Efficacy of Machine Learning Models in Predicting Ocean pH Levels

Aadyant Maity
With Help from Efthimios Gianitsos

Abstract:

This paper will investigate the efficacy of machine learning models, including Linear Regressor, MLP regressor, Support Vector Machine (SVM), and Random Forest Regressor, in accurately predicting ocean pH levels. By utilizing a comprehensive dataset of oceanographic variables, we evaluate the models' performance on training and development sets. The findings highlight the relative unimportance of eight oceanographic variables in the prediction of ocean pH levels using the machine learning models mentioned above. These insights contribute to a better understanding of ocean acidification impacts and will aid in the development of mitigation strategies because scientists will be able to focus their efforts on the three important variables given by the models.

Introduction:

Ocean acidification, a consequence of escalating carbon dioxide emissions, poses a grave threat to marine ecosystems. Organisms that depend on calcification to build shells for protection (clams, oysters, scallops) are among those that are heavily impacted by ocean acidification. The growth of calcium carbonate slows down and even reverses in severe cases of acidification leading to an unstable marine ecosystem [1]. Accurate prediction of ocean pH levels is crucial for comprehending the magnitude of this phenomenon and developing effective preventative measures. By harnessing a comprehensive dataset compiled by NOAA, comprising numerous oceanographic variables, we were able to train four highly accurate models to predict ocean pH levels. Firstly, our analysis reveals that G2aou and G2talk emerge as the two most important features, emphasizing their critical role in accurately predicting ocean pH levels.

Moreover, the Random Forest Regressor stands out as the most accurate model, achieving an impressive accuracy rate exceeding 98%. These findings underscore the potential of machine learning models in enhancing our data collection process and guiding the formulation of effective mitigation strategies.

Methodology:

Dataset and Preprocessing

The dataset used in our project was obtained from the National Oceanic and Atmospheric Administration (NOAA). It has been compiled over the span of about 50 years with the help of 96 cruises spanning the Earth's oceans. The data is accurate to 0.005 in salinity, 1% in oxygen, 2% in nitrate, 2% in silicate, 2% in phosphate, 4 µmol kg-1 in TCO2, 4 µmol kg-1 in talk [2]. The data was preprocessed to remove any columns that were not directly related to ocean pH levels. Some extraneous columns were: halogenated transient tracers, CCl4, and TCO2. The columns that we used for our project were: temperature (sample temperature), salinity (salts in sample), oxygen (dissolved oxygen in sample), aou (amount of oxygen consumed by biological processes in that area), talk (alkalinity of the sample after all non water particles are filtered out), cfc11 (concentration of chlorofluorocarbon-11 in sample), cfc12 (concentration of chlorofluorocarbon-11 in sample), phosphate (concentration of phosphate in sample), pcfc12 (partial pressure of chlorofluorocarbon-12 in sample), nitrate (concentration of sample), silicate (concentration of silicate in sample), and phtsinsitutp (pH of sample in measured temperature and pressure, value being predicted) [2]. After we preprocessed our data to only contain the columns mentioned above, we had to delete any row that had an empty cell in any of the columns. These preprocessing steps resulted in data being cut from approximately 50,000 rows

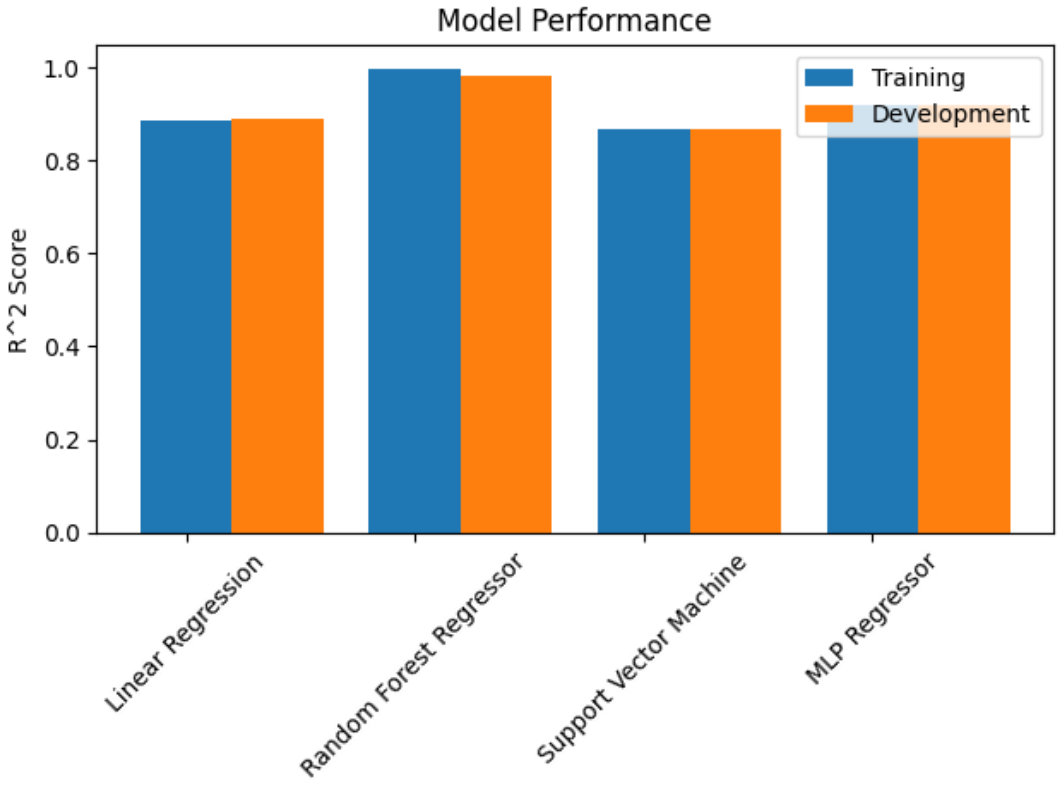
to 25,000 rows. Additionally, we drew the hypothesis that the talk column would be the most important feature in the prediction of the phtsinsitutp column [3].

Models

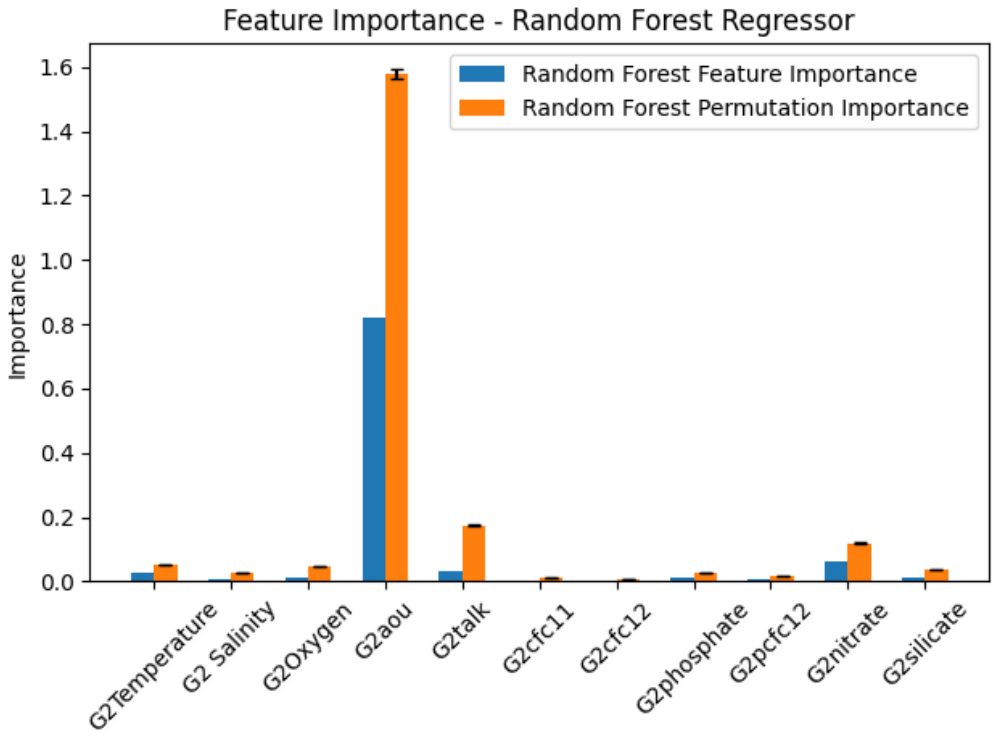
The data was split into two sets: training and development. 90% of the data went into the training set and the remaining 10% went into the development set. The training data was used to train the models to effectively predict the phtsinsitutp column based on the ten columns of data outlined above. The training data would be used in the algorithm and would produce results based on the statistical significance of the training data in the model training process. The preprocessing was handled by the split_data() and preprocess() functions. The analysis showed us the importance of each variable in the prediction of the phtsinsitutp column and gave us valuable knowledge on the necessity of relatively unimportant features when it came to model accuracy. We used Linear Regressor, Random Forest Regressor, MLP Regressor, and Support Vector Machine to generate our results. The function gave us training and development data R^2 values for each model. We also calculated the Random Forest Feature and Permutation Importances for each variable that we used in our training data.

Results:

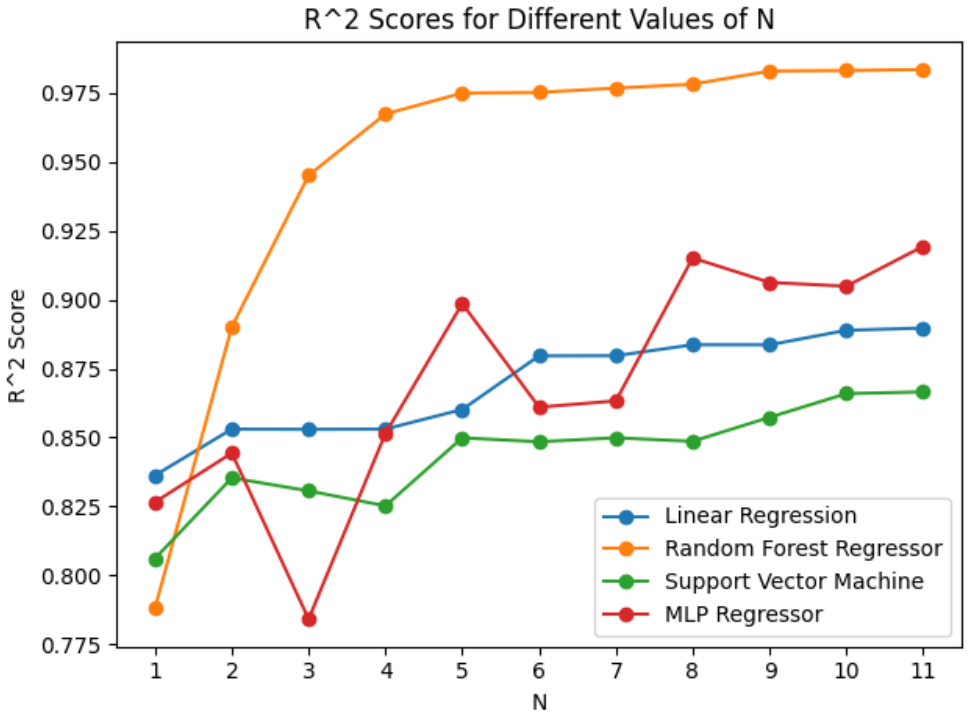
The table below shows us the accuracy of each model in the development (orange) and training (blue) stages. We can see that the Random Forest Regressor was the most accurate model in both stages. This is important for our project also because the Random Forest Regressor is unique in its ability to natively rank features.



The feature importance table shows us that of the 11 oceanographic variables used in our project, eight variables have close to no importance in the prediction of ocean pH levels. We see that G2aou, G2talk, and G2nitrate have relatively high feature and permutation importances. It is also important to note that the permutation importance data has not been normalized to sum up to 1 because it is purely based on relative scale unlike the feature importance data. Additionally, each side of the error bars represent one standard deviation.



The table below shows us how the models' accuracies change as we add the N most important features as ranked in the feature importance table. We can see that only the Random Forest Regressor has a monotonic curve as N is increased. This is important to note as it supports our decision to use it as the standard when determining the importance of the features in our dataset because it shows its reliability in detecting patterns that are not visible to humans.



Conclusion:

In this study, we investigated the efficacy of machine learning models in predicting ocean pH levels using a comprehensive dataset of oceanographic variables. The findings from our analysis offer novel insights into the relative importance of these variables in accurately predicting ocean pH levels and highlight the potential of machine learning models for addressing the pressing challenges of ocean acidification. Beyond the immediate implications for ocean acidification research, our study contributes to the broader field of environmental data analysis. By demonstrating the effectiveness of machine learning models in processing vast and complex oceanographic datasets, we highlight the potential of these models for advancing other areas of environmental science. However, it is essential to acknowledge the limitations of this research. While our findings provide valuable insights, the models' predictions are still influenced by the quality and comprehensiveness of the dataset used. Therefore, continuous efforts to improve

data collection and integration will further enhance the accuracy and reliability of the machine learning models in predicting ocean pH levels. Finally, our study demonstrates the potential of machine learning models as valuable tools in understanding and addressing ocean acidification. The identification of critical oceanographic variables and the impressive accuracy achieved by the Random Forest Regressor pave the way for more focused and effective conservation strategies. As we continue to face the challenges posed by climate change, this research contributes to the growing body of knowledge aimed at safeguarding marine ecosystems and preserving the delicate balance of our oceans. The combination of advanced data analysis techniques with environmental research holds great promise for informing evidence-based policies and actions to protect our planet's most valuable natural resources.

Acknowledgements and References:

[1] - Environmental Protection Agency. (n.d.). *Effects of Ocean and Coastal Acidification on Marine Life*. EPA. https://www.epa.gov/ocean-acidification/effects-ocean-and-coastal-acidification-marine-life

[2] - Lauvset, Siv K.; Lange, Nico; Tanhua, Toste; Bittig, Henry C.; Olsen, Are; Kozyr, Alex; Alin, Simone R.; Álvarez, Marta; Azetsu-Scott, Kumiko; Barbero, Leticia; Becker, Susan; Brown, Peter J.; Carter, Brendan R.; Cotrim da Cunha, Leticia; Feely, Richard A.; Hoppema, Mario; Humphreys, Matthew P.; Ishii, Masao; Jeansson, Emil; Jiang, Li-Qing; Jones, Steve D.; Lo Monaco, Claire; Murata, Akihiko; Müller, Jens Daniel; Pérez, Fiz F.; Pfeil, Benjamin; Schirnick, Carsten; Steinfeldt, Reiner; Suzuki, Toru; Tilbrook, Bronte; Ulfsbo, Adam; Velo, Antón; Woosley, Ryan J.; Key, Robert M. (2022). Global Ocean Data Analysis Project version 2.2022 (GLODAPv2.2022) (NCEI Accession 0257247). [indicate subset used]. NOAA National Centers for Environmental Information. Dataset. https://doi.org/10.25921/1f4w-0t92. Accessed 6/26/23

[3] - Orenda Technologies. (2023, May 18). *Total alkalinity vs. ph, and their roles in Water Chemistry*. Blog. https://blog.orendatech.com/total-alkalinity-role-water-chemistry

Dataset - Index of /data/oceans/ncei/ocads/data/0257247 (noaa.gov)