Proceedings of the 5th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/20/20231100

# Prediction of the age of abalones based on machine learning algorithms

**Muzi Li**

School of Geography and Bioinformatics, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu, China, 665000

lmz18008795022@163.com

**Abstract.** Abalone is an important seafood, widely used in food, medicine, and other fields. The age of abalone is one of the important factors that determine its quality and market value. However, the traditional age determination method requires the dissection of abalone, which is time-consuming and expensive. Therefore, it is important to find a fast and work out age prediction method. This article uses a machine learning algorithm to predict the age of abalone. The authors collected data on characteristics such as sex, length, diameter, height, and weight for 4177 abalone observations. This data set is admirably large. In the following study, the authors compare the effects of prediction using different machine learning algorithms, including linear regression, decision trees, random forests, and support vector machines. It is worth mentioning that the authors have done sufficient research and evaluation of these algorithms to find out the best prediction scheme. The results show that the random forest algorithm is the best, and its average absolute error is only 1.44 years. The performance of random forest algorithm is the best.

**Keywords:** machine learning, abalone, age prediction, random forest

## 1. Introduction

Machine learning is an important method of artificial intelligence, which can help computers learn and discover data patterns autonomously. At present, machine learning has become one of the core technologies in modern computer science. In machine learning, it is very important to choose suitable algorithms and models. According to different data types, problem types, requirements and application scenarios, different models and algorithms can be selected to complete the task, including neural networks, decision trees, support vector machines, random forests, naive Bayes, etc. In the study of age prediction of abalone, Elvira et al. used nuclear regression model to predict age and other characteristics of abalone, and the results showed that the model had good performance [1]. On the other hand, Wei et al. proposed a deep learning method to predict the age of abalone. They combined convolutional neural network (CNN) and cyclic neural network (RNN) to predict the age of abalone, and the results showed that this method has a good effect [2]. In addition, there are many other research teams exploring the use of machine learning to predict the age of abalone. These studies aim to improve the accuracy of the model, reduce the error rate, and explain the factors affecting the age of abalone from more dimensions. Based on previous studies, this paper further compared the prediction effects of machine learning models such as linear regression, support vector regression, kernel regression and random forest on the age of

DOI: 10.54254/2755-2721/20/20231100

abalone. These results are of great reference value for the management and decision-making of Marine aquaculture industry.

## 2. Data set

This study's data set comes from the UCI machine learning repository and comprises 4177 abalone observations [1]. Each observation included 8 characteristics, one of which was the target variable — age of the abalone. The other characteristics included sex, length, diameter, height, weight, etc.

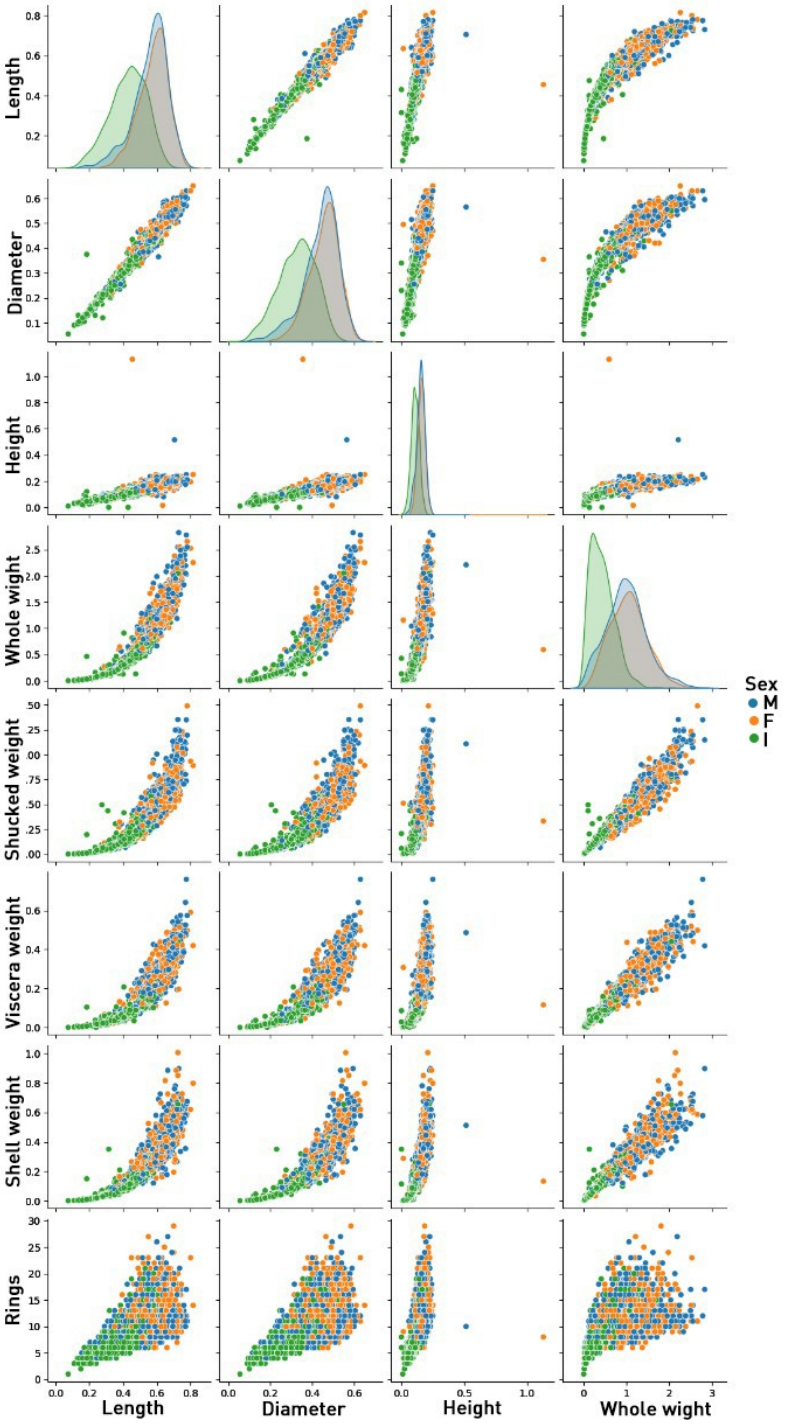


**Figure 1(a).** Scatterplot of abalone feature relationship.

DOI: 10.54254/2755-2721/20/20231100

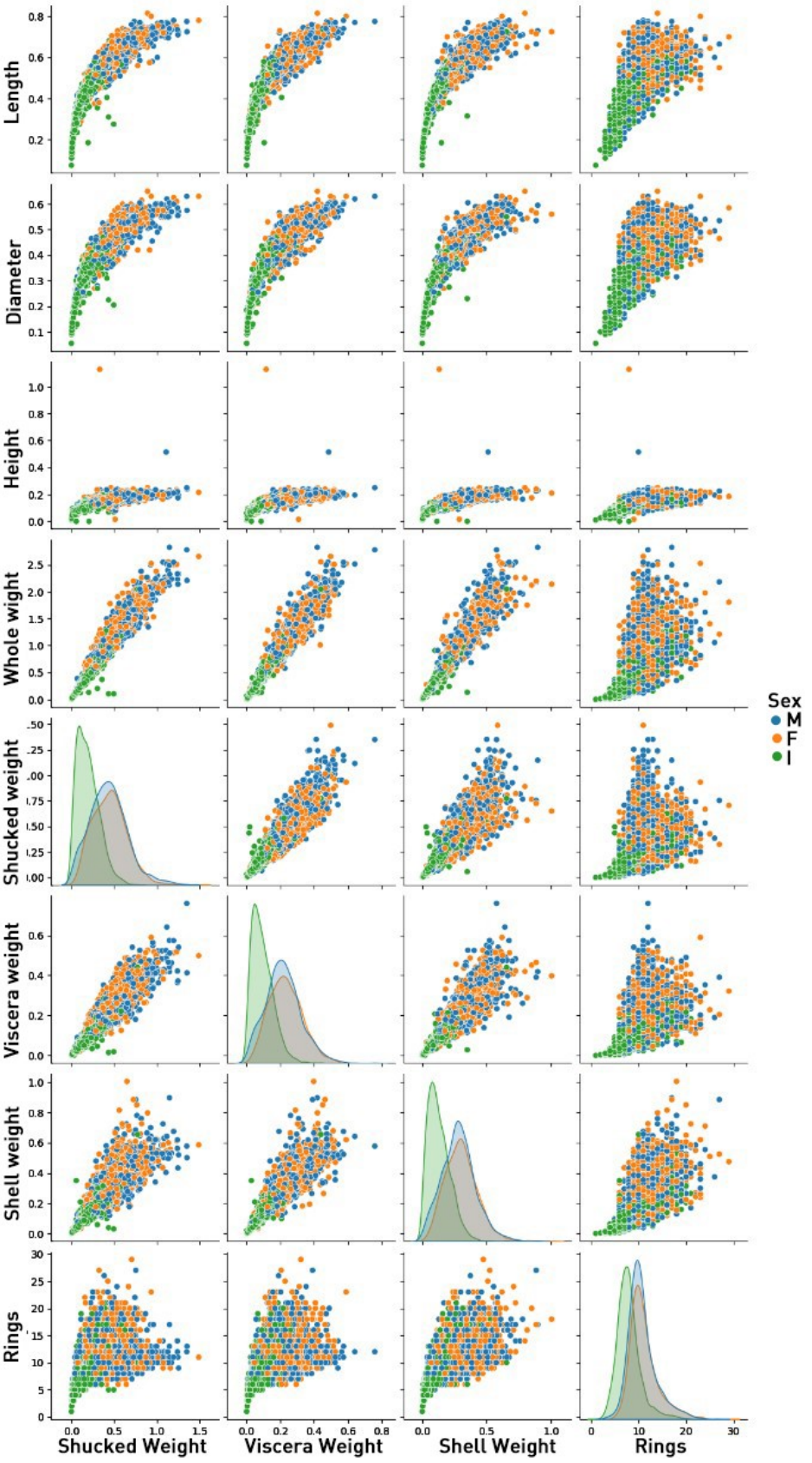


**Figure 2(b).** Scatterplot of abalone feature relationship.

DOI: 10.54254/2755-2721/20/20231100

Through Figure 1, the author draws the following conclusions [3]:

1. The first line of the left figure shows that there is a clear linear link between abalone length and diameter and height of abalone. There is a large nonlinear link between abalone length and the four varieties of abalone weight.
2. The last line of the right picture shows that there is a positive association between abalone ring rings and many attributes, with the linear relationship between rings and height being the most logical.
3. Observing the histogram on the diagonal, it can be seen that the values of various features in juvenile abalone (sex value "I") are significantly smaller than those in other adult abalones. However, there is no significant difference in the distribution of characteristic values between male abalone (sex value "M") and female abalone (sex value "F")

To quantitatively analyze the linear correlation between features, the author calculates the correlation coefficient matrix between features and visualizes the correlation using Figure 2.

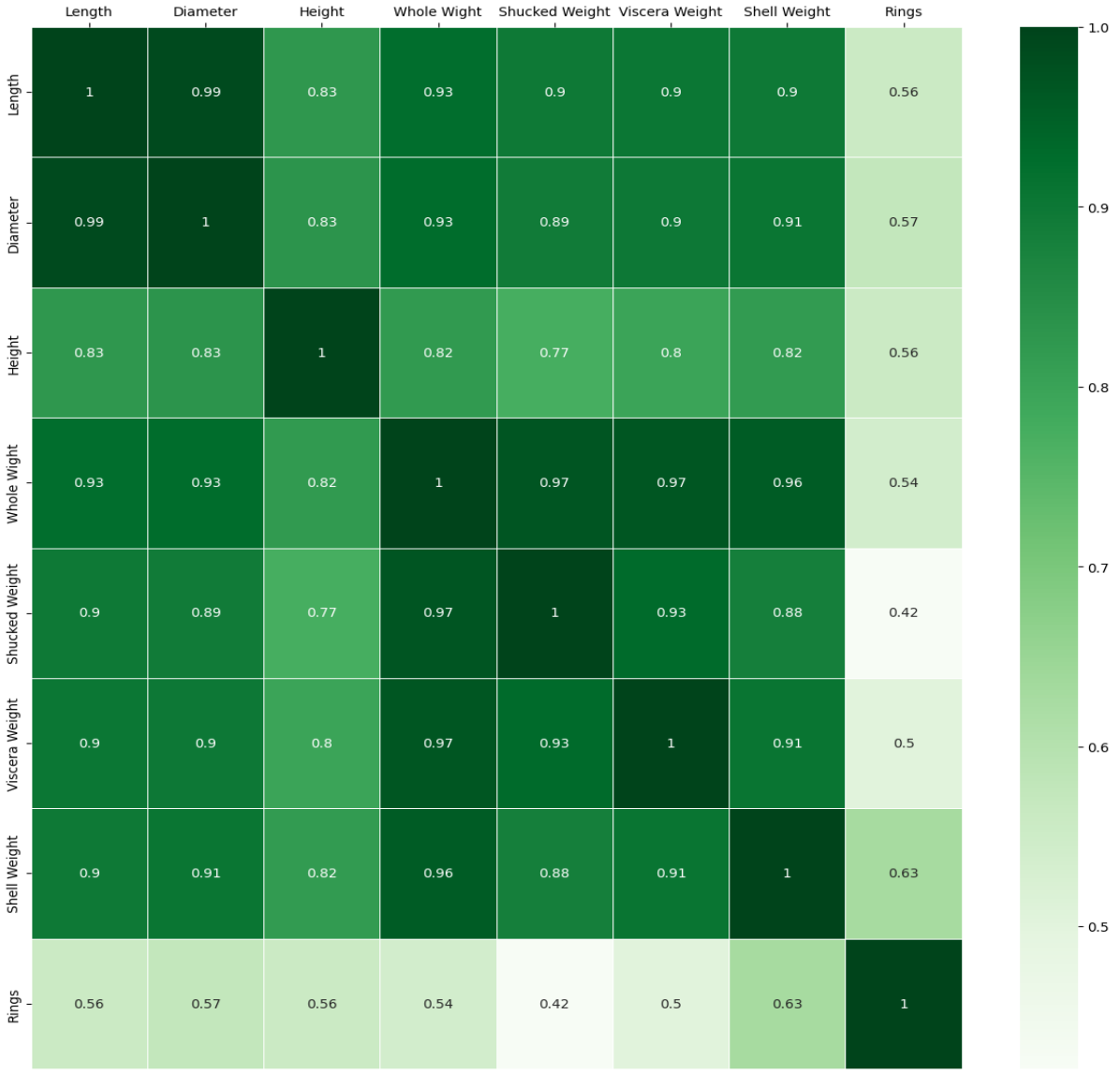


**Figure 2.** Heat map of abalone feature relationship.

DOI: 10.54254/2755-2721/20/20231100

```
[ ]  Sex_onehot = pd.get_dummies(data["Sex"], prefix="Sex")
     data[Sex_onehot.columns] = Sex_onehot
     data.head()
```

| | Sex | Length | Diameter | Height | Whole Wight | Shucked Weight | Viscera Weight | Shell Weight | Rings | Sex_F | Sex_I | Sex_M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | M | 0.455 | 0.365 | 0.095 | 0.5140 | 0.2245 | 0.1010 | 0.150 | 15 | 0 | 0 | 1 |
| 1 | M | 0.350 | 0.265 | 0.090 | 0.2255 | 0.0995 | 0.0485 | 0.070 | 7 | 0 | 0 | 1 |
| 2 | F | 0.530 | 0.420 | 0.135 | 0.6770 | 0.2565 | 0.1415 | 0.210 | 9 | 1 | 0 | 0 |
| 3 | M | 0.440 | 0.365 | 0.125 | 0.5160 | 0.2155 | 0.1140 | 0.155 | 10 | 0 | 0 | 1 |
| 4 | I | 0.330 | 0.255 | 0.080 | 0.2050 | 0.0895 | 0.0395 | 0.055 | 7 | 0 | 1 | 0 |

```
data["ones"] = 1
data.head()
```

| | Sex | Length | Diameter | Height | Whole Wight | Shucked Weight | Viscera Weight | Shell Weight | Rings | Sex_F | Sex_I | Sex_M | ones |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | M | 0.455 | 0.365 | 0.095 | 0.5140 | 0.2245 | 0.1010 | 0.150 | 15 | 0 | 0 | 1 | 1 |
| 1 | M | 0.350 | 0.265 | 0.090 | 0.2255 | 0.0995 | 0.0485 | 0.070 | 7 | 0 | 0 | 1 | 1 |
| 2 | F | 0.530 | 0.420 | 0.135 | 0.6770 | 0.2565 | 0.1415 | 0.210 | 9 | 1 | 0 | 0 | 1 |
| 3 | M | 0.440 | 0.365 | 0.125 | 0.5160 | 0.2155 | 0.1140 | 0.155 | 10 | 0 | 0 | 1 | 1 |
| 4 | I | 0.330 | 0.255 | 0.080 | 0.2050 | 0.0895 | 0.0395 | 0.055 | 7 | 0 | 1 | 0 | 1 |

**Figure 3.** Data preprocessing.

Before using machine learning algorithms, as shown in Figure 3, the author does some pre-processing on the data. First, the author converts sex characteristics into numerical variables, with 0 representing females and 1 representing males. Next, the author normalized the numerical variables to eliminate dimensional differences between them [4]. Finally, the author divided the data set into two parts [5]: training and testing, with 70% going to the training model and 30% going to the testing model.

## 3. Model constructing

This study compares four machine learning technologies, namely linear regression, decision tree, random forest and support vector machine, and analyze their differences in performance. By evaluating and comparing the performance of these four technologies, we can better understand their advantages and disadvantages in specific problems, and thus provide useful guidance for practical application.

### *3.1. Linear regression*

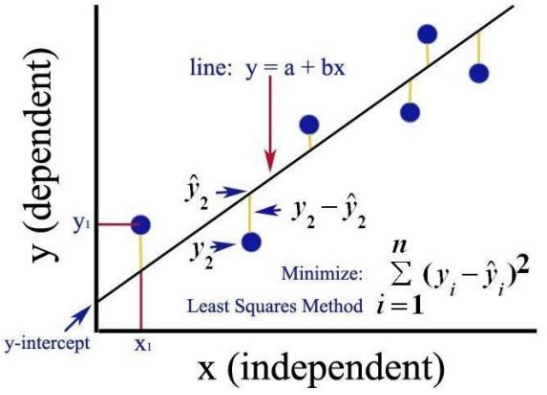


**Figure 4.** Linear regression schematic [6].

DOI: 10.54254/2755-2721/20/20231100

As shown in Figure 4, Linear regression is a common prediction algorithm, which can predict the target variables by fitting a linear model. From a machine learning perspective, the goal is to construct an algorithmic model (function) to map attributes (X) to labels (Y). During the learning process of the algorithm, it attempts to find a function with the best fitting relationship among the parameters. In this study, the paper used a multiple linear regression model, the basic form of which is:

$$y = b0 + b1x1 + b2x2 + .. + bnxn \quad (1)$$

Where, y represents the target variable — age of abalone; b0, b1, b2,... bn is the regression coefficient; x1, x2... xn is the characteristic variable.

### *3.2. Decision tree*

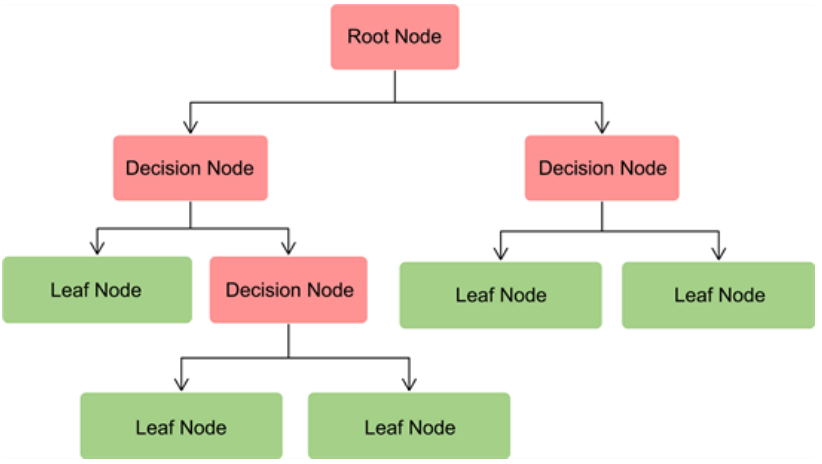


**Figure 5.** Decision tree schematic [7].

Decision tree is a classification and prediction method based on tree structure. The main principle, as seen in Figure 5, is to partition the data set into multiple subsets so that the samples in each subset are as similar as feasible, and the samples in other subsets are as diverse as possible. The decision trees in this study were built using Classification and Regression Trees (CART) methods.

### *3.3. Random forest*

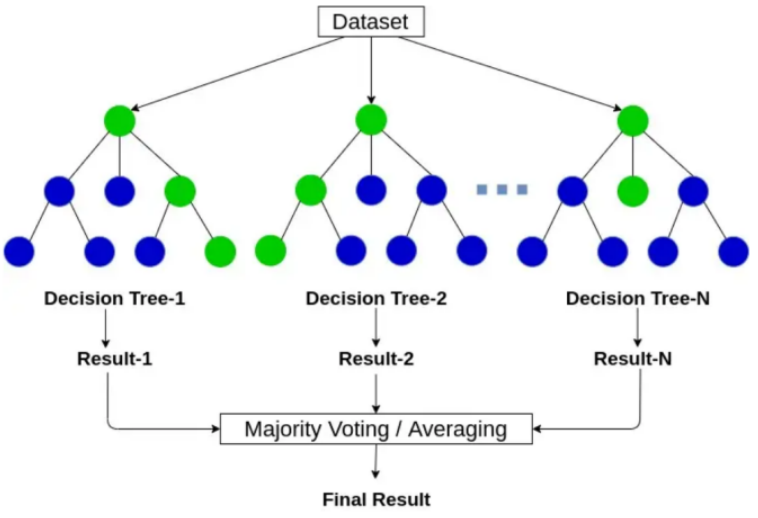


**Figure 6.** Random forest schematic [8].

DOI: 10.54254/2755-2721/20/20231100

Random forest is an ensemble learning-based machine learning technique that mixes numerous decision trees to perform classification or regression problems [9].

As shown in Figure 6, the core principle of random forest is random sampling and feature random selection. In the process of learning the decision tree, a portion of the original dataset is randomly sampled by the bootstrap sample method to obtain multiple decision trees. Meanwhile, when selecting the splitting node, random features are selected for partitioning, and the importance of the feature is calculated. Then, by integrating the predicted results of multiple decision trees, the final classification/regression result is obtained. This paper used the random forest algorithm to predict the age of abalone.

### *3.4. Support vector machine*

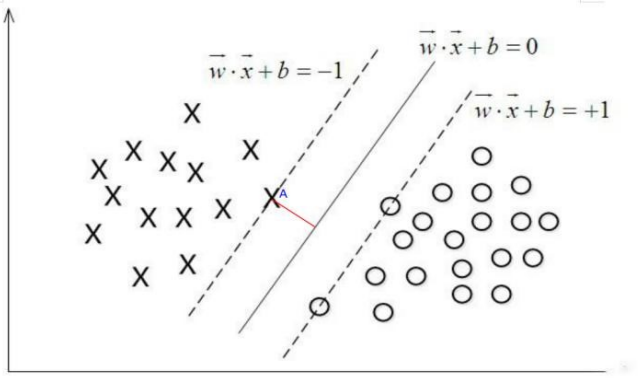


**Figure 7.** Support vector machine schematic [10].

As shown in Figure 7, SVM is a common and sophisticated machine learning technique that is used for classification and regression [11]. The basic idea behind it is to transfer the data to a high-dimensional feature space and then determine the ideal hyperplane that divides the classes with the greatest margin. SVM represents data as points in an n-dimensional space, where n is the number of features. The algorithm seeks a hyperplane that best divides the data points into two classes. This hyperplane is represented as a line, and the margin refers to the distance between the line and the closest data points from either class. The ideal hyperplane maximizes the margin between the two classes, allowing for improved generalization to new, previously unknown data. A Support Vector Regression was utilized in this work to predict the age of abalone.

## 4. Training and results

In regression models, the mean absolute error (MAE) is a measure used to measure the mean absolute difference between the predicted and actual values. It calculates the average of the absolute difference between the predicted value and the actual value, without considering the wrong direction in the error [12]. Therefore, MAE is one of the commonly used evaluation indicators in regression models.

The formula for calculating MAE is:

$$MAE = (\frac{1}{n}) * \sum |y - \hat{y}| \quad (2)$$

where y is the true value; ŷ is the predicted value; n is the number of samples and the symbol $\sum$ represents the summation.

MAE is a useful metric because it gives an idea of how far off the predictions are from the actual values on average. In this paper, it was used to measure the discrepancy between the predicted and actual age of abalone.

In this study, four machine learning algorithms including linear regression, decision tree, random forest and support vector machine are used to analyze the performance differences [13]. It is found that

DOI: 10.54254/2755-2721/20/20231100

random forest algorithm has the best performance, and its mean absolute error (MAE) is 1.44 years, which is significantly better than the other three algorithms. Table 1 shows the specific results obtained for age prediction of abalone. These findings indicate that random forest algorithm has high prediction accuracy and good generalization in this research field, and it can be considered to be applied to related practical problems in the future.

**Table 1.** Abalone age prediction results.

| Model | Discrepancy |
|---|---|
| Linear regression | 1.80 years old |
| Decision tree | 1.61 years old |
| Random Forest | 1.44 years old |
| Support vector machine | 1.64 years old |

## 5. Conclusion

Machine learning techniques were utilized in this study to predict the age of abalone, and according to the findings, it was demonstrated that the random forest approach exhibited the most optimal performance. The random forest algorithm has the following advantages: (1) It can process high-dimensional data and nonlinear relations and can be used as an effective method to predict the age of abalone; (2) Overfitting can be effectively avoided; (3) It has certain robustness to outliers and missing values. However, there are some limitations to this study. First, the paper only considered the influence of basic characteristic variables on the age of abalone, and ignored other possible influencing factors, such as environmental factors and feeding methods. Second, the sample size is relatively small, including only 4,177 samples, and there may be some bias. Future studies may consider adding more characteristic variables, such as the size and color of abalone, to improve the accuracy of the prediction. Furthermore, the sample size can be further expanded to improve the credibility of the study.

## Acknowledgment

This paper was the culmination of a long and difficult journey for me. During this journey, I faced challenges and difficulties, but at the same time, more and more people reached out to me and gave me firm support to help me through the difficulties step by step. In the process of writing my thesis, I am most grateful to my professor Guillermo Goldsztein. He is not only my tutor, but also my good friend I benefited a lot from his Applied Statistics in Data Science course, which made me fall in love with this field, the field of machine learning, completely. Further, I would like to thank my assistant teacher, Roy, who guided me in the right direction with their professional knowledge. Lastly, I am extremely grateful to my parents for their support.

## References

[1] Sam Waugh. 1995. Abalone Data Set. UCI Machine Learning Repository: Abalone Data Set
[2] Irvine M A, et al. Adult data set. UCI Machine Learning Repository, 1998.
[3] Yizhen Han. Machine Learning Project — Predict the Age of Abalone, 2019.
[4] Misman M F, Samah A A, Aziz N, et al. Prediction of Abalone Age Using Regression-Based Neural Network[C]// 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS). 2019.
[5] Hastie T, et al. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media, 2009.
[6] Linear Regression Schematic (Descending into ML): Linear Regression https://zhuanlan.zhihu.com/p/74616776
[7] Decision Tree Schematic Prediction of Wheat Rust Diseases Using Data Mining Application Prediction of Wheat Rust Diseases Using Data Mining Application (scirp.org)
[8] Random Forest Schematic. The relationship between decision trees and random forests. https://blog.csdn.net/qq_39777550/article/details/107312048
[9] Breiman L. 2001. Random forests. Machine learning, 45(1): 5-32.

DOI: 10.54254/2755-2721/20/20231100

[10] Support vector machine Schematic Support vector machine series (2) — Mathematical representation of the original SVM problem https://zhuanlan.zhihu.com/p/28201195
[11] Drucker H, et al. 1997. Linear Support Vector Regression Machines.Advances in neural information processing systems, 155-161.
[12] Han C W. 2017. Predicting the Age of Abalone using Morphological Neural Networks[C]// Smart Technologies in Data Science and Communication 2017.
[13] Jabeen, K., & Ahamed, K. I. (2016). Abalone age prediction using artificial neural network. IOSR J Comput Eng, 18(05), 34-38.