Analysis of Oceanic Oxygen Concentration and Influencing Factors

Saiem Ilyas, Riyaan Ahmed, Trisha Bhatnagar, Yoatam Gebremicael

Introduction

The oxygen content in seawater is extremely crucial for the health of the ocean and marine organisms.. It also influences biogeochemical cycles and Earth's climate. In this study, we are interested in determining how oxygen levels in the ocean are influenced by temperature, salinity, and depth.. With the given dataset, we hope to determine the key factors that impact oxygen levels and develop models to help us further understand these factors.

Data Analysis

Data Cleaning:

- **Handling Missing Values:** Missing data were addressed by replacing missing values in the **CTDTEMP**, **CTDSAL**, and **depth** columns with their column means. **micromoles_of_oxygen_per_unit_mass_in_sea_water** column, was filled with the mean of the same column.
- **Outliers**: Outliers were removed using the zscore method. Those greater than 3 standard deviations from the mean were classified as outliers and were removed from **best_Oxygen**, **micromoles_of_oxygen_per_unit_mass_in_sea_water**, **CTDTEMP**, and **CTDSAL** columns.

Exploratory Data Analysis (EDA):

- **Hypothesis Testing:** To test the correlation between the concentration of oxygen and temperature, pearson's correlation was used. The test resulted in a correlation coefficient of -0.38 with a pvalue of 0, showing a moderate negative correlation between the two variables, showing that as temperature increases, oxygen concentration decreases.

- **Data Transformation:** StandardScaler was used to normalize the dataset to improve the efficiency of our models. The transformation normalized feature range, making the machine learning model more effective.

Visualization

- **Pairplots and Correlation Matrix**: Used to analyze relationships between variables, showing that depth had a strong negative correlation with oxygen concentration.
- **Scatter Plots**: Showed a clear inverse correlation between depth and oxygen, supporting the idea that depth plays an essential role in oxygen levels.
- **Time-Series Plot**: Analyzed how oxygen levels changed over time, which could hint at seasonal patterns. The time series plot showed a lot of variation in the oxygen levels over time, with a sharp increase from around 2021-11 to 2022-06.

Model Development

Random Forest Regression

The Random Forest model was initially trained on default parameters and resulted in a strong R^2 value of 0.8497. After hyperparameter tuning, the final model resulted in an R^2 value of 0.8222 and a Mean Absolute Error (MAE) of 0.5202. Depth was the most important feature, followed by salinity and temperature.

The initial R^2 value(0.8497) was high, showing good performance. However, high scores are sometimes a sign of overfitting, where the model performs well on training sets but struggles with new, unseen sets. Hyperparameter tuning was used to improve generalization, leading to a slight drop in the R^2 value(0.8222). The decrease is a trade-off because now the model will perform better on different subsets of data, so it is more stable and reliable in practical use.

Linear Regression

A linear regression model was also used, but it did not work well, with R^2 value of -0.6235. This shows the relationship between features and oxygen concentration is non-linear, making Random Forest a better fit for this problem.

Cross Validation

To ensure the Random Forest model was stable, 5-fold cross-validation was performed, resulting in a mean score of 0.8471. This process helped approximate the model's performance over different subsets of data, ensuring that our results would generalize well to unseen data. It gives us a good balance between computational expense and model stability.

Main Findings

- Our analysis confirmed that depth has the most significant impact on oxygen concentration. As depth increases, oxygen levels decrease, likely due to reduced light penetration, lower photosynthesis, and limited mixing between water layers.

- While salinity and temperature do impact oxygen solubility in seawater, their impact is less than depth.

- The Random Forest model, which can handle non-linear relationships, outperformed the linear regression model, which suffered from multicollinearity and the issue of the data being non-linear.

Drawbacks and Limitations

- The dataset may suffer from missing factors, like ocean currents or biological activity, that can influence oxygen levels.

- The analysis did not take into account how oxygen was distributed over various longitudes and latitudes since it was challenging to merge the data by these two variables. Future analysis could fix this by using improved techniques for merging the data or by collecting more data.

- Although cross validation was used, more testing on unseen data would help strengthen the model's reliability and reduce the risk of overfitting.

- Although oxygen levels over time were analyzed with a time series plot, more insight into seasonality and long-term trends in oxygen levels is needed.

Conclusion

This study was able to conclude that depth is the most essential variable on ocean oxygen concentration, followed by temperature and salinity with lesser influences. The Random Forest model had great predictive capability and performed better than linear regression due to the non-linear nature of the data. Follow-up research could build on this study by adding additional variables, such as ocean currents and biological processes, to enable a more in-depth understanding of the dynamics involved in oxygen concentration in seawater.