Jake Stephens

Project

August 20, 2021

For my project I chose to analyze abalone data I found on the machine learning data repository website. The goal of my project is to find a good way to predict the number of rings on the shell of the abalone. In order to do this, I used several different methods and compared them. First, I used basic linear regression by choosing 1 field at a time to see which field had the greatest correlation with the number of rings. My results shown in figure 1 clearly depict that shell weight is has the highest correlation due to the fact that it has the lowest SSE of all the fields for linear regression. I was thinking that the shell weight would be the best column for quadratic and cubic methods, but I ran them against all fields just to see the numbers. After analyzing the results from the linear regression, quadratic, and cubic methods of predicting the number of rings, it is abundantly clear to me that the shell weight has the highest correlation or closest relationship to the number of rings compared to the other fields. Also, it is clear to me that the higher the degree, the more accurate the method became. For linear regression the SSE of shell weight was 13,343, but for quadradic 13,013, and for cubic it was even lower at 12,865. Although it is easy to tell by looking at the number, that the cubic method (degree of 3) has the highest accuracy, however it is hard for me to explain why by looking at the graph (see figures 2-4.) You can see by looking at the graphs that the lines have different slopes that they follow, but I find it hard to see visually how the cubic method fits the data better than the other methods.

For my next experiment I wanted to use Knn to try and predict the number of rings. Except I manipulated the data slightly. Instead of predicting the number of rings (as Knn is for predicting labels), I divided the data into 2 groups, Child, and Adult. An abalone is considered an adult if it has more than 10 rings, and a child if it has 10 or less rings. After computing the maturity of the abalones, I ran the dataset through Knn to try and predict if an abalone would be an adult or not. To my surprise, the Knn classifier worked very well on this, with accuracy’s around ~75%. With k=13 being the best choice with an accuracy of 76.5%. This accuracy may also be influenced by the fact that there are significantly more children than adults in the data set. There are 2,730 children and 1,447 adults. Possibly with a more evenly distributed dataset, the accuracy may be different. But I do believe that since the data is off balance and the accuracy is still high, that the Knn classifier is a good method to use to predict the maturity of an abalone.

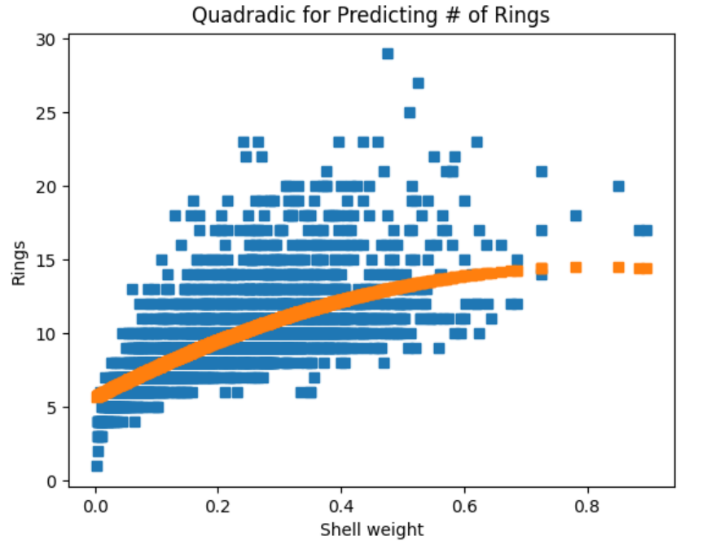
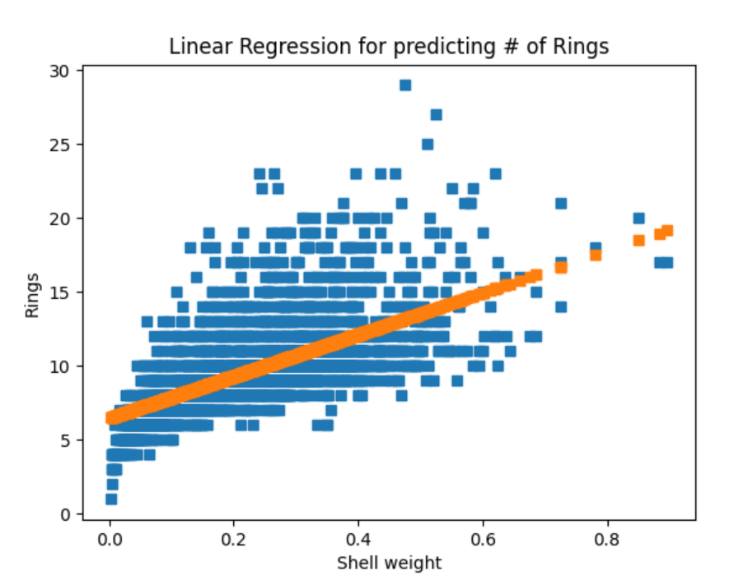
**How to run:** To run this program, run the launch.py file.

Dataset: <https://archive.ics.uci.edu/ml/datasets/abalone>

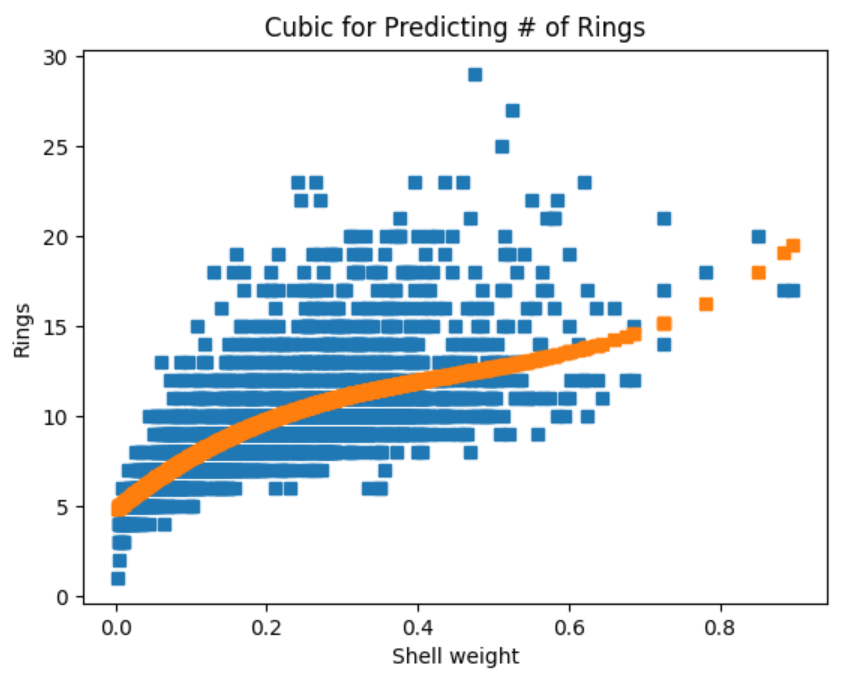
Text

Description automatically generated

***Figure 1***



***Figure 2 Figure 3***



***Figure 4***