

Application of the ID3 algorithm to optimize anchovy fishing

Ivan Petrlik¹, Pedro Lezama², Henry Bermejo³, Jorge Mayhuasca⁴, JoeMendoza⁵, Saúl Ascue⁶

^{1,3} University Cesar Vallejo

^{2,4,5} National University Federico Villarreal

⁶ University Autónoma of Perú

¹ipetrlika@ucvvirtual.edu.pe, ²plezama@unfv.edu.pe, ³hbermejot@ucvvirtual.edu.pe, ⁴jmayhuasca@unfv.edu.pe, ⁵2015706348@unfv.edu.pe, ⁶SASCUE@autonoma.edu.pe

Abstract

Data mining is a very important analysis that is widely used in multiple business processes, such as agriculture, medicine, fishing among others, and it is a very important element in decision-making systems. It is also primarily concerned with classification due to the dynamic varieties of available data sets. Decision tree-based classification is the foundation of all classification algorithms and is widely used by experts in all types of research. This research implements the ID3 algorithm in the fishing extraction process, allowing to optimize the efficiency and effectiveness of the anchovy fishing process.

Keywords: Data mining, algorithm, efficiency, effectiveness, ID3.

1. Introduction

Improving the extraction of fishing processes, especially anchovy (*Engraulis ringens*), is an increasingly important challenge, since there are only 2 seasons annually because this fish has enormous importance in the marine ecosystem, and you have to focus on when and when fishing should take place. Although there is indeed an average of 86 days per allowed fishing season (information obtained from produces from 2016 to 2021), vessels must focus their effort, to obtain the results of compliance goals.

Year	Fishing seasons	Days per season
2016	Season1 Norte	58
2016	Season2 Norte	83
2017	Season1 Norte	101
2017	Season2 Norte	65
2018	Season1 Norte	116
2018	Season2 Norte	78
2019	Season1 Norte	95
2019	Season2 Norte	87
2020	Season1 Norte	81
2020	Season2 Norte	81
2021	Season1 Norte	100
Average		86

Table 1: Duration of the fishing seasons since 2016. Source Produce

Therefore, it is important to establish a predictive model that allows optimizing the fishing, and for this the ID3 algorithm is used, to perform the analytical processing of the fishing variables and look for the

classification relationship of the most important elements, such as the day of fishing, the size of the boat among others. This research carries out this evaluation, reaching a level of assertiveness greater than 90%.

2. Literature review

[1] in their research entitled "Standardizing catch per unit effort by machine learning techniques in longline fisheries: a case study of bigeye tuna in the Atlantic Ocean", they applied an SVM (support vector machine) to standardize catch rates of fisheries, in which optimization was made with the particle swarm model (PSO-SVM), genetic algorithms (GA-SVM) and grid search algorithms. The results showed a performance above the average using the PSO and GA algorithms in a controlled environment.

[2] in their research entitled "Focused small-scale fisheries as complex systems using deep learning models", proposed three deep learning models to predict the catch of fish, where the model consists of a monthly indexed of each fishing zone and time series; For the first model, the activation function ReLu, MaxPooling, and dense layers are considered; the second model consists of three layers, LSTM, and two dense layers; For the third model, 5 neural layers were applied combining them with LSTM and dense layers.

The proposed models gave a general yield greater than 75%, the yield has an improvement depending on the type of variables used where the yield exceeds 90% for the variables SST, Chl-a, clams, shrimp, warrior crab, sharks; and it is less than 70% in other cases.

[3] in their research entitled "Sequential Fish Catch Forecasting Using Bayesian State Space Models", presents a comparison of predictive models to optimize fishing effort, these models are based on the Hamiltonian Monte Carlo method; Within the Temenos Gradient Increase Decision Tree, SSN-LLM, SSM-2TM, and CFM-RW models.

As a result, they obtained an optimization of up to 70%, with a dependence on the types of the period, it should be presumed that the model is susceptible to sudden changes in fishing behavior.

[4] in their research entitled "ID3 Decision Tree Classification: An Algorithmic Perspective based on Error rate", bases their research on the analysis of the ID3 algorithm where the learning cycle from the initial node is analyzed (root node), entropy and gain; to explore its performance based on the error rate.

The results obtained were from a dataset with 4627 records and 217 attributes, where an average RMSE (Mean Square Error) of 0.77, the average MAE (Relative Absolute Error) of 0.68, average ICI (Incorrectly Classified Instance) average 0.59.

[5] in their research entitled "Mass Incidents Prediction Based on ID3-SMOTE Algorithm", seeks to determine the level of danger of mass incidents that occurred in China, for which the classification of the level of danger is given by Chinese national standard (Small effect (A), Moderate effect (B), Relatively large (C), Significant / Critical (D), Specially significant / Huge (E)); The research consists of the use of the ID3 algorithm for prediction and the SMOTE algorithm to reduce the scarcity of data for some variables.

Two results were presented, the first one made use of the ID3 algorithm where classification A obtained an accuracy of 85%, B - 77%, C - 61%, D - 47%, E - 29%; As can be seen given the scarcity of data in categories C, D, and E, a deficient precision is obtained; In the second case, Using the SMOTE algorithm to improve data quality (Sample Increase), a significant increase in categories C, D and E is obtained in the organ supercar probe with 60% accuracy.

[6] in their research "Classification models to recognize patterns of desertion in university students", proposed the elaboration of 3 predictive models (C4.5, ID3, Neural Networks) to determine the risk of student dropout, The research took place at the National University of San Agustín (UNAS), Peru.

To determine the efficiency of the models, a performance measure was taken of accuracy, precision, and sensitivity, with the result that both the neural networks and the ID3 algorithm obtained a similar performance (Accuracy: Neural Network, 0.52 - ID3, 0.64; Precision: Red Neural, 0.61 - ID3, 0.6 and Sensitivity: Neural Network, 0.61 - ID3, 0.93), while the C4.5 model obtained 0.5 in the three performance tests.

[7] in their research "Application Research of ID3 Attribute Optimization Algorithm Based on Correlation Coefficient", propose to improve the ID3 algorithm based on the concept of spearman's rank correlation; For this, it was used as a dataset of 500 records of the Xi'an Shiyou University Hospital, where it is sought to determine the health status of each patient, on variables such as blood pressure, blood lipids, among others; as well as making a classification of diseases such as hypertension, hyperlipidemia, and hyperglycemia.

algorithm classification form (Classification Rules) as well as the decision making of the same and the simplification of the results when carrying out the construction of the model.

[8] in their research "EKNNIS: Ensemble of KNN, Naïve Bayes Kernel and ID3 for Efficient Botnet Classification Using Stacking", developed several algorithmic models to determine which has the best classification, for the development the EKNNIS technique was used which consists of 2 phases (Pre-processing and Classification). This development is based on determining the malware within normal traffic, the data used was provided by CTU University which consists of 14 variables, following the EKNNIS approach (Pre-growth phase) these variables were subjected to an algorithm which I assigned weights depending on their relevance, of which only 8 were considered.

As a result, they obtained predictions higher than 99%, having naive Bayes at the head with 99.99%, followed by KNN with 99.98 and the ID3 algorithm with 99.96.

3. Methodology

3.1. Data preparation

3.1.1. Step 1

To use the ID3 algorithm, it was necessary to use and adapt the data from Produce (a regulatory entity for anchovy fishing in Peru). These data can be accessed with the link https://extranet.produce.gob.pe/aplicaciones/desembarque/reports/reporte_desem.php automating the download and storing it daily.

3.1.2. Step 2

The transformations of the data obtained in step 1 are carried out, the following criteria are taken into consideration.

- Fishing Date.- It is the day on which the fishing is being carried out, it must be within the dates of the season
- Operation Day.- It is the number of the fishing day from the beginning of the season to the end of it, its behavior is not homogeneous in all seasons.
- Season.- It is the season that the regulatory entity defines it, there are two seasons per year (I, II). Season I in typical situations starts on any day in April and ends between June and July. Season 2 begins in November and ends between January and February of the following year.
- Vessel Registration.- These are the identifiers of the vessels that have quotas allowed for the extraction of anchovy.
- RSW (Refrigerated seawater) .- vessel cooling system, not all vessels have this attribute, in addition, when RSW is activated at the time of fishing, the actual metric tons (gross tons) of fishing are not considered. of the vessel, the net tons are considered, reducing the volume of fishing.
- Vessel size.- is the classification based on its gross registered tons (GRT), but we must remember that in this matter, the term tonnage does not imply weight but volume, the assigned values are: large fleet, medium fleet, and small floats.
- Capacity ranges.- is the weight classification that it can have in gross tons and net tonnes of capacity for the storage of fish.

CAPACITY RANGES
[200 - 300> TM
[300 - 400> TM
[400 - 500> TM
[500 - 600> TM
< 200 TM
>= 600 TM

Table 2: Capacity Ranges. Source: Self-made

- Real gross.- is the maximum quantity of real tons of the vessel's fishing storage, without taking into consideration the RSW.
- Real net tons.- is the maximum amount of real tons of the vessel's fishing storage, considering the RSW in the vessels that own it.
- Boat season quota.- It is the percentage of fishing that is assigned to each boat by the regulatory body.
- Hull Type.- is the type of hull of the boat, it can be Naval Steel, Fiberglass , and Wood.
- Unloading ports.- are the ports where the boats unload the fishing.
- Company boat.- These are the companies that have permits issued by the regulatory entities for anchovy fishing.

3.1.3. Step 3.

To prepare the dataset, the data of the days of operation, season, vessels, and the assigned quotas are integrated, with the real tons, defining the level of fishing compliance. For this, the following formula is proposed.

$$Compliance\ Level = \frac{Tons\ unloaded}{Real\ gross\ or\ net\ tons}$$

3 Compliance-levelscenarios are considered:

Goal	Target Value	Interpretation
> 0.8	1	Meets the defined goal
<= 0.5, 0.8=>	2	Partially meets the definiteness goal
< 0.5	3	Does not meet the defined goal

Table 3: Compliance Scenarios. Source: self-made

3.2. **Generation of the predictive model using the ID3 algorithm (70%, 30%).**-2 evaluations will be used:

3.2.1. **Case 1:** 823,030 fishing records are analyzed, where 70% are provided for learning (576,121) and 30% (246,909) for verification.

3.2.1.1. Obtaining 7 levels of analysis and the dependencies found between the variables, which can be seen in the following image.

➤ Brutas reales o netas reales

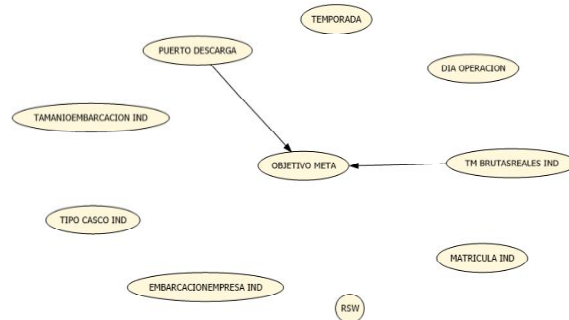


Image 4: Network of dependencies level 2. Source own elaboration

➤ Tipo casco

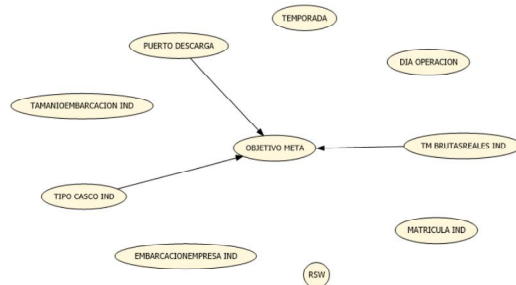


Image 5: Network of dependencies level 4. Source own elaboration

➤ Dia Operación

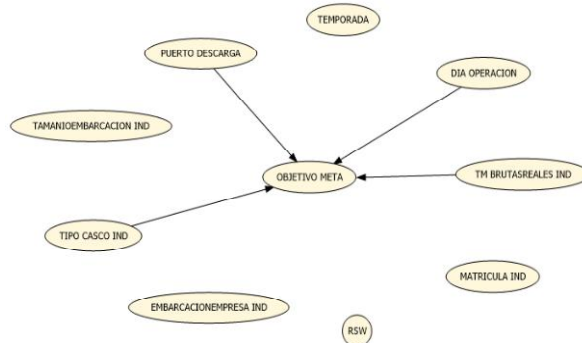


Image 6: Network of dependencies level 4. Source own elaboration

➤ Temporada

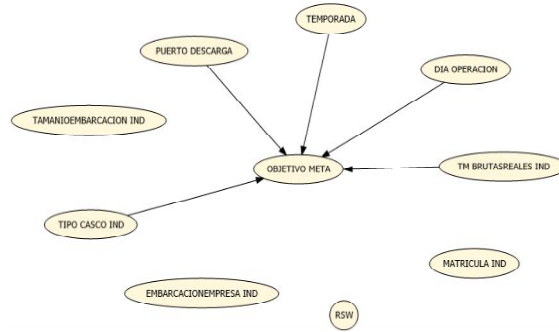


Image 7: Network of dependencies level 5. Source own elaboration

➤ Matrícula Embarcación

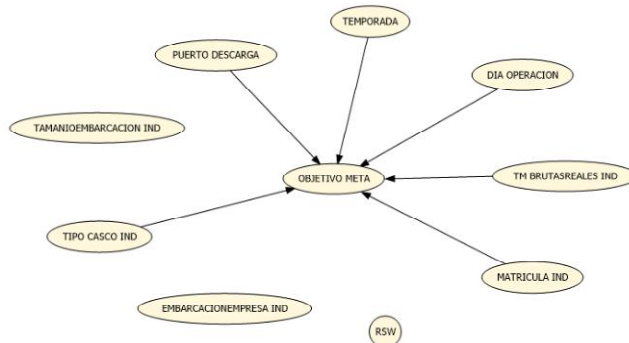


Image 8: Network of dependencies level 6. Source own elaboration

➤ Tamaño de la embarcación

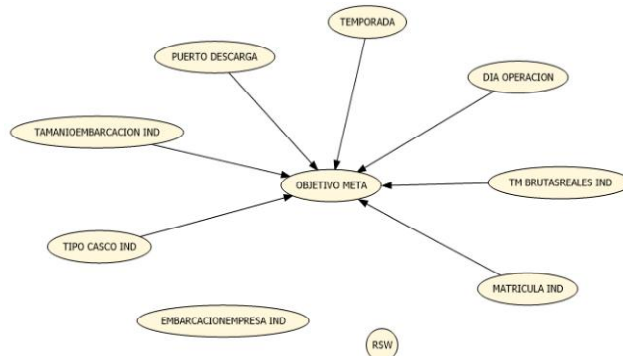


Image 9: Network of dependencies level 7. Source own elaboration

➤ RSW

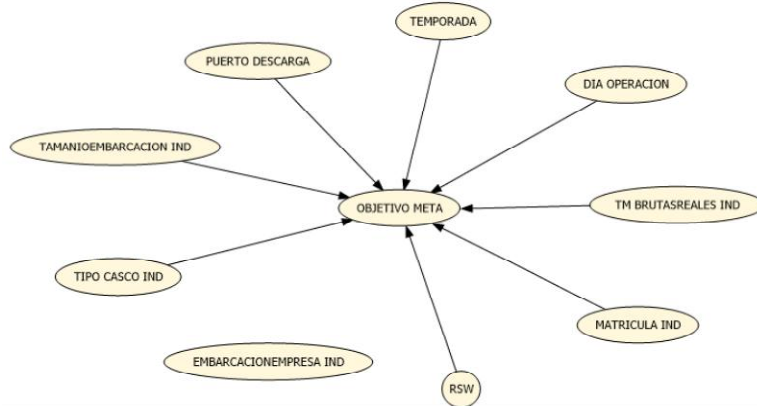


Image 10: Network of dependencies level 8. Source own elaboration

➤ Empresa de la embarcación

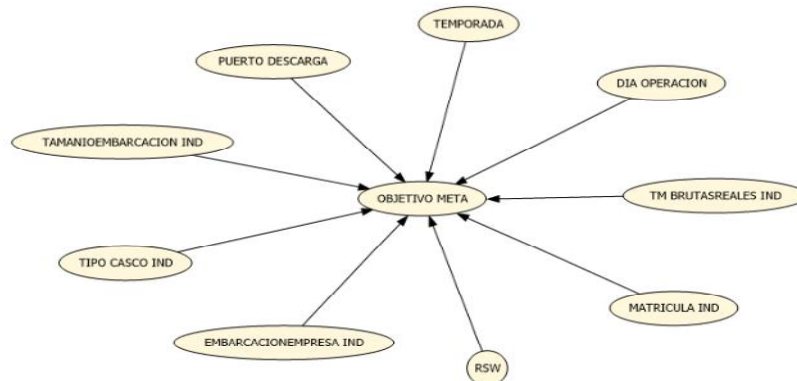


Image 11: Network of dependencies level 9. Source own elaboration

3.2.2. Case 2: Analysis of real and prediction values considering 100% of the data for learning
3.2.2.1. By having 100% learning, 8 levels can be seen in the following graph:

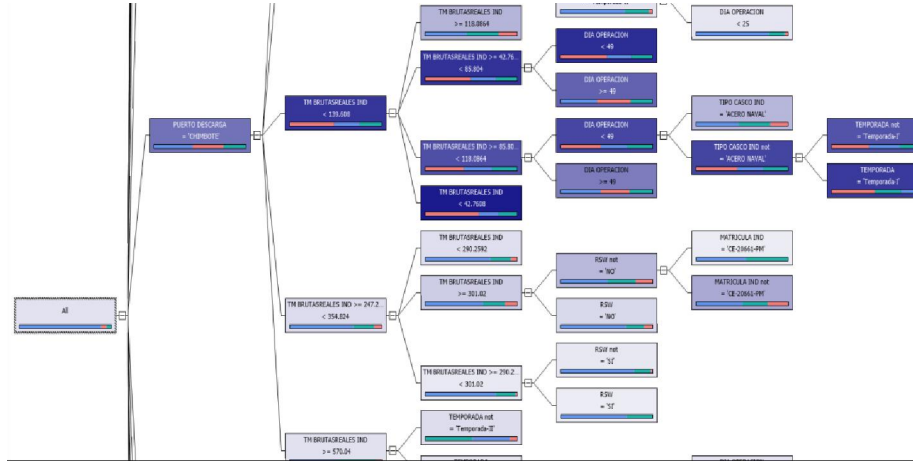


Image 12: Decision tree generated. Source: self-made

4. Result and discusión

4.1. Analysis of actual and prediction values (70%, 30%)

Provided	3 Real	2 Real	1 Real	Assertivenesslevel
3	211,680	8,750	5,102	97.13%
2	1,071	1,322	495	9.30%
1	5,190	4,141	9,158	62.07%
Total	217,941	14,213	14,755	89.98%

Table 4: Actual and prediction values (70%, 30%). Source: self-made

From the table the following conclusions are obtained:

- For objective 3 there is an assertiveness level of 97.13%.
- For objective 2 there is an assertiveness level of 9.30%
- For objective 1 there is an assertiveness level of 62.07%.

Considering the level of assertiveness of objectives 1, 2, 3 and the total of evaluations, there is a general level of assertiveness of 89.98%

To evaluate the level of assertiveness of the predictions, a comparison is made with 100% of real data, obtaining the following results.

Provided	3 Real	2 Real	1 Real	Assertivenesslevel
3	705,797	28,003	16,504	97.14%
2	3,989	5,137	1,800	10.82%
1	16,777	14,350	30,673	62.63%
Total	726,563	47,490	48,977	90.11%

Table 5: Real and prediction values (100%). Source: self-made

From the table the following conclusions are obtained:

- For objective 3 there is an assertiveness level of 97.14%.
- For objective 2 there is an assertiveness level of 10.82%
- For objective 1 there is an assertiveness level of 62.63%.

Considering the level of assertiveness of objectives 1, 2, 3 and the total of evaluations, there is a general level of assertiveness of 90.11%

Compared with the values obtained in item 4.2. Analysis of real and prediction values (70%, 30%), the following results are obtained:

Goal	Assertivenesslevel70%,30%)	Assertivenesslevel(100%)	Variation
3	97.13%	97.14%	0.01%
2	9.30%	10.82%	1.52%
1	62.07%	62.63%	0.56%
TOTAL	89.98%	90.11%	0.13%

Table 6: Comparison level of assertiveness (70%, 30%) Vs level of assertiveness (100%). Source: self-made.

5. Conclusions

In this research, 823,030 fishing records are analyzed, to optimize the efficiency and effectiveness of the fishing extraction processes, using the ID3 algorithm, a level of assertiveness was reached between 89.98% and 90.11%.

The classification rules obtained are also highly confirmed by the algorithm, since 2 tests are carried out, one considering 70% of the records for analysis and 30% for validation, and the other using 100% of the records, reaching a variation between both evaluations of 0.13%.

6. References

- [1] S. Yang, Y. Dai, F. Wei and S. Huiming. Standardizing catch per unit effort by machine learning techniques in longline fisheries: a case study of bigeye tuna in the Atlantic Ocean. *Ocean and Coastal Research*, 83, 93 (2020)
- [2] R. Cavieses Nuñez, M. A. Ojeda Ruiz, A. Flores Irigollen, E. Marín Monroy, M. Ibañez Lucero and C. Sánchez Ortiz. Focused small-scale fisheries as complex systems using deep learning models. 342, 353 (2021)
- [3] Y. Kokaki, N. Tawara, T. Kobayashi, K. Hshimoto and T. Ogawa. Sequential Fish Catch Forecasting Using Bayesian State Space Models. 776, 781 (2018)
- [4] A. Rajeshkanna and K. Arunesh. ID3 Decision Tree Classification: An Algorithmic Perspective based on Error rate. 787, 790 (2020)
- [5] T. Shi, X. Wei , and X. Shao. Mass Incidents Prediction Based on ID3-SMOTE Algorithm, 115-118 (2016)
- [6] J. Zárate Valderrama, N. Bedregal Alpaca & V. Cornejo Aparicio. Classification models to recognize patterns of desertion in university students. 168, 177 (2021)
- [7] M. Gang, C. Liumei and Z. Quancheng. Application Research of ID3 Attribute Optimization Algorithm Based on Correlation Coefficient. 279, 283 (2021)
- [9] A. Niranjan, K. M. Akshobhya, P. D. Shenoy and K. R. Venugopal. EKNIS: Ensemble of KNN, Naive Bayes Kernel, and ID3 for Efficient Botnet Classification Using Stacking. 1, 6 (2018)

Authors



Ivan Petrlik

Doctor in Systems Engineering, Master in Systems Engineering. Research Professor RENACYT – UCV - CONCYTEC. Software Analyst and Programmer with more than 18 years of experience in national and private companies. Member of the digital literacy commission of the College of Engineers of Peru. Member of the Information Technology and Communications Committee. Member of the Community of Knowledge R+D+I+Systems of the

Faculty of Industrial Engineering and Systems of the
Universidad Nacional Federico Villarreal and Member
of the Circle of Ambassadors of Peace - Geneva -
Switzerland



Pedro Lezama

Doctor in Systems Engineering, Master in Systems Engineering with a mention in information technology management, Systems Engineer, PMP Certificate, PMI-ACP, SAMC, SPOC, SMC, SFC, Expert in Project Management and University Professor. with more than 14 years of general experience in the implementation and development of large-scale Information Software.



Henry Bermejo

Doctor in Public Management and Governance, Master in Systems Engineering with mention in Information Technology, Master in University Teaching. University Professor with more than 20 years of experience in different National and State Universities. Collegiate and active member of the College of Engineers of Peru - La Libertad. Post Degree in Human Resources. Specialist in Networks and Communications - CISCO. Chief in the Information Technology Area in the Company BUSINESSOFT Tecnología de Sistemas and with wide participation in Telecommunications companies.



Jorge Mayhusca

Doctor in Engineering, Master in Systems Engineering, Industrial Engineer, Research Professor, Principal Professor appointed by the Faculty of Industrial and Systems Engineering. Director of the Academic Department of Systems engineering. Director of the Professional School of Industrial Engineering. Member of the scientific committee of the Faculty of Industrial and Systems Engineering. Chairman of the quality committee of the Professional School of Systems Engineering.



Joe Mendoza

Systems Engineer at the National University Federico Villarreal, a Computer technician with a mention in Computer Security at the IDAT institute. Consultant in Power BI, Tableau, and analytics. Technical support and help desk specialist. Agile Scrum methodology specialist. Network and Communications Specialist - Cisco Certified Network Associate (CCNA - CISCO).



Saúl Ascue

Systems engineer, more than 2 years of experience in business intelligence and analytics, specialized in Power bi, Tableau Python, and R