Burcu Ibicioglu
2115708

Machine Learning Assignment Report: Turtle Age Prediction

Predicting Turtle Age Longevity has historically been estimated using intrusive methods such as counting growth rings on shell cross-sections. Nevertheless, this process could cause harm to the turtles. The objective of this project is to develop a machine learning model that predicts turtle age non-invasively using morphological features. By using regression techniques, this approach aims to provide conservationists and wildlife researchers with a more ethical and practical solution.

Data and Preprocessing

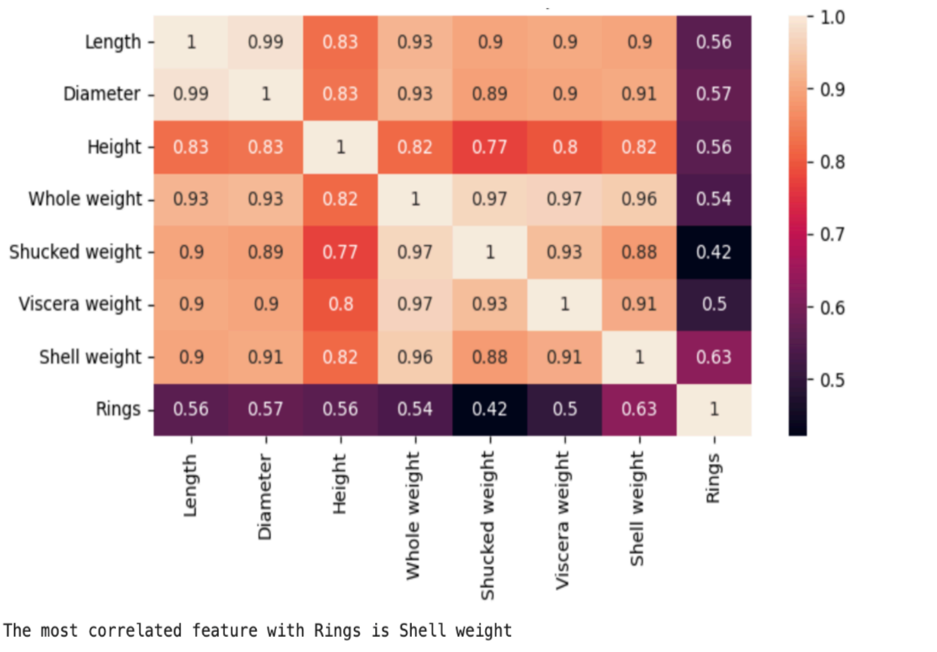
Overview

The dataset provided in this study contains multiple physical features of turtles, including:

- **Numerical Features:** Length, Diameter, Height, Whole Weight, Shucked Weight, Viscera Weight, Shell Weight
- **Categorical Feature:** Sex
- **Target Variable:** Rings (Turtle age)

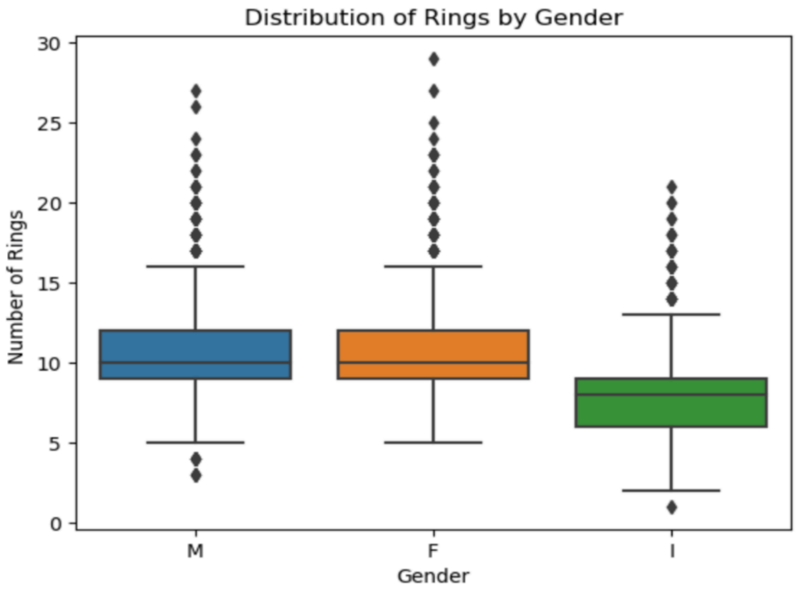
Data Exploration

I started my analysis by creating a heatmap to understand the data and correlations between variables. Through this map, I was able to explore which features had the highest association with the target variable; **Rings**. As you can see in the *Correlation Heatmap* below, **Shell Weight** is the most correlated feature with **Rings.**



Correlation Heatmap

I have also assessed the relationship between **Gender** and number of **Rings** (turtle age) *Box Plot (1),* which showed me that Gender does not play a significant role on the turtle age.



Box Plot (1)

Data Splitting

I then split the dataset as follows:

- 80% for training
- 10% for validation
- 10% for testing

In order to keep the balance between the three, I applied binning to the **Rings** variable using **KBinsDiscretizer** minimizing the possible overfitting that may occur.

Model Training and Tuning

For this part, I trained and analyzed three models of regression;

Linear Regression	2. Polynomial Regression	3. K-Nearest Neighbors (KNN) Regression
As a baseline model, it produced the following results: • **R^2 = 0.57**, **MSE = 5.19**, **MAE = 1.62** • Although interpretable, it was unable to capture complex relationships.	Polynomial regression with degrees 2–5 was tested to model non-linear connections. Degree 2 produced the greatest outcomes: • **R^2 = 0.59**, **MSE = 4.98**, **MAE = 1.57** • Higher degrees (4 and 5) caused overfitting.	Models were tested with **k = {1, 3, 6, 10}**, with **k=10** performing best: • **R^2 = 0.55**, **MSE = 5.50**, **MAE = 1.62** • Lower k-values caused overfitting.

Results and Discussion

Model	R^2 Score	MSE	MAE
Linear Regression	0.57	5.19	1.62
Polynomial (Degree 2)	0.59	4.98	1.57
KNN (k=10)	0.55	5.50	1.62

The performances of the models were compared using the evaluation metrics on the left table. Polynomial regression (degree 2) was the best-performing model, providing the optimal balance between accuracy and generalization.

Cross-Validation

A **10-fold cross-validation** on Polynomial Regression (degree 2) resulted in an **average R^2 of 0.58**, confirming its reliability.

Feature Importance Analysis

For the polynomial regression model, the **most influential features** were identified through coefficient analysis. **Shell Weight, Whole Weight, and Length** were the strongest predictors of turtle age. In order to visualize these relationships, I created a feature importance bar chart (see appendix).

Error Analysis

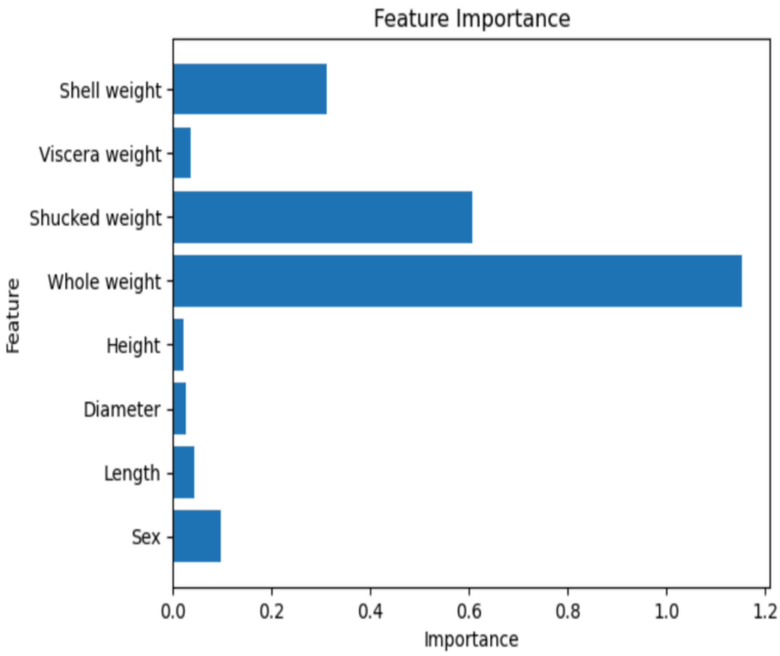
I then discovered two systematic errors:

- **Overestimations:** The ages of larger turtles are often overestimated.
- **Under-Predictions:** Turtles that were smaller were often overlooked.

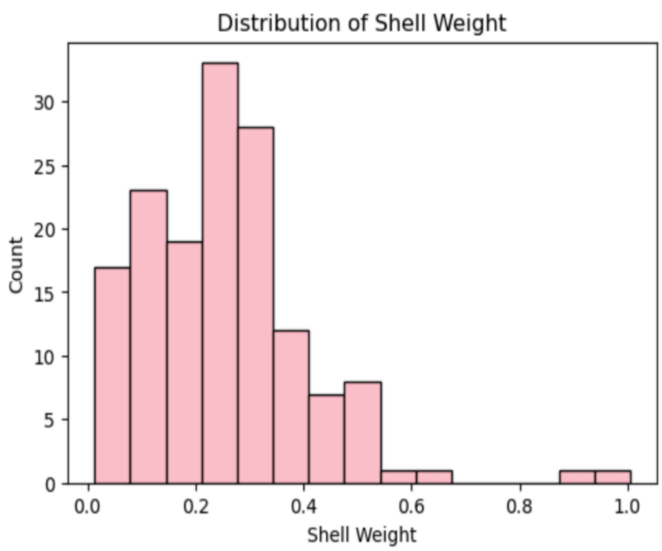
Conclusion

While my initial results suggested multiple models with similar performance, cross-validation confirmed that **Polynomial Regression (degree 2)** is the most effective method to predict turtle's ages non-invasively. For the future improvements of the dataset, more data can be collected, particularly for extreme age groups, and incorporating additional biological indicators to refine the model further.

Appendix



Feature importance bar chart



Histogram of Distribution of Shell Weight