The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 10 Issue 1, 2022
pp: 144 - 148

Abalone Age Prediction Using Machine Learning

[1]P. Manasa, [2]S. Ummi Honey, [3]S. Indumathi, [4]Y. Haritha
[1,2,3,4] UG Student, Department of Electronics and Communication Engineering,
Dr K V Subba Reddy College Of Engineering For Women, Kurnool, Andhra Pradesh, India

Abstract
One of the most common kinds of shellfish are abalone. Their shells are frequently used in jewelry, and their flesh is prized as a delicacy. The cold coastal areas are home to the marine snail known as the abalone. The value of an item is heavily influenced by its age. Cutting the shell through the cone, staining it, and counting the number of rings through a microscope are the boring and time-consuming methods used to determine an abalone's age. The age of abalone is predicted using other, less difficult measurements. Sex, length, diameter, height, whole weight, shucked weight, and shell weight are the physical parameters used. Machine learning is used to make the age prediction. One of the most common kinds of shellfish are abalone. Their shells are frequently used in jewelry, and their flesh is prized as a delicacy. In this work, I think about how to figure out how old an abalone is based on its physical characteristics. Because other approaches to estimating their ages take time, this issue is interesting. As a result, working hours could be saved if a statistical method proves reliable and accurate enough. Depending on the species, an abalone can live for up to 50 years. Their growth rate is primarily influenced by water flow and wave activity-related environmental factors. Those living in sheltered waters typically develop more slowly than those living in exposed reef areas due to differences in the availability of food [1]. Because of this, it is difficult to determine an abalone's age, and their size also depends on whether or not food is available. Additionally, abalone occasionally form so-called "stunted" populations, whose growth characteristics differ significantly from those of other abalone populations.

1. Introduction

Abalone is a type of snail that can be eaten, and its price varies depending on its age. The goal is to use physical measurements to figure out the age of the abalone. Traditionally, the age of an abalone is determined by staining the shell, cutting through the cone, and counting the number of rings under a microscope—a tedious and time-consuming procedure. The abalone dataset was first published in 1995, and other measurements, which are easier to obtain, are used to predict the age. Since then, a lot of research has been done using a variety of algorithms and techniques, the first of which was decision trees. On the abalone test dataset in 1999, CLOUDS, a decision tree-based algorithm, was used to achieve a 26.4% accuracy [6] In most classification problems, selecting a split point for the dataset at each internal node involves sorting the values of each numerical attribute, calculating the gini index (a decision tree-based classifier evaluation metric) at each possible split point, and choosing the split point with the lowest gini value. It was discovered that this brute force approach was difficult and computationally demanding [6]. As a result, CLOUDS employed a superior strategy known as SSE [6], in which quantiling techniques were used to divide the

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 10 Issue 1, 2022
pp: 144 - 148

dataset's range into intervals for each attribute, and an estimation of gini values at the boundaries of these intervals was compared to the minimum of the actual gini values at the boundaries.

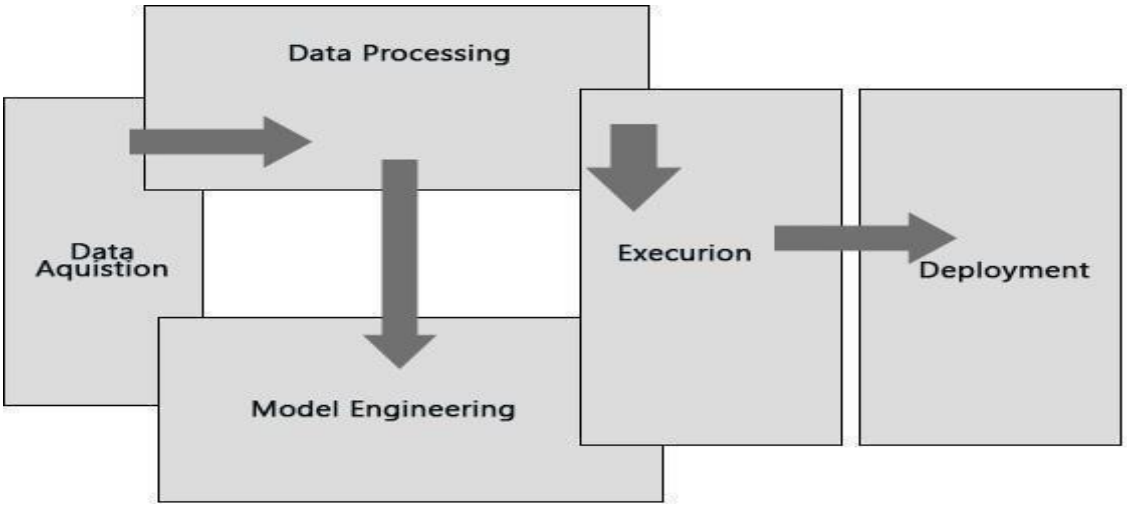

Fig.1 Architecture of Machine Learning

2. Literature Review

After that, only the intervals in which the estimated gini value was lower than the minimum gini value were subjected to the brute force approach. On the dataset, it was discovered that this method was extremely accurate in predicting the correct gini value, in contrast to the brute force approach, which would have required many read operations on smaller memories. However, neither the classification accuracy nor the tree size of the abalone dataset were significantly improved by using SSE in comparison to the sorting method, which achieved an accuracy of 26.3%[6]. Another decision tree method, C4.5, was only 21.5 percent accurate.

A preprocessed dataset with eight reduced classes, all numeric attributes, and one-fourth of the dataset left out for testing is used to run the K-means clustering algorithm, which achieves an accuracy of 61.78 percent [7]. The experiment also helps to determine the relative contribution of various attributes to the accuracy of classification (in order of importance): Height, length, sex, weight by whole, by shell, by viscera, by diameter, and by shucked weight.

Abalones with rings ranging from 3 to 14 were successfully classified into those three classes using an ordered probit model and an Ordinary Least Squares (OLS) regression model that take manipulated attributes as input for estimation of the number of rings and, consequently, age. However, the regressor's estimation was inaccurate.

A supervised learning architecture for neural networks known as Cascade Correlation (CasCor) starts with a minimum number of units and automatically adds and trains new units

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 10 Issue 1, 2022
pp: 144 - 148

one by one. It has a dynamic topology as a result. CasCor does not require backpropagation and can learn more quickly than conventional neural networks. The CasCor network is also able to incrementally detect more complex features because the unit's input weights are frozen when it is added. 8] On the abalone dataset, this architecture with 100 hidden units (but a similar neural network) produced a classification accuracy of 24.43 percent, significantly higher than the 19.73 percent produced by a conventional neural network with the same number of hidden units. 4] When extrapolating the findings, it is safe to say that employing CasCor will outperform a comparable conventional neural network. CasPer is an improvement over CasCor that addresses its generalization issue and propensity for large network formation. Using RPROP and varying learning rates for various units, this is accomplished. With only 50 hidden uni, an accuracy of 30.78% was achieved with just the RPROP gradient descent in CasPer as an improvisation.

3. Proposed System

Seaborn is a Python data visualization library based on Matplotlib. It provides a high- level interface for drawing attractive and informative statistical graphics. This article deals with the distribution plots in seaborn which is used for examining univariate and bivariate distributions. In this article we will be discussing 4 types of distribution plots namely:

1. Joinplot
2. Distplot
3. Pairplot
4. Rugplot

The purpose of a bivariate analysis is to ascertain how two variables relate to one another. Two measurements were taken for each observation in this analysis.

The samples used in this scenario could be independent or paired with different treatments. By and large, in a bivariate examination, the factors utilized can be connected or free (free). When two measurements are taken from the same sample, they are interrelated. In contrast, independent indicates that measurements are taken on two distinct sample groups. Correlation is a sign of how two variables change. We have talked about Pearson's Correlation coefficients and the significance of correlation in the previous chapters. To determine which variable has a high or low correlation with another variable, we can plot a correlation matrix. A correlation matrix is a table that displays the correlation coefficients between variables. The correlation between two variables is displayed in each cell of the table. A correlation matrix is used to summarize data, provide input for more in-depth analysis, and serve as a diagnostic for in-depth analysis.

The Ciência & Engenharia - Science & Engineering Journal
ISSN: 0103-944X
Volume 10 Issue 1, 2022
pp: 144 - 148

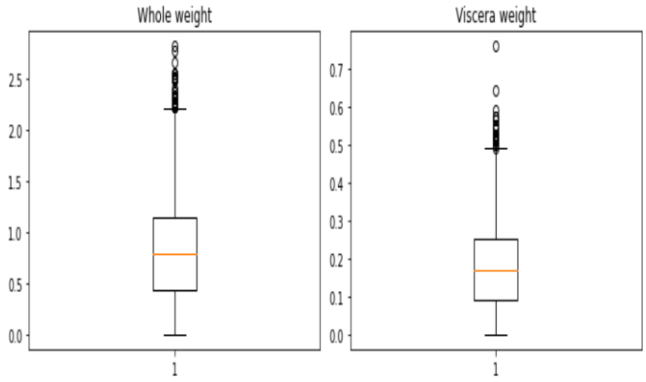

Fig.2 correlation matrix

A machine learning model is defined as a mathematical representation of the output of the training process. Machine learning is the study of different algorithms that can improve automatically through experience & old data and build the model. A machine learning model is similar to computer software designed to recognize patterns or behaviors based on previous experience or data. The learning algorithm discovers patterns within the training data, and it outputs an ML model which captures these patterns and makes predictions on new data

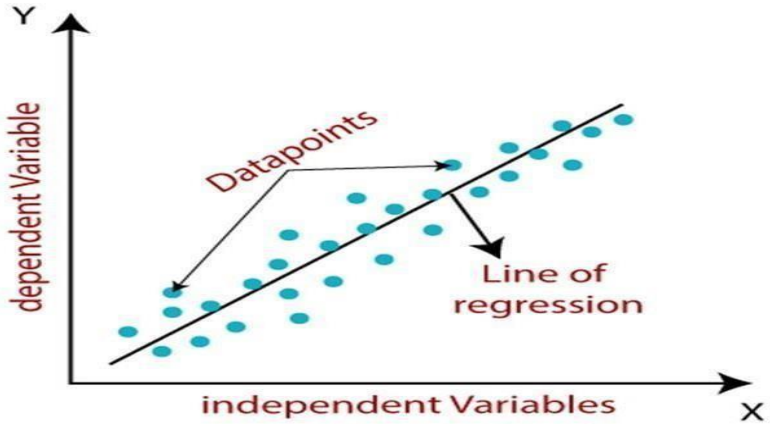

Fig.3 Proposed data set

Leo Breiman and Adele Cutler invented the popular machine learning algorithm known as random forest, which combines the results of multiple decision trees into a single one. Because it can handle both classification and regression issues, its popularity has been fueled by its adaptability and ease of use. The three primary hyperparameters of random forest algorithms must be set prior to training. The number of trees, the number of features sampled, and the size of the node are examples of these. Regression and classification problems can then be solved with the random forest classifier.

The random forest algorithm is made up of a collection of decision trees. Each tree in the ensemble is made up of a data sample called the bootstrap sample that comes from a training set with replacement. The out-of-bag (oob) sample, which we'll discuss later, is made up of one third of that training sample and is used as test data. After that, feature bagging introduces yet another instance of randomness, broadening the diversity of the dataset and lowering the correlation between decision trees. The prediction will be determined differently

depending on the kind of problem. Individual decision trees will be averaged for a regression task, while a majority vote will be used for a classification task.

5.Conclusion

The use of machine learning to determine the age of an abalone is the subject of this project. Physical measurements like sex, length, diameter, height, whole weight, shucked weight, shell weight, and rings can be used to predict an abalone's age. Anaconda Software and a Python IDE called Spyder are used to solve the problem. The target feature was predicted using five classification algorithms and two regression algorithms in this report. Cross validation was used for each method. In hyper parameter tuning, we use grid search to determine which models have the best performance based on accuracy. The results are then presented and discussed in the form of accuracy scores, confusion matrix, and classification report (recall, precision, and F1-score). The RF model outperforms all other classification models in accuracy, recall, and f1-score, as shown by the preceding analysis. However, compared to other models, it is not significantly higher. The restriction necessitates additional investigation.

References

1. Abalone: https://en.wikipedia.org/wiki/Abalone
2. Hossain, M, & Chowdhury, N Econometric Ways to Estimate the Age and Price of Abalone. Department of Economics, University of Nevada (2019).
3. UCI Machine Learning Repository, Abalone dataset: https://archive.ics.uci.edu/ml/datasets/Abalone
4. A.B.Karthick Anand Babu, Design and Development of Artificial Neural Network Based Tamil Unicode Symbols Identification System. IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 2, (2012).
5. Alsabti, K., Ranka, S., & Singh, V CLOUDS: A decision tree classifier for large datasets (1999).
6. K. Jabeen ve K. Ahamed, "Abalone Age Prediction using Artificial Neural Network," IOSR Journal of Computer Engineering, vol. 18, no. 05, pp. 34–38, (2016).
7. Misman, M. F., Samah, A. A., Ab Aziz, N. A., Majid, H. A., Shah, Z. A., Hashim, H., & Harun, M. F. (2019, September). Prediction of Abalone Age Using Regression-Based Neural Network. In 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS) (pp. 23-28). IEEE.
8. Sahin, E., Saul, C. J., Ozsarfati, E., & Yilmaz, A. Abalone Life Phase Classification with Deep Learning. In 2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMI), pp. 163-167 (2018).
9. UCI Machine Learning Repository, Abalone dataset: https://archive.ics.uci.edu/ml/datasets/Abalone
10. Bhatia, N. Survey of nearest neighbor techniques. arXiv preprint arXiv:1007.0085 (2010).
11. Kerber R ChiMerge: discretization of numeric attributes. In: Proceedings of the tenth national conference on artificial intelligence (1992)