Final Analysis
Problem Statement:

Ocean salinity is a crucial variable in understanding Earth's climate systems, yet it is becoming increasingly difficult to predict due to the compounding effects of rising global temperatures. Warmer climates accelerate ice melt, altering precipitation-evaporation patterns, leading to significant changes in ocean salinity that disrupt marine ecosystems, ocean currents, and global weather patterns. Accurately predicting salinity levels is essential to understanding and mitigating the acceleration of global warming that drives extreme climate events, but the complexity and variability of salinity, salt levels, make this a challenging task.

To address this challenge by utilizing various machine learning approaches—such as linear regression, random forest regression, KNNeighbors, and decision tree models—it is possible to analyze large datasets of salinity and temperature, thus being able to predict ice melt and evaporation trends. Developing such models would enable the prediction of salinity levels under different climate scenarios, providing valuable insights into the potential impacts on ocean currents and climate stability. This project not only demonstrates the application of coding and data analysis to real-world problems but also contributes to the broader effort to understand and adapt to the impacts of rising global temperatures. I will predict Temperature with Salinity, in hopes that this can be used as a way to predict future salt levels, ultimately giving us some type of idea of the future of climate change given that increasing sea temperatures cause increased global warming.

Dataset Description: This data set is from Kaggle in CSV format. The original dataset had numerous unnecessary columns that I erased in order to solely focus on Salinity and Temperature and depth, for a brief moment. The information provided on the data set is from a California Oceanic Investigation department, it includes up to 50,000 samples of data from the 1900s- present. This data set will be very insightful on how predictive models can set the stage of future climate effects.

Exploratory Data Analysis: The minimum temperature of the dataset is 1.44 degrees celsius, with a max of 31.14. This shows us the difference in temperature that can occur in the region, California. Salinity ranges from 28.431-37.034. According to ATI North America, 37 Salinity is considered very high, this could be in correlation to the high water temperature seen above, we will further investigate this idea. I also sorted and dropped columns just to have a better idea of the numbers I was working with to see what I should expect. The mean and median temperature are about the same, the same goes for Salinity. This tells us that the data for both are evenly distributed and symmetrical.

Initial Visualizations:
The Histplot of temperature gives us perspective on how temperature ranges. Comparing it to the Salinity histplot, we see that most of the Salinity values are between 32 and 35, giving us insight on how high the Salinity is in the region. With this information, it makes sense that there

are usually higher temperatures, but there is something off. The temperature is usually below 20 degrees according to the data, this tells us that the correlation between Salinity and temperature could be different than what studies have suggested, the idea that high salinity=high temperature. With the Salinity vs Temperature scatter plot, we see that all of the temperature values result in around 32-34 salinity. We also see a slight increase in temperature as salinity increases, the cluster curves slightly. Comparing Salinity and Depth comes to mind when thinking about the ocean, it is not all at the same depth. There is some correlation between Salinity and Depth. According to the data, at peak depth salinity is around 35 and depth is lower with salinity below 33 and above about 35.8. This suggests that depth and salinity has some correlation, but we are not investigating this aspect, but this would be a good aspect to investigate to connect to global warming; however, we can't control the oceans depth and increased temperature will cause increased sea levels, so we have to figure out the temperature aspect first to approach this new idea.

Modeling and Interpretations:

Linear Regression: The analysis of the linear regression model reveals that salinity has a notable relationship with temperature, with a coefficient of -2.1347. This coefficient indicates that for every one-unit increase in salinity, the model predicts a decrease in temperature by 2.1347°C. This negative relationship suggests that higher salinity levels are associated with lower water temperatures, supporting the hypothesis that salinity might influence temperature in a way that could slow the impacts of global warming. However, the underlying complexity of this relationship is not fully captured, as indicated by the model's performance metrics.

The Mean Squared Error (MSE) of the model is 13.29, which represents an improvement over the baseline MSE of 17.789. The reduction in MSE signifies that the model is a significant step forward, indicating that salinity has some predictive power over temperature. Nevertheless, the MSE of 13.29 still reflects an average error of 3.65°C in predictions, suggesting that the model is not perfect. Given the context that increased salinity can contribute to coral reef deterioration—an important issue in the fight against climate change—the error margin of 3.65°C is significant. This level of error means that predictions based on this model could still be misleading, causing scientific efforts to appear more effective than they actually are. The lack of precision in the model could result in false confidence in strategies aimed at mitigating global warming, potentially hindering progress.

The correlation coefficient (r) of 0.51 further suggests a moderate positive linear relationship between salinity and temperature. While the relationship is statistically significant, it is not strong enough to offer precise predictions, especially with the 3.65°C error margin. This moderate correlation suggests that while salinity does have an impact on temperature, other factors likely play a more substantial role in determining oceanic temperatures.

In conclusion, the model indicates that increased salinity is associated with a decrease in temperature, offering some insight into potential ways to mitigate global warming. However, the 3.65°C error and the moderate correlation suggest that there are other significant factors influencing temperature that the model does not account for. To better inform global warming efforts, further refinement of the model and the inclusion of additional variables are necessary to reduce prediction error and improve accuracy. Without these improvements, the effectiveness of climate mitigation strategies based on this model may be overstated.

The K-Nearest Neighbors model:

The K-Nearest Neighbors (KNN) has MSE values of 10.13993 on the training data and 10.3215 on the testing data. The corresponding error fluctuations are approximately 3.18°C for the training set and 3.22°C for the testing set. These values indicate that the model is, on average, off by about 3.2°C when predicting the temperature from salinity. The MSE values on both the training and testing sets are quite close to each other, suggesting that the model is generalizing well. There is minimal overfitting, as evidenced by the small difference between training and testing errors. This indicates that the model is not overly complex and does not memorize the training data, but instead captures relationships between the features and the target variable. With an error of around 3.2°C, the model’s predictions are off by about 9% of the total range of temperature (which spans from 1°C to 34°C). Depending on the application, this level of error would be considered acceptable. The analysis indicates that salinity can predict temperature to a reasonable extent, but it is not the sole factor influencing temperature. The KNN regression model was able to predict temperature with an error of approximately 3.2°C on both training and testing data, showing that salinity provides a meaningful relationship with temperature. However, the 3.2°C prediction error suggests that while salinity is a strong predictor, it does not account for all the variability in temperature.

Salinity appears to capture some of the main trends in temperature, as the model's error is relatively small, particularly when compared to the baseline model that simply predicts the mean temperature. This suggests that salinity contributes to the temperature variations, but there are likely other environmental factors or variables that influence temperature that are not captured in the model.

The fact that the model performs similarly on both the training and testing sets, without significant overfitting, further implies that salinity is a significant predictor of temperature, but that its influence is not exhaustive. Other variables, possibly including atmospheric pressure, humidity, or geographical factors, could be influencing temperature in ways that the model is not able to capture with just salinity.

With an r value of .885, there is a strong positive linear relationship between Salinity and Temperature.

In conclusion, while salinity can explain a portion of the temperature variations, its predictive power is limited. The model’s performance highlights the need for additional data and factors to improve accuracy. Incorporating other features or exploring more advanced models that account for multiple influences on temperature could improve predictions and better capture the complexity of temperature dynamics.

Decision Tree: The graph displaying tree depth versus accuracy indicates some important insights into the decision tree model's performance. As the tree depth increases, both the training accuracy and test accuracy decrease, which is unusual because typically, increasing depth allows the model to better fit the training data. This suggests that the model might be struggling with overfitting or underfitting. The decrease in accuracy could also point to issues in the data, such as noise or the absence of key features, which cause the model to perform poorly as its complexity grows.

The Mean Squared Error (MSE) values of 10.28 for training and 10.41 for testing indicate that the model is performing reasonably well, as the MSE is fairly consistent between the training and test sets. The small difference between the training and test MSE values suggests that the model is not overfitting, and its performance generalizes well across unseen data. However, the MSE values also imply that the model’s predictions are still off by an average of around 3.2 degrees. This may not be ideal for certain applications, especially when high precision is required.

Additionally, the correlation coefficient (r) of 0.85 indicates a strong positive linear relationship between salinity and temperature. This suggests that the model is capturing significant patterns in the data, with salinity being a good predictor of temperature. However, despite the high r-value, the increasing depth and the consistent MSE values show that there might be room for improvement in the model. The goal should be to reduce the MSE while maintaining or improving the correlation between the features and target variable.

Random Forest:

The Random Forest model’s results suggest that salinity has a moderate relationship with temperature, but there is room for improvement. The MSE of 9.492 indicates an average prediction error of about 3.08°C, which is significant for accurate climate predictions. The R value of 0.685 shows a moderate positive linear relationship, meaning salinity is a meaningful predictor but not the sole factor influencing temperature. The R^2 value of 0.469 means that nearly 54.1% of temperature variation is unexplained, highlighting the complexity of predicting ocean temperature.

Given the amount of data, I am unable to use multiple parameters when grid searching for runtime reasons, this would help the model be more accurate. With the parameters given, MSE stayed about the same while r increased to over .90, this suggests that the parameters and gridsearch allowed the model to find a stronger positive linear relationship. This suggests that the model is now more precise.

Salinity’s feature importance of 1 confirms it as the most influential factor in the model, but the moderate R^2 and MSE suggest that other environmental factors are also important. While the model demonstrates that salinity can predict temperature to some degree, it doesn't fully capture the broader dynamics of ocean temperature and climate change. More variables need to be included to enhance prediction accuracy and better inform climate change strategies.

Exploratory Data Analysis (EDA):

The dataset reveals that temperature ranges from 1.44°C to 31.14°C, and salinity varies between 28.431 and 37.034. It was observed that the data is symmetrically distributed, with similar values for the mean and median of both temperature and salinity. Initial visualizations, including histograms and scatter plots, highlighted a moderate positive relationship between salinity and temperature. However, while the scatter plot showed a slight increase in temperature as salinity increases, the correlation was not as strong as expected.

Modeling and Results:

1. Linear Regression:
   - The linear regression model showed a negative relationship between salinity and temperature with a coefficient of -2.1347, suggesting that as salinity increases, temperature decreases.
   - The Mean Squared Error (MSE) was 13.29, indicating an average error of about 3.65°C in predictions. This error margin is significant and reduces the accuracy of the model for precise climate predictions.
   - The correlation coefficient (r) of 0.51 suggested a moderate relationship, meaning salinity is a meaningful predictor, but other factors likely influence ocean temperature.
2. K-Nearest Neighbors (KNN):
   - The KNN model showed reasonable performance, with MSE values of 10.14 on the training set and 10.32on the test set. The error margin of around 3.2°C indicates a moderate prediction accuracy.
   - The r value of 0.885 indicates a strong positive correlation between salinity and temperature, but the model still has limitations, as it doesn’t account for all variables influencing temperature.
3. Decision Tree:
   - The decision tree model exhibited some unusual behavior with decreasing accuracy as tree depth increased, possibly indicating overfitting or underfitting.
   - The MSE values of 10.28 (training) and 10.41 (testing) were close, suggesting minimal overfitting.
   - The correlation coefficient (r) of 0.85 shows a strong relationship between salinity and temperature, but further refinement is needed to improve precision.
4. Random Forest:
   - The Random Forest model showed a moderate relationship with salinity having a significant influence on temperature predictions. The MSE of 9.492 indicates an average error of 3.08°C, suggesting that while the model is useful, it still has room for improvement.
   - The r value of 0.685 suggests that salinity is an important factor, but the model fails to capture a substantial amount of temperature variation (R^2 value of 0.469).

    - Grid search was employed to optimize the model, which resulted in a stronger linear relationship (r > 0.90) and greater precision, but still did not fully account for all environmental factors.

Key Findings:

- Salinity is a meaningful predictor of ocean temperature, with salinity's feature importance being the highest among all features in the Random Forest model. However, temperature predictions are still significantly impacted by unexplained variance.
- The models show that while salinity and temperature are correlated, other environmental variables (e.g., atmospheric pressure, humidity, depth, etc.) are also likely influencing temperature.
- Despite the models' ability to predict temperature reasonably well, the error margin of approximately 3°C suggests that they are not sufficiently accurate for precise climate predictions.
- The analysis points to the need for additional features and refinement in modeling approaches to reduce error and better capture the complexity of the relationship between salinity and temperature.

Next Steps and Recommendations:

- Incorporate additional environmental variables (e.g., atmospheric pressure, humidity, ocean depth, geographical factors) to improve the accuracy of predictions and capture a more comprehensive picture of oceanic temperature dynamics.
- Expand the dataset to include data from other oceans to reduce regional biases and improve the model’s generalizability.
- Explore more advanced machine learning models or techniques to better account for non-linear relationships and multiple variables influencing temperature.
- Refine hyperparameters and perform further grid search to optimize the performance of existing models, especially the Random Forest model, which showed potential after parameter adjustments.

This analysis demonstrates that salinity can be used to predict ocean temperature to some degree, but further improvements in data collection, feature selection, and modeling techniques are necessary to make more accurate predictions.