Comparative Analysis of Supervised Learning Algorithms for Predictive Modeling of Abalone Age

Adaveni Hari Priya

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

Abstract— *Abalone is a marine snail found in the cold coastal regions. Age is a vital characteristic that is used to determine its worth. Currently, the only viable solution to determine the age of abalone is through very detailed steps in a laboratory. This paper exploits various machine learning models for determining its age. The model was built based on a dataset obtained from the UCI Machine Learning Repository. A comprehensive analysis of various machine learning algorithms for abalone age prediction is performed which include, K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree and Support Vector Machine (SVM). The three classifiers tested to evaluate their effect on its performance. Comprehensive experiments were performed using our data set.*

I. INTRODUCTION

Abalone is a sort of single-shelled marine snails. It is otherwise called the snail of the ocean with a huge beefy body including expansive strong food, which has turned into a costly connoisseur delicacy because of its short fishing seasons [1]. The worth of abalone is monetarily connected with its age. Hence, the age of the abalone assumes an essential part for the two ranchers and shoppers to decide its cost. The age of the abalone is profoundly connected to its costs as it is the sole variable used to decide its worth [4]. In any case, deciding the time of abalone is a profoundly elaborate cycle that is normally done in a research center. Deciding its age is expected in logical examinations like sea life science on abalone. Ordinarily, the quantities of layers of shell rings are estimated to appraise the abalone age [7]. The rings framed relates to the development of the abalone. The most common way of deciding the abalone age begins with chopping down the shell of an abalone. The shell cone is then stained so the rings will be more noticeable to count [8]. Abalone age can be assessed in light of the quantity of dim groups on the right-hand side of the segment. Because of the vulnerability of totally staining all rings, specialists have chosen to add a worth of 1.5 during the ring build up to work on the estimation of the abalone ages. This customary strategy influences ranchers as far as expenses and time handling[4]. Subsequently, to ad lib this cycle, AI can be carried out to help specialists utilizing an enormous number of information containing actual estimations of abalone to foresee its age in a short measure of time.

II. DATA MINING

Information mining includes the utilization of convoluted information examination devices to find beforehand obscure, intriguing examples and associations with regards to enormous informational collection. These instruments can incorporate measurable models, numerical calculation and AI strategies [2][5]. Information mining is most significant examination step of information revelation in data set (KDD) process. The fundamental objective of information mining is to remove the helpful data from immense crude information and switching it over completely to a reasonable structure for its successful and productive use. In like manner, information mining undertakings can be partitioned into two classifications: clear and prescient arrangement procedures [3][6]. Information order is the most common way of coordinating information into classifications/bunches so that information objects of same gathering are more comparative and information objects from various gatherings are exceptionally divergent.

Grouping calculation relegates each example to a specific class to such an extent that characterization mistake will be least. It is utilized to remove models that precisely characterize significant information classes inside the given dataset [5]. Characterization methods can deal with handling of enormous volume of information. It can anticipate all out class marks and orders information in view of model worked by utilizing preparing set and related class names and afterward can be utilized for grouping recently accessible test information. Hence, it is illustrated as an indispensable piece of information investigation and is acquiring ubiquity. Arrangement utilizes regulated learning approach. In managed learning, a preparation dataset of records is accessible with related class names [6][9]. Order process is separated into two primary advances. The first is the preparation step where the grouping model is assembled. The second is the actual order, wherein the prepared model is applied to relegate obscure information object to one out of a given arrangement of class name.

III. METHODOLOGY

A comprehensive analysis of various machine learning algorithms for abalone age prediction is performed which include, K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree and Support Vector Machine (SVM).

3.1 Support Vector Machine (SVM) Classification

SVM is a non-paramatric classifier. SVM is essentially a linear binary classifier that gives test data for classification a class label. It aims to widen the gap between the two classes as much as possible. By using kernel approaches, it can also be used to non-linear data. In an infinitely dimensional space, an SVM creates a hyperplane that is utilized for regression or classification. The closest training data points are used to determine the greatest spacing between two classes to create this hyperplane. Support vectors are the name given to these data points. SVM is a very potent model that can still provide ideal accuracy with fewer training examples.

3.2 Naive Bayes Classification

Naive Bayes Classifier is the straightforward Measurable Bayesian Classifier. It is called naive as it accepts that all factors contribute towards order and are corresponded together. This supposition that is called class contingent freedom[5]. Naive Bayes is an old style order calculation that depends on Baye's Hypothesis. It accepts that the indicators are autonomous. The primary goal is to amplify the back probabilities for each class. Hypothetically, it brings about a base blunder rate and has brought about functioned admirably for different true applications. A naive Bayes classifier thinks about that the presence (or nonattendance) of a specific component (characteristic) of a class is irrelevant to the presence (or nonappearance) of whatever other element when the class variable is given. The Naive Bayes Classifier procedure depends on Bayesian Hypothesis and it is utilized when the dimensionality of the sources of info is high [6]. Bayesian grouping depends on Bayes Hypothesis and Bayes Hypothesis is expressed as beneath:

Let X is an information test whose class mark isn't known and allow H to be some speculation, to such an extent that the information test X might have a place with a predefined class C. Bayes hypothesis is utilized for ascertaining the back likelihood P(C|X), from P(C), P(X), and P(X|C). Where

P(C|X) is the back likelihood of target class.

P(C) is known as the earlier likelihood of class.

P(X|C) is the probability which is the likelihood of indicator of given class.

P(X) is the earlier likelihood of indicator of class.

$$(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

The Naive Bayes classifier [2] functions as follows:

1. Let D be the preparation dataset related with class labels. Each tuple is addressed by n-layered component vector, X=(x1, x2, x3,.....,xn).
2. Consider that there are m classes C1, C2, C3...., Cm. Assume that we need to group an obscure tuple X, then the classifier will foresee that X has a place with the class with higher back likelihood, molded on X. i.e., the Gullible Bayesian classifier appoints an obscure tuple X to the class Ci if and provided that P(Ci|X) > P(Cj|X) For 1 ≤ j ≤ m, and i≠j, above back probabilities are figured utilizing Bayes Hypothesis.

3.3 K- Nearest Neighbor (KNN) Classification

The KNN Calculation is the least complex of all AI calculations. It depends on the rule that the examples that are comparative, for the most part lies in close area [5]. KNN is occurrence based learning technique. Case based classifiers are likewise called sluggish students as they store all of the preparation tests and don't fabricate a classifier until a new, unlabeled example should be grouped [6]. Apathetic learning calculations require less calculation time during the preparation stage than eagerlearning calculations, (for example, choice trees, brain organizations and bayes organizations) however more calculation time during the grouping process[6]. Closest neighbor classifiers depend on advancing by likeness, for example by contrasting a given test and the accessible preparation tests which are like it. For an information test X to be grouped, its K-closest neighbors are looked and afterward X is relegated to class mark to which greater part of its neighbors has a place with. The decision of k additionally influences the exhibition of k-closest neighbor calculation. In the event that the worth of k is too little, KNN classifier might be defenseless against over fitting due to commotion present in the preparation dataset. Then again, assuming k is too enormous, the closest neighbor classifier may misclassify the test in light of the fact that its rundown of closest neighbors might contain a few information focuses that are situated far away from its area.

IV. EXPERIMENTAL RESULTS

The trial assessment was led utilizing the Python programming language with the help of the Python Scikit-learn library. The dataset used in this study was the abalone dataset, Anticipating the time of abalone from actual estimations got from the UCI ML storehouse [10].The dataset comprises of 4177 examples with nine ascribes. Each example addresses eight mathematical characteristics and one downright quality. The objective information is fixed by the quantity of rings for each example of abalone. After the documents have been made, the test dataset contained 1,253 examples while train dataset involved 2,923 examples. We haphazardly split the datasets into preparing and testing sets as indicated by a proportion of 70:30. The preparation dataset was used to perform model execution for the expectation of abalone age. The model appraisal was finished by estimating model execution in view of the outcomes. After the outcomes were gotten, further examination will be finished to break down whether the model could be worked on in light of the size of boundaries.

Table 1
Experimental results

Algorithm	Accuracy	Precision	Recall
Naive Bayes	80.75	80.7	80.72
K-Nearest Neighbors	81.58	81.6	81.6
Support Vector Machine	83.65	83.63	83.6

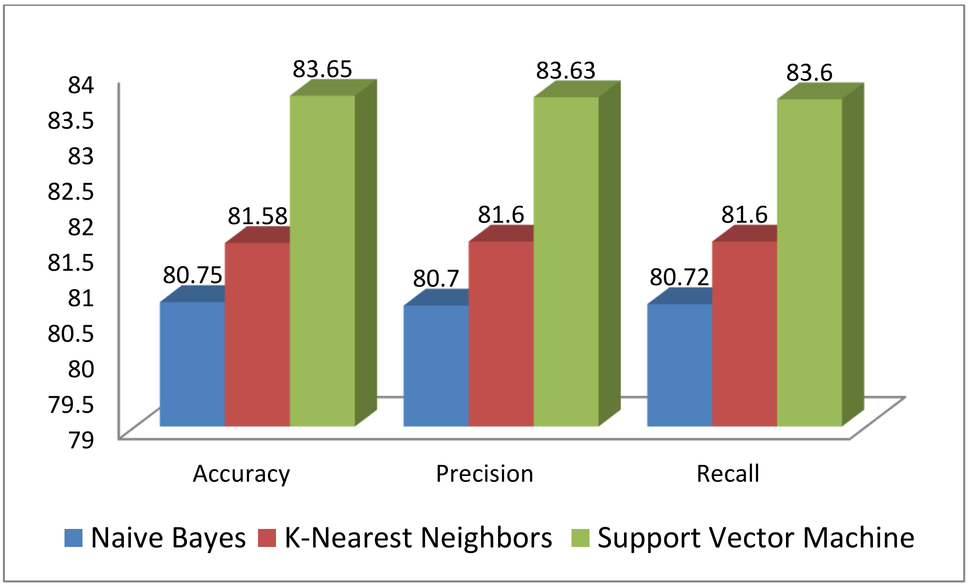

Figure-1: Performance of Classifiers

We observe from the figure-1, the Naive Bayes achieved an accuracy of 80.75%, precision of 80.7%, and recall of 80.72%, Although it provided relatively good results. The K-Nearest Neighbors exhibited slightly better performance than Naive Bayes, with an accuracy of 81.58%, precision of 81.6%, and recall of 81.6%. While it showed improvement, it still fell short compared to Support Vector Machine. The Support Vector Machine demonstrated the highest performance among the three algorithms, with an accuracy of 83.65%, precision of 83.63%, and recall of 83.6%. It consistently outperformed the other algorithms and proved to be the most reliable choice for age prediction of abalone in this study.

Overall, Support Vector Machine (SVM) achieved the highest accuracy, precision, and recall among the three algorithms, with 83.65%, 83.63%, and 83.6% respectively. K-Nearest Neighbors (KNN) performed slightly better than Naive Bayes, but SVM outperformed both in all metrics.

The results indicate that SVM is the most reliable algorithm for predicting the age of abalone in this particular context. It demonstrates the highest overall performance in terms of correctly classifying abalone into their respective age categories.

V. CONCLUSION

The results indicate that Support Vector Machine offers the highest accuracy, precision, and recall, making it the recommended algorithm for predicting the age of abalone in this specific context. These findings emphasize the importance of selecting appropriate algorithms for specific tasks, as different algorithms can yield different levels of performance. Overall, this study highlights the potential of machine learning algorithms, particularly Support Vector Machine, in accurately predicting the age of abalone. Further research and experimentation can help refine and improve these results, leading to more accurate predictions in real-world applications.

REFERENCES

[1] Abalone: https://en.wikipedia.org/wiki/Abalone

[2] G. Ravi Kumar, S. Rahamat Basha, Surya Bhupal Rao, “A Summarization on Text Mining Techniques for Information Extracting from Applications and Issues”, Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 324-332, ISSN: 0973-8975.

[3] G. Ravi Kumar, K. Tirupathaiah and B. Krishna Reddy, “Client Churn Prediction of Banking and fund industry utilizing Machine Learning Techniques”, International Journal of Computer Sciences and Engineering, Volume-7, Issue-6, e-ISSN: 2347 — 2693, PP: 871-875, June 2019,

[4] Hossain, M, & Chowdhury, N Econometric Ways to Estimate the Age and Price of Abalone. Department of Economics, University of Nevada (2019).

[5] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques.2nd ed. San Francisco: Morgan Kaufmann, 2005.

[6] J. Han and M. Kamber,” Data Mining concepts and Techniques”, the Morgan Kaufmann series in Data Management Systems, 2 nd ed. San Mateo, CA; Morgan Kaufmann, 2006.

[7] 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).

[8] Mayukh, Hiran. Age of Abalones Using Physical Characteristics: A Classification Problem. no. Ml, 2010, pp. 1–4Hossain, M. and Chowdhurry, N. M. (2019) ‘Econometrics Ways to Estimate the Age and Price of Abalone’, Munich Personal RePEC Archive, pp. 1–18.

[9] Surya Bhupal Rao, S.Rahamat Basha, G. Ravi Kumar, “A Comparative approach of Text Mining: Classification, Clustering and Extraction Techniques”, Journal of Mechanics of Continua and Mathematical Sciences, Special Issue, No.-5, January (2020) pp 120-131, ISSN: 0973-8975.

[10] UCI Machine Learning Repository, Abalone dataset: https://archive.ics.uci.edu/ml/datasets/Abalone.