**Mahamadou Mahi Diallo**

**Northcentral University**

**TIM-8555: Predictive Analysis**

**Dr. Mohammad Yavarimanesh**

**October 13, 2024**

**Abalone Data Analysis Report**

This analysis focused on the Abalone dataset, which contains biological measurements such as length, diameter, height, and various weight categories. The primary objective was to preprocess the data, explore statistical properties, and assess whether key features followed a normal distribution.

**Data Preprocessing**

To reduce the impact of the features and bring them to a similar scale the Min-Max scaling normalization technique was applied to the dataset. This preprocessing was important because the features in the dataset are of different scales, for example length and weight. Naturally normalizing the data increased comparability and made it possible to apply some specific statistical models.

**Statistical Testing**

We have used the Kolmogorov-Smirnov test to analyze the normal distribution properties of given data and we have also used QQ plots for the binomial distributions.

* **Height**: The QQ plot for Height showed that the data points are close to the reference Gaussian distribution, thus being able to deduce that, Height has a distribution that is potentially normal.
* **Length**: For Length the K-S test produced a K-squared value of 1.145 and a significance probability equal 0.564. We did not reject the null hypothesis because the p-value was greater than the alpha of 0.05 Thus, it is probable that Length also follows normal distribution.

**Insights and Observations**

The statistical testing indicated that several of the attributes signed for the model indeed conform to Gaussian probability distribution, including Height and Length. This means that common parametric statistical models which presuppose normality could be used to this sampled data set with reasonable accuracy. It is easier to differentiate among the regularities and contrast the diverse measures within the biological field due to the normalized characteristics.

**Recommendations**

As for future work, further analyses might extend the current work by exploring interactions between other characteristics, for example, age, gender, or body measurements. There it stated that regression analyses could be used to make estimates on dependent variables, such as weight given length or height. Some other analysis could also look at how better to categorize abalone to different families or groups depending on their biological attributes.

References

Alita, D., Putra, A. D., & Darwis, D. (2021). Analysis of classic assumption test and multiple linear regression coefficient test for employee structural office recommendation. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *15*(3), 295-306. [https://journal.ugm.ac.id/ijccs/article/view/65586](about:blank)

Gelman, A., Hill, J., & Vehtari, A. (2021). *Regression and other stories*. Cambridge University Press. [https://books.google.com/books?hl=en&lr=&id=fILoDwAAQBAJ&oi=fnd&pg=PR11&dq=Regression+Modeling+and+Assumptions&ots=WyWXuupmN1&sig=erZuAgalMfgDP8NWgH6Yha6yBl8](about:blank)

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Linear regression. In *An introduction to statistical learning: With applications in python* (pp. 69-134). Cham: Springer International Publishing. [https://link.springer.com/content/pdf/10.1007/978-3-031-38747-0\_3.pdf](about:blank)