVIEW

Fish stocks are traditionally arranged as common property resource that categorizes the variance, weight, and size of fish. It is done to control overfishing and to identify fish groups that are not potential for the commercial market or part as an indicator of contributing to the balance of the ecosystem in a marine sanctuary. Marine over exploitation and excessive fishing have been an issue that affects the marine ecosystem because of the devastation of corals and some fish indicators that help in maintaining the health of the reef[5].

Overfishing contributed an environmental effect, including the number of fish to catch, coral cover, algal cover coastal length, and biophysical activities [6].The effort in maintaining the coastal ecosystem's health is constant monitoring of the population and density of fish, the type of fish, biomass of the fish, and water environment, which can provide a basis for decision making of the stakeholders involved in local and market fishing.[5]-[7].

3. METHODOLOGY

3.1Datasets

In this study, the data collected from the conducted fish identification assessment are validated and clustered according to the size (cm), count, trophic group, family, species, name or variety, and the particular group, as shown inTable I below.

Table 1: Sample data from the fish assessment

Fish Size	Count	Trophic Group	Family	Species	Variety	Group
6	8	Herbivore	Pomacentridae	moluccensis	DamselFishes	Major
12	2	Omnivore	Chaetodontidae	octofasciatus	ButterflyFishes	Indicator
15	1	Benthic Carnivore	Labridae	mesothorax	Wrasses	Major
11	1	Benthic Carnivore	Labridae	celebicus	Wrasses	Target

The variable fish size from Table 1 determines the length (cm) and maturity of fish. The variable Count determines the population or density and abundance in an area. Trophic Group variable determines the sustenance of the fish to wit: herbivore, planktivore, carnivore, and omnivore. The variable family refers to the scientific classification family of fish and so the variables species and variety. The variable group belongs to specific categories to wit: (1) Major fish where species in this group has a low commercial value; (2) Target fish which refers to a group of fish with economic significance as they are caught for consumption and commerce; and (3) Indicator fish as fishes that tolerate a narrow range of environmental conditions and could provide information on the present health of the marine ecosystem.

The fish identification was already made based on findings from the initial research conducted by the DENR. To leverage the use of technology, data mining algorithms were used and were instrumental in checking the accuracy of the data. The Naïve Bayes, KNN, and C4.5 algorithms were simulated using the WEKA software application. When classifying the data using Naïve Bayes, simulation results revealed a 91.24% prediction accuracy of the correctness of classified instances. Further, the KNNalgorithm revealeda high 98.05% prediction accuracy, and a 96.86% prediction was obtained using the C4.5 algorithm. The graphical representation of the correctly and incorrectly classified instances are shown in Figure 1.

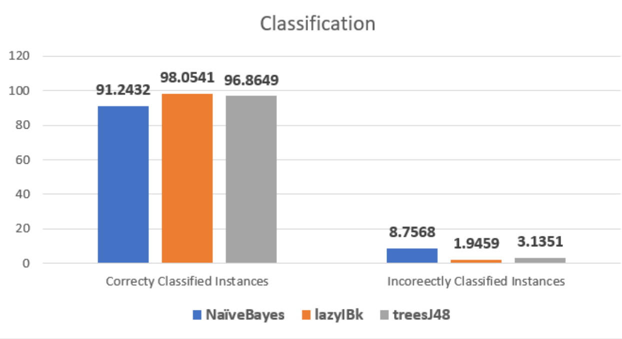
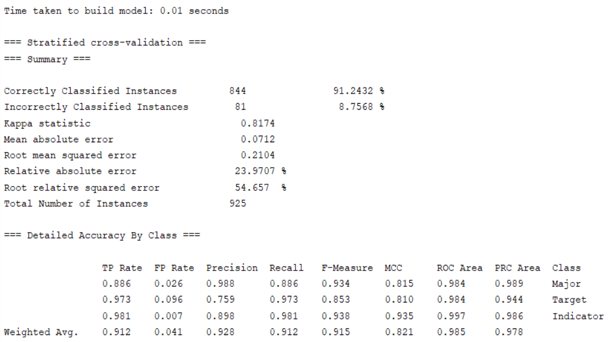


Figure 1: Indexed algorithms and percentage accuracies

3.2 Naïve Bayes (NB) Algorithm

Naive Bayes Theorem is a formula that calculates a probability by counting the frequency of values and combinations of importance in the historical data. The Bayes theorem is shown in equation (1) below where P (c|x) is the posterior probability of class (target) with the given predictor (attribute), P(c) is the prior probability of a class, P (x|c) is the likelihood which is the probability of predictor in a given class, and P(x) is the prior probability of a predictor obtained from the study of [8].

$$P(c|x) = \frac{P(c|x)P(c)}{P(x)} \qquad (1)$$



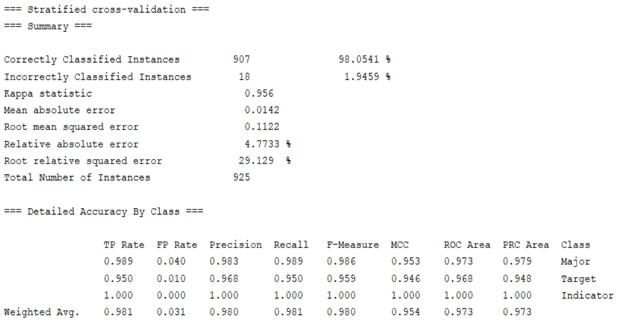
```
Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         844               91.2432 %
Incorrectly Classified Instances        81                8.7568 %
Kappa statistic                          0.8174
Mean absolute error                      0.0712
Root mean squared error                  0.2104
Relative absolute error                 23.9707 %
Root relative squared error             54.657  %
Total Number of Instances              925

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.886    0.026    0.988      0.886    0.934      0.815    0.984     0.989     Major
                 0.973    0.096    0.759      0.973    0.853      0.810    0.984     0.944     Target
                 0.981    0.007    0.898      0.981    0.938      0.935    0.997     0.986     Indicator
Weighted Avg.    0.912    0.041    0.928      0.912    0.915      0.821    0.985     0.978
```

Figure 2: Simulation result using the Naïve Bayes algorithm

The classification result on the simulated dataset using the Naïve Bayes algorithm is shown in Figure 2. The data obtained revealed a true positive rate of 0.912 correctly classifying 844 instances or 91.2432% accuracy.

3.3 KNN Algorithm

The KNN algorithm is another most commonly used methods for prediction and classification due to its simplicity and efficiency[9]. Figure 3 shows that a 907 correctly classified instances with 98.05% prediction accuracy were revealed using the KNN algorithm. It has the lowest incorrectly classified instances of 18 or equivalent to 1.9459%. The ROC level when KNN algorithm was used is 0.973 and is lower when compared to the two algorithms.

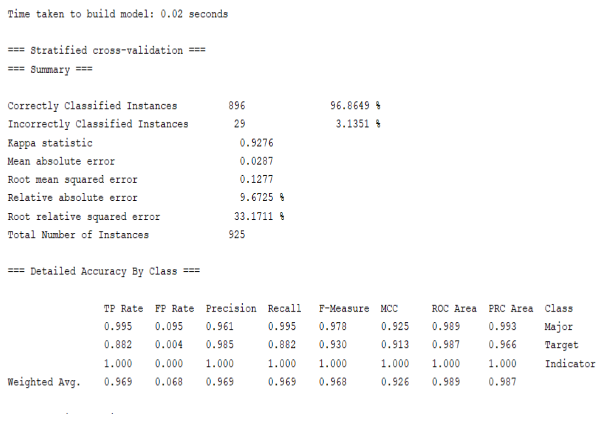


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         907               98.0541 %
Incorrectly Classified Instances        18                1.9459 %
Kappa statistic                          0.956
Mean absolute error                      0.0142
Root mean squared error                  0.1122
Relative absolute error                  4.7733 %
Root relative squared error             29.129  %
Total Number of Instances              925

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.989    0.040    0.983      0.989    0.986      0.953    0.973     0.979     Major
                 0.950    0.010    0.968      0.950    0.959      0.946    0.968     0.948     Target
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Indicator
Weighted Avg.    0.981    0.031    0.980      0.981    0.980      0.954    0.973     0.973
```

Figure 3: Simulation result using KNN algorithm

3.4 C4.5 Algorithm

The C4.5 algorithm which is similar to a node structure of a tree [10]was instrumental in classifying the same dataset revealing96.86% correctly classified instances and ROC level at 0.989, root relative squared error of 33.17%, and 29 incorrectly cases classified as shown in Figure 4.

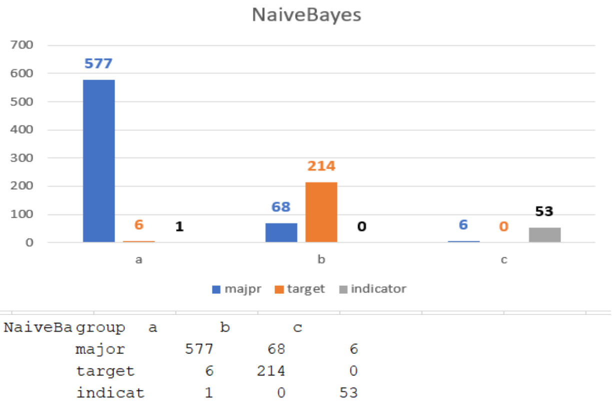


```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         896               96.8649 %
Incorrectly Classified Instances        29                3.1351 %
Kappa statistic                          0.9276
Mean absolute error                      0.0287
Root mean squared error                  0.1277
Relative absolute error                  9.6725 %
Root relative squared error             33.1711 %
Total Number of Instances              925

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.995    0.095    0.961      0.995    0.978      0.925    0.989     0.993     Major
                 0.882    0.004    0.985      0.882    0.930      0.913    0.987     0.966     Target
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Indicator
Weighted Avg.    0.969    0.068    0.969      0.969    0.968      0.926    0.989     0.987
```

Figure 4: Simulation Result Using C4.5

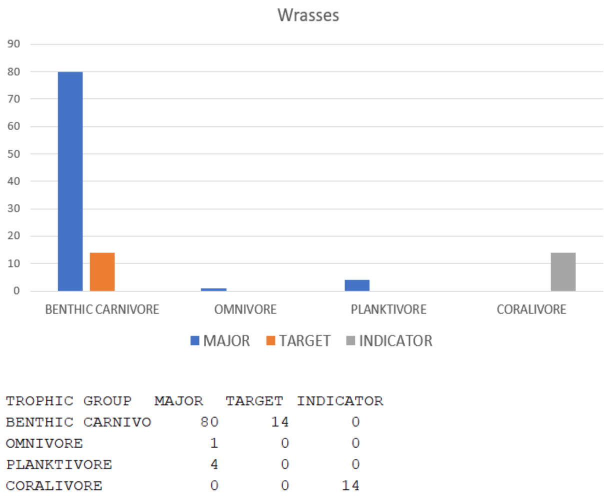
4. RESULT AND DISCUSSION

In the journal of Comparative Analysis of Bayes and KNN by [8],the Bayes Theorem determines the probability of an event occurring given the likelihood of another incident that has already happened with a high incorrectly classified instances of 8.7568% as compare to KNN and C4.5. The indexed simulation resultsare shown in Figure 5.Incorrectly classified instances denote that some of the fishes of the same variety or name were classified differently in the categorical group and trophic group, such that in the wrasses as shown in Figure 6.



NaiveBagroup	a	b	c
major	577	68	6
target	6	214	0
indicat	1	0	53

Figure 5: Naïve Bayes Result



TROPHIC GROUP	MAJOR	TARGET	INDICATOR
BENTHIC CARNIVO	80	14	0
OMNIVORE	1	0	0
PLANKTIVORE	4	0	0
CORALIVORE	0	0	14

Figure 6: Sample Fish Group (Wrasses)

The main advantage gained in employing the KNN method is that the target function will be approximated locally. Since the objective function is approximated locally for each query to the system, systems can concurrently solve multiple problems and deal successfully with changes in the problem arena hereto classifying correct instance of 98.9541% with less incorrectly classified instances of 1.9459%.

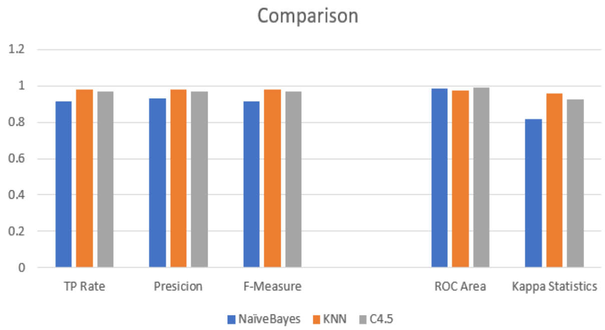


Figure 7: Complexity Comparison

Figure 7 shows that both the KNN and C4.5 algorithms' run-time complexity matches the C4.5 depth, which cannot be greater than the number of attributes, and has a higher rate of True Positive instances, as well as the combined measures of precision and recall, as compared to Naïve Bayes.

5. CONCLUSION

This study analyzes the strength of data mining, particularly the KNN, C4.5, and Naïve Bayes algorithms, to predict the accuracy and evaluate the dataset obtained in the research conducted by the DENR referring to the potentiality of Tubalan Marine Sanctuary. The evaluation and comparison of both researches made by DENR and in this paper are necessary to come up with a satisfactory decision as to the assessment of the coastal marine ecosystem. Findings by the DENR include that most of the "target fish" or fish with high economic value are part of the "indicator fish," which is a factor contributing to a healthy reef. Further, Parrotfish and Surgeonfish were identified as high economic value targets for fishing and, at the same time, a contributor to a healthy reef [11][12]. Without careful monitoring and the conduct of community awareness in Tubalan Marine Sanctuary, overfishing and degradation of marine habitats might occur. Extent to data mining algorithms used and with various measure metrics, the simulation result revealed a 98.05% accuracy rate using the KNN algorithm which is higher compared to C4.5 and Naïve Bayes. It denotes that the KNN algorithm is efficient in predicting the accuracy of the novel dataset obtained by the DENR and prior research findings are worthy of implementation and use.

REFERENCES

[1] R. J. King, R. Batista-Navarro, M. Nicolas, V. Hilomen, and G. Solano, "Ecological niche modelling tool for aquatic life population distribution using maximum entropy model," *2017 8th Int. Conf. Information, Intell. Syst. Appl. IISA 2017*, vol. 2018-Janua, pp. 1–6, 2018, doi: 10.1109/IISA.2017.8316390.

[2] S. Takemura, "Fisheries management toolbox: A fishers' self-assessment scheme toward sustainable development of coastal communities," *2018 Ocean. - MTS/IEEE Kobe Techno-Oceans, Ocean. - Kobe 2018*, pp. 3–6, 2018, doi: 10.1109/OCEANSKOBE.2018.8559217.

[3] D. XI, "Report on Habitat Assessment in Tubalan Cove, Malita, Davao Occidental," Davao City.

[4] V. Ribeiro, A. Rocha, R. Peixoto, F. Portela, and M. F. Santos, "Importance of statistics for data mining and data science," *Proc. - 2017 5th Int. Conf. Futur. Internet Things Cloud Work. W-FiCloud 2017*, vol. 2017-Janua, pp. 156–163, 2017, doi: 10.1109/FiCloudW.2017.86.

[5] S. Lin, X. Hongjun, Y. Dequan, and L. Shouju, "Fisheries ecosystem management based on optimization algorithm," *Proc. - 2009 Int. Conf. Environ. Sci. Inf. Appl. Technol. ESIAT 2009*, vol. 3, pp. 19–22, 2009, doi: 10.1109/ESIAT.2009.272.

[6] L. Gao and A. Hailu, "Integrating recreational fishing behaviour within a reef ecosystem as a platform for evaluating management strategies," *Proc. - Int. Conf. Adv. Inf. Netw. Appl. AINA*, pp. 1286–1291, 2010, doi: 10.1109/AINA.2010.67.

[7] L. Sun and Y. Sun, "Marine coastal ecosystem health assessment: A case study in Jiaozhou Bay, China," *2nd Int. Conf. Bioinforma. Biomed. Eng. iCBBE 2008*, pp. 4354–4357, 2008, doi: 10.1109/ICBBE.2008.588.

[8] M. S. Vijayarani, M. M. Muthulakshmi, and A. Professor, "Comparative Analysis of Bayes and Lazy Classification Algorithms," *Int. J. Adv. Res. Comput. Commun. Eng.*, vol. 2, no. 8, 2013.

[9] X. Wang, Z. Jiang, and D. Yu, "An improved KNN algorithm based on kernel methods and attribute reduction," *Proc. - 5th Int. Conf. Instrum. Meas. Comput. Commun. Control. IMCCC 2015*, pp. 567–570, 2016, doi: 10.1109/IMCCC.2015.125.

[10] R. Li, X. M. Wei, and X. W. Yu, "The improvement of C4.5 algorithm and case study," *Isc. 2009 - 2009 Int. Symp. Comput. Intell. Des.*, vol. 2, pp. 190–192, 2009, doi: 10.1109/ISCID.2009.195.

[11] A. J. Cheal, M. Emslie, M. A. MacNeil, I. Miller, and H. Sweatman, "Spatial variation in the functional characteristics of herbivorous fish communities and the resilience of coral reefs," *Ecol. Appl.*, vol. 23, no. 1, pp. 174–188, 2013, doi: 10.1890/11-2253.1.

[12] D. R. Bellwood and J. H. Choat, "A functional analysis of grazing in parrotfishes (family Scaridae): the ecological implications," *Environ. Biol. Fishes*, vol. 28, no. 1–4, pp. 189–214, 1990, doi: 10.1007/BF00751035.