




# DATA 2204: K-NN Regression – Assignment #1

**Name:** Raj Dholakia  
**Student ID:** 100813041



# Table of Contents

- Analysis Statement
- Key Features of the Dataset
- Understanding the Correlations
- Understanding the Learning Curve
- Standard vs Optimized k-NN Regression Models
- What should be the next steps?



# Analysis Statement

Create a **standard and optimized k-NN model** to predict the age ('Rings') of abalones using their physical features.

# Key Features of the Dataset

- The dataset consists of seven (7) independent variables and one dependent - 'Rings'.
- 'Rings' provides an estimate of the age of the abalone.
- An abalone is an edible marine mollusc.
- There are a total of 4177 samples in the dataset.
- The dataset has no missing values.
- The mean of 'Rings' (dependent variable) is 9.93, hence an RMSE/MAE value of less than 0.993 or 10% should reflect a good model.



## Abalone Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: Predict the age of abalone from physical measurements



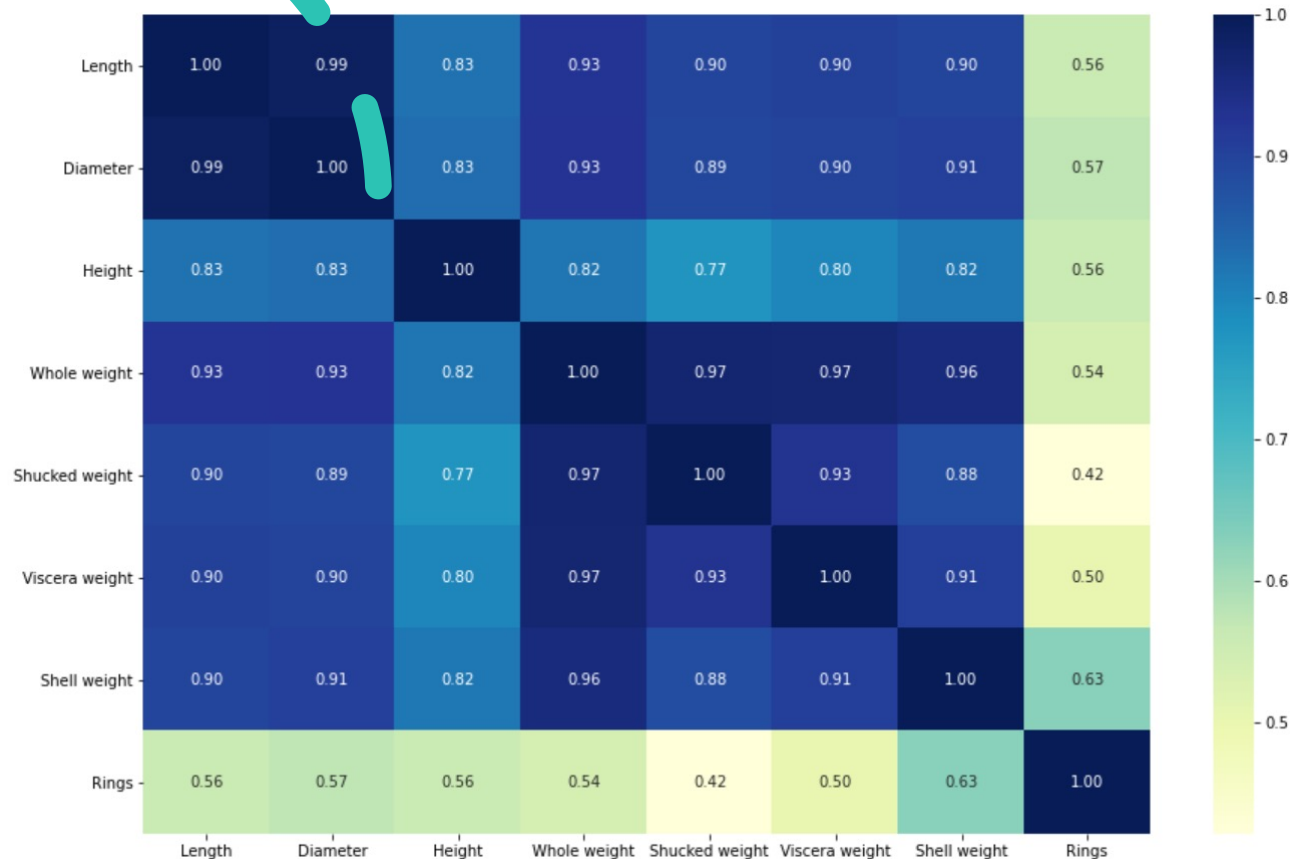
Data Set Characteristics:	Multivariate	Number of Instances:	4177	Area:	Life
Attribute Characteristics:	Categorical, Integer, Real	Number of Attributes:	8	Date Donated	1995-12-01
Associated Tasks:	Classification	Missing Values?	No	Number of Web Hits:	1096813



# Understanding the Correlations

Some insights gained from the heatmap of the data:

1. All the independent variables have **strong correlations**, with the minimum being 0.77 between 'Shucked weight' and 'Height'. All the other variables have a strong correlation of 0.8 and higher.
2. 'Rings', the dependent variable, is **moderately correlated** to all the independent variables. 'Shell weight' has the largest correlation value of 0.63, while 'Shucked weight' has the lowest correlation value of 0.42.



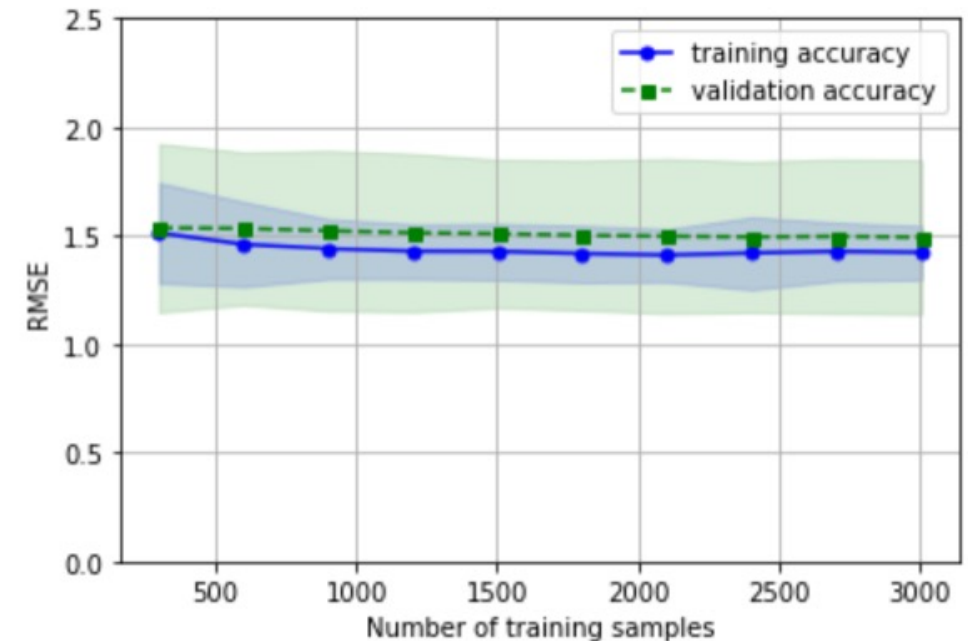
# Understanding the Learning Curve

Some insights gained from the Learning Curve of the k-NN Standard Model:

1. From the key statistics, we had determined that a good model would be one with a **RMSE value** lesser than 0.99 ( $\cong 1.00$ ). However, here we can see that the lowest is around **1.40**.
2. As the RMSE value is not close to the ideal value, we should expect a **high bias** in the model.
3. As the training and validation accuracies are close to each other we can see that the **variance is low**.
4. Notice how the learning **curve is flat**, which means that adding more of the data is not improving the model by a significant amount.

Note: the above insights can be confirmed from the values of the avg. bias and avg. variance.

k-NN Regressor Learning Curve



Bias Variance Trade-Off

Estimator: kNN

Average Bias: 4.88

Average Variance: 0.44

# Standard vs Optimized k-NN Model

Considering the standard (original) model:

- The model is only able explain **53% ( $R^2$ )** of the variability in the dataset. This shows that the model is **not performing well** on the dataset.
- As the values of  **$R^2$  and Adj\_  $R^2$  are the same**, we can assume that the model's performance is being reflected accurately.
- There is a difference of 0.68 ( $\cong 30\%$ ) between **RMSE** and **MAE**. This underpins **presence of some outliers**.

Comparing the models:

- There is **negligible difference** between the metrics obtained for both the models. Hence, the **performance** of both the standard (original) and optimized models is the **same**.
- This lack in improvement of the model can be attributed to the limited number of hyperparameters present in the k-NN Regression algorithm.

Original Model

n\_neighbors: 17

$R^2$ : 0.53

Adj\_  $R^2$ : 0.53

Mean Absolute Error: 1.52

Mean Squared Error: 4.83

Root Mean Squared Error: 2.20

Optimized Model

Estimator: k-NN Regression Model

Fitting 10 folds for each of 192 candidates, totalling 1920 fits

```
Best params: {  
  "algorithm": "auto",  
  "n_neighbors": 24,  
  "weights": "distance"  
}
```

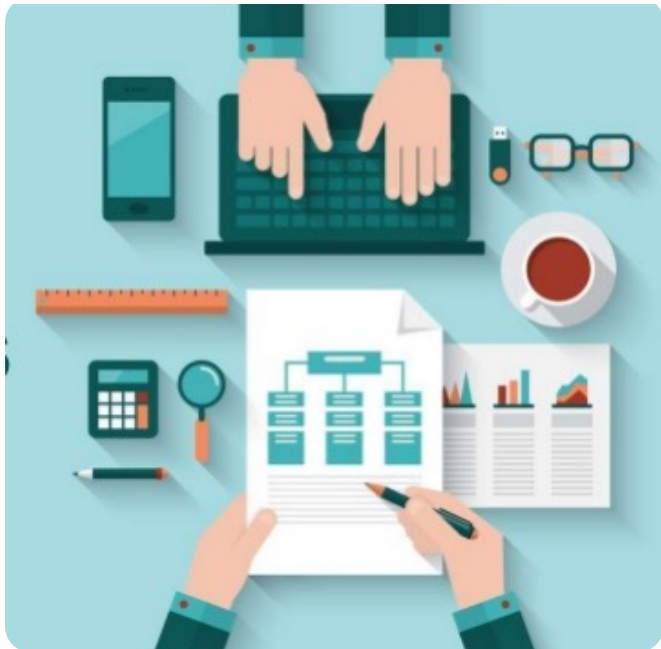
$R^2$ : 0.54

Adj\_  $R^2$ : 0.53

Mean Absolute Error: 1.51

Mean Squared Error: 4.81

Root Mean Squared Error: 2.19



# What should be the next steps?

*Mr. John Hughes* can take into consideration the following options to better predict the age ('Rings') of abalones using a k-NN Regressor:

- 1. Add more independent features to the dataset:** the current variables in the dataset are strongly correlated to each other. Furthermore, the correlation between the independent and dependent variables is towards the lower end of the spectrum. This limits the ability to use the independent variables (features) of the dataset to predict the dependent variable (response). Adding more features could provide better results as more of the variability in the dependent variable could be explained.
- 2. Get more data samples:** increasing the size of the dataset will provide a better understanding of the relationships between the variables. However, only if more independent features are added to the dataset. The learning curve showed that adding more of the same features does not improve the performance of the model. Though, the performance of the model might improve if more data points are added.
- 3. Feature engineering:** this would be the third step in improving the performance of the model. Additional independent features and more data will provide more opportunities to carry out feature engineering.





Thank you

Raj Dholakia

raj.dholakia@dcmail.ca

Student ID: 100813041