Spring 2025

Stockton Graduate Research Symposium

Thinh Le

Title Ideas:

Predicting the Presence of Marine Megafauna Using Machine Learning and Environmental Data

Timeline:

Symposium abstract due by: April 7th

Poster completed by print due date: April 22, 2025

Stockton University Graduate Research Symposium: Monday, April 28, 2025

Research Interests:

- I love to learn more about the friends in the sea since I want to join the research program at the Cape May Whale Watch Center.
- Or maybe something about the sea environment.
- I especially love taking photos and working with images (e.g. image classification), but I think it will require a lot of knowledge in computer vision.

Objectives/Goals of the Project:

- Identify key features that significantly contribute to accurate prediction of marine species classes using Random Forest classification algorithm in machine learning. Visualize model performance through accuracy and loss graphs.
- Showcase skills in data manipulation, visualization, and analysis using Python.

ABSTRACT

This study uses the Random Forest algorithm, one of the most popular machine learning methods, to predict the presence of marine species based on environmental factors. The research data was collected from the Cape May Whale Watch and Research Center in Cape May, New Jersey, which includes marine megafauna observations from 2012 to 2024. The data consists of both text and numerical information, all of which were analyzed and cleaned using Python libraries. Additionally, the data was encoded to ensure the Random Forest model could be trained effectively. The results of this study aim to identify which environmental factors are

most important for predicting the presence of large marine animals, as shown in the feature importance chart. Model performance was also visualized using graphs of accuracy and loss. Future research will expand the dataset by exploring additional factors and testing alternative machine learning methods to further improve prediction accuracy.

KEYWORDS: Machine Learning, Random Forest, Marine Megafauna

INTRODUCTION

Environmental factors can influence the appearance and distribution of marine species (Mills et al., 2023). For example, the distribution and abundance of marine mammal prey are affected by environmental and temporal factors, such as water conditions, which in turn influence the movement and migration patterns of marine mammals. However, identifying the most important environmental factors from a large amount of collected observational data can take a lot of time for researchers. To simplify this process, this study used the random forest machine learning algorithm (Bruce et al., 2020).

Random forest is a powerful machine learning algorithm that can handle large datasets with many variables. It works by creating many decision trees. A decision tree is a model that splits the data into different groups based on certain conditions and makes a prediction, and the final prediction is made by combining the results from all the trees.

MATERIALS & METHODS

The research data was collected from the Cape May Whale Watch and Research Center in Cape May, New Jersey, and includes marine megafauna observations from 2012 to 2024. Megafauna included marine mammals, sharks, and sea turtles. The data is stored in a comma-separated values (CSV) file, containing personnel, environmental and sea conditions, images and species specific marine animal behaviors. However, only the environmental data was used for analysis and species prediction.

The data was processed and analyzed using Python libraries, including removing missing values and ensuring the integrity of categorical data from predefined values. Additionally, the data was correctly formatted according to its data types, such as floating-point numbers and integers, and any unnecessary white spaces were removed. Finally, the data was encoded and normalized to prepare it for training the Random Forest model.

The number of decision trees selected to train the Random Forest model is shown through a line chart comparing accuracy and log loss. Additionally, the importance of each environmental factor is displayed in two bar charts, sorted from low to high. Factors with less contribution to the algorithm’s predictions are removed, and the model is retrained using only the more important factors, and its accuracy is compared.

RESULTS

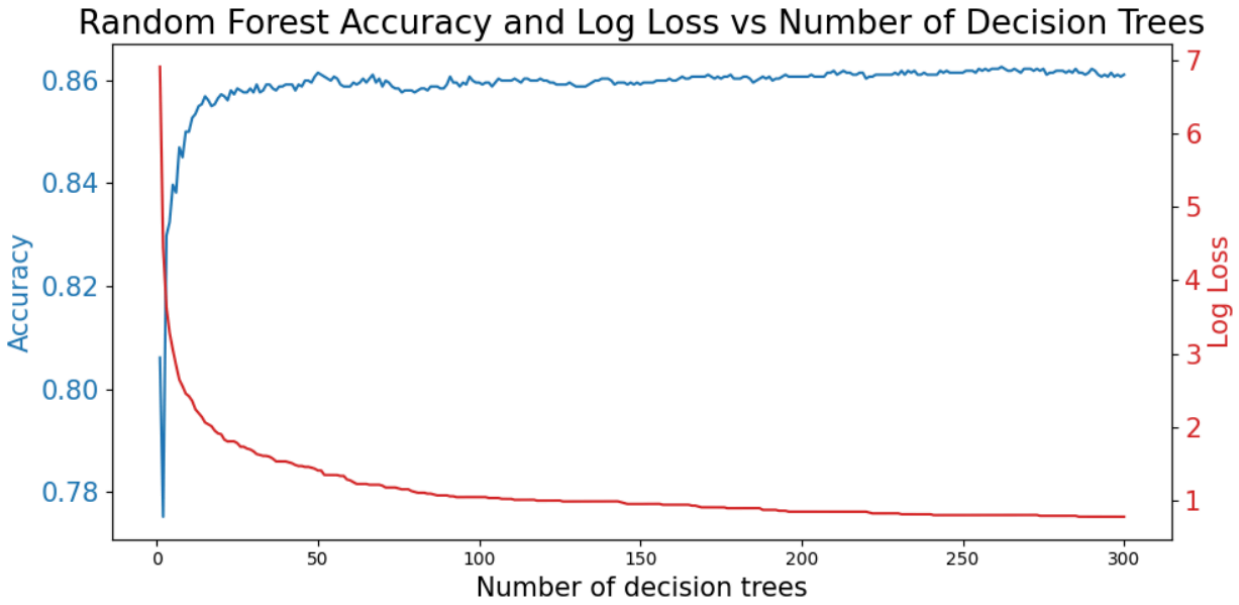

Figure 1: Line chart to identify the ideal number of decision trees for the random forest model by comparing accuracy and log loss

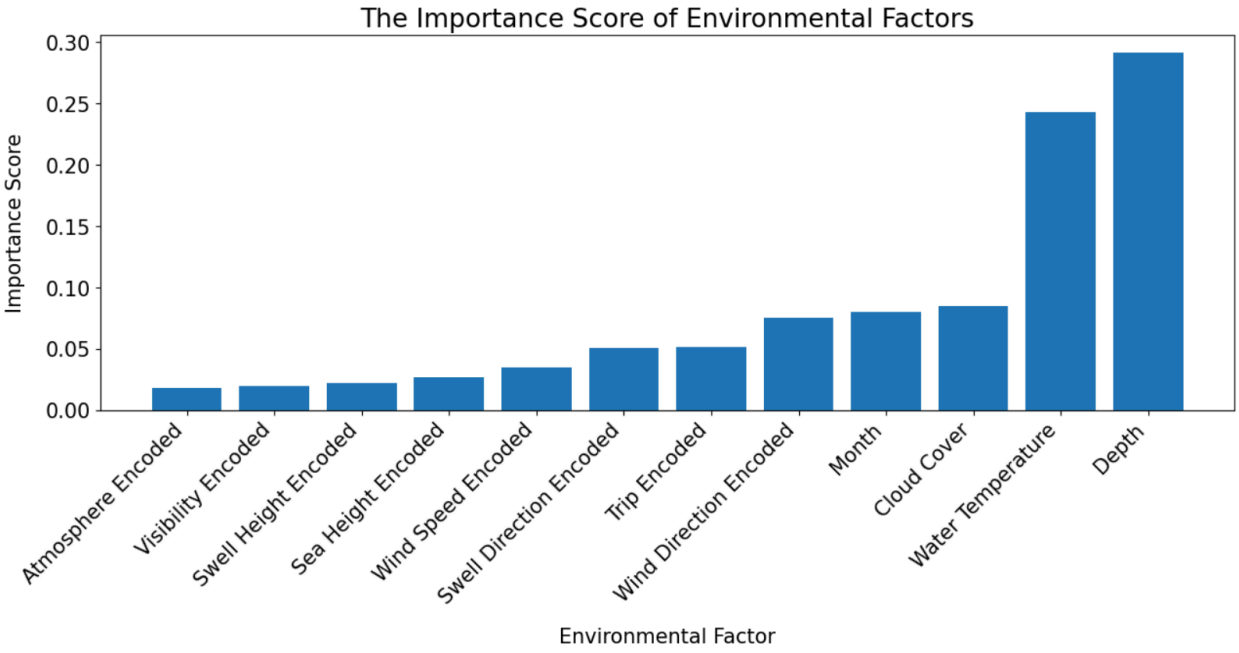

Figure 2: The importance score of all environmental factors, sorted from low to high importance score. Water temperature and sea depth are two factors with the highest importance score.

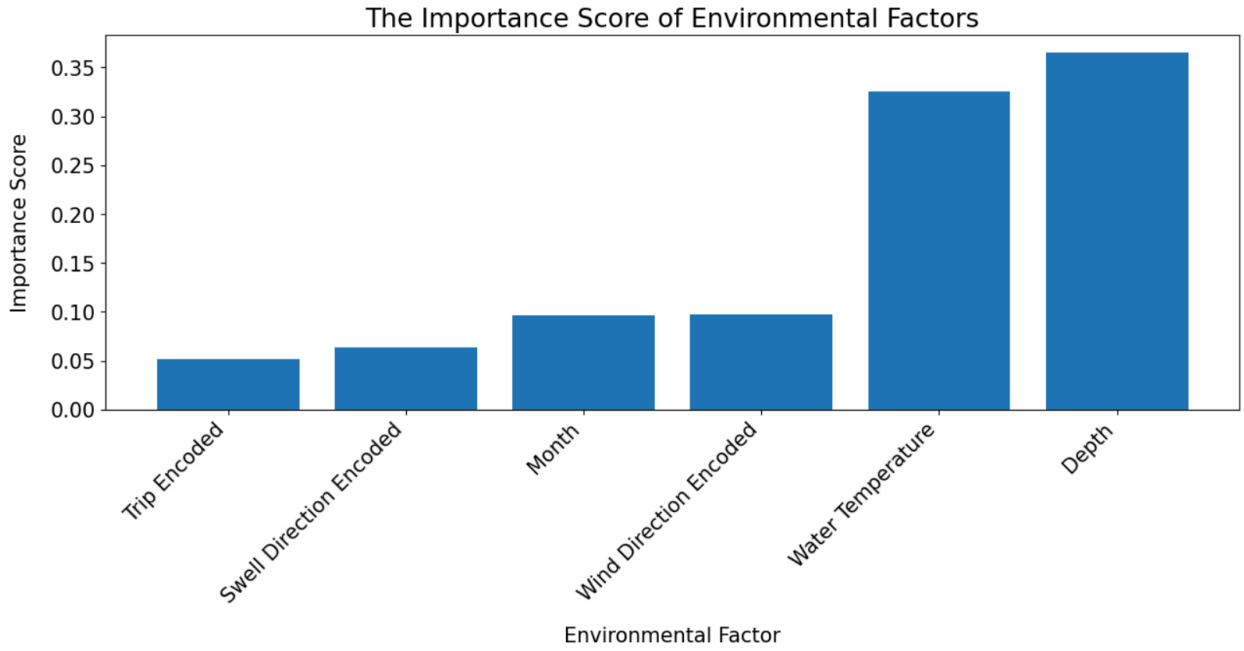

Figure 3: The importance scores of environmental factors, sorted from low to high, after removing environmental factors with low importance scores. Water temperature and sea depth are still the two factors with the highest importance scores

DISCUSSION

The results from this study provide key insights into the importance of various environmental factors in predicting marine species.

Figure 1 illustrates the relationship between the number of decision trees in the random forest model and its performance, represented by accuracy and log loss. As expected, increasing the number of trees led to improved accuracy and reduced log loss. The optimal number of trees for this model was found to be around **150**, where the model achieved a strong balance between accuracy and log loss. Beyond this point, additional trees did not significantly contribute to performance improvement, suggesting that it is sufficient for the model to learn the underlying patterns in the data without overfitting.

Figure 2 focuses on the importance scores of all environmental factors in the dataset, with factors sorted from low to high importance. The model indicates that water temperature and sea depth are the most influential variables in predicting the species. When evaluating the model's performance on the test data, the accuracy reached **87.26%**.

Furthermore, Figure 3 illustrates the impact of removing environmental factors with low importance scores on the model's performance. After eliminating these less influential factors, the model's accuracy slightly decreased to **86.48%**, indicating that the removal of features with low importance score did not have a significant negative effect on model performance. Water

temperature and sea depth remain the most important factors, highlighting their key role in the model's predictions.

By following this approach, researchers can get more accurate predictions about which environmental factors affect marine species. This makes it a valuable tool for studying marine ecosystems and predicting how marine species might respond to changing environmental conditions.

FUTURE DIRECTIONS

Future research will expand the dataset by exploring additional factors and testing alternative machine learning methods to further improve prediction accuracy. Moreover, testing alternative machine learning methods, such as Gradient Boosting or Neural Networks, could further improve prediction accuracy and provide valuable insights into the effectiveness of different algorithms in comparison to Random Forest.

Additionally, to increase the accuracy of species identification, the use of photo-identification techniques can be integrated into the model. By combining this technique with machine learning, future research could achieve better classification and more accurate predictions.

REFERENCES

Bruce, Peter, Andrew Bruce, and Peter Gedeck. 2020. Practical Statistics for Data Scientists: 50 Essential Concepts, 2nd ed. O'Reilly Media.

Mills, K. E., Osborne, E. B., Bell, R. J., Colgan, C. S., Cooley, S. R., Goldstein, M. C., Griffis, R. B., Holsman, K., Jacox, M., & Micheli, F. (2023). Chapter 10: Ocean ecosystems and marine resources. In USGCRP (U.S. Global Change Research Program), *Fifth National Climate Assessment.* https://doi.org/10.7930/NCA5.2023.CH10

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Journal of Machine Learning Research. 12:2825--2830.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my teachers for their invaluable guidance, support, and permission to use the research data. Their expertise and encouragement were essential to the success of this work, and I am truly thankful for their mentorship in coding and writing this paper.