Final Report

## CSEN 140 Data Mining

##### By Team 10: Ephraim Esson w1644482, Kieran Greeley 1634712, Luis Zamudio 1643926

## Abstract

This project explores the application of machine learning models to predict wave quality for surfers, using environmental data from the Sea Forecast and Waves Classification dataset. After consolidating data from six sources, we experimented with ten regression models, tuning hyperparameters such as training size, max depth, and number of estimators. Evaluation was based on RMSE and R² across training and validation sets. Our results suggest that the Decision Tree model achieved the highest performance. However, this may be due to overfitting. This work has potential implications for real-time surf forecasting using lightweight, interpretable models.

## Introduction

Predicting surf conditions is a long-standing challenge for surfers. Despite technological advances, existing forecasting tools often lack reliability and precision. Machine learning offers an opportunity to improve predictive accuracy by leveraging large-scale environmental data. This project investigates whether a machine learning model can reliably predict surf quality using multiple relevant factors, such as tide, wind, swell characteristics, and lunar phase. Our approach involved aggregating heterogeneous datasets, training a range of regression models, and comparing their performance to identify the most effective predictive method.

## Background / Related Work

Prior research in oceanographic modeling has increasingly turned toward machine learning methods. One relevant study by Li et al. (2024) proposed a predictive framework called Orca, which uses large language models (LLMs) in combination with real-time buoy data to estimate significant wave height. Orca significantly outperformed traditional methods, such as the Global Wave Database, in both speed and accuracy (MSE: 0.0838 vs. 0.2). While our goal differs slightly in that we target surf quality rather than wave height alone, the success of such systems demonstrates the feasibility of data-driven approaches in ocean prediction tasks. This informed our decision to explore a wide range of regression models and assess their real-world viability.

## Approach

To begin, we consolidated data from six CSV files:

* beach.csv – Identifies beach locations.
* spot.csv – Specifies data collection spots within beaches.
* day\_forecast.csv / hour\_forecast.csv – Provide weather and wave predictions on daily and hourly scales.
* tide.csv – Contains tide conditions over time.
* sea\_condition\_fact.csv – Capture environmental attributes.

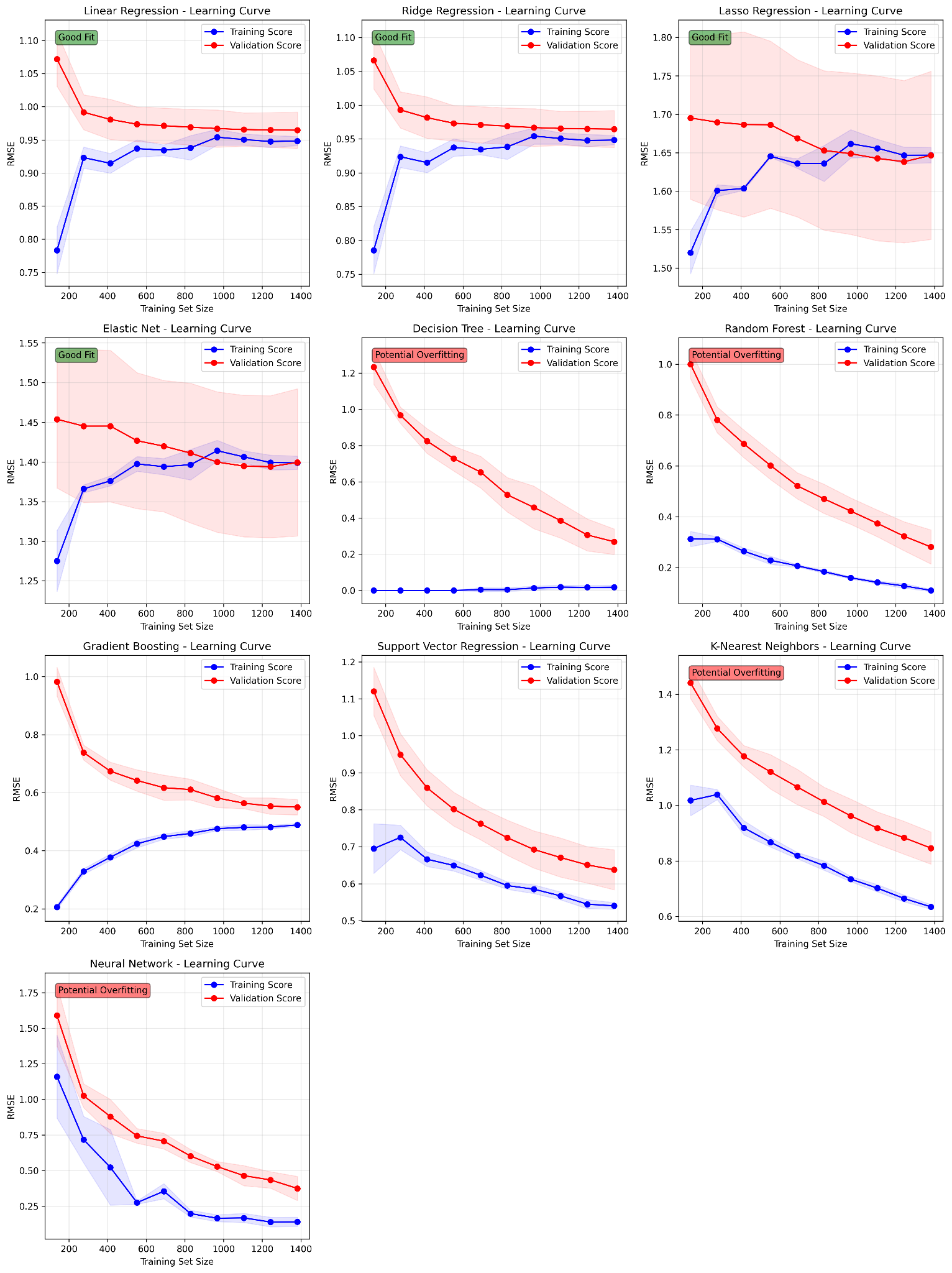
Using the provided model.png schema, we joined these files into a single structured dataset. Preprocessing steps included removing irrelevant columns, converting timestamps into numerical values, and applying one-hot encoding to categorical features. The final result was a clean, feature-rich dataset suitable for regression analysis.

We selected ten models for evaluation: Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Decision Tree, Random Forest, Gradient Boosting, Support Vector Regression, K-Nearest Neighbors, and Neural Network. Each model was trained on an 80/20 train-test split. Hyperparameter tuning was model-specific; for example, we tested different values for alpha in regularized regressions and varied the number of estimators and maximum depths for tree-based models. All models were implemented using scikit-learn and trained in a consistent environment to allow fair comparison.

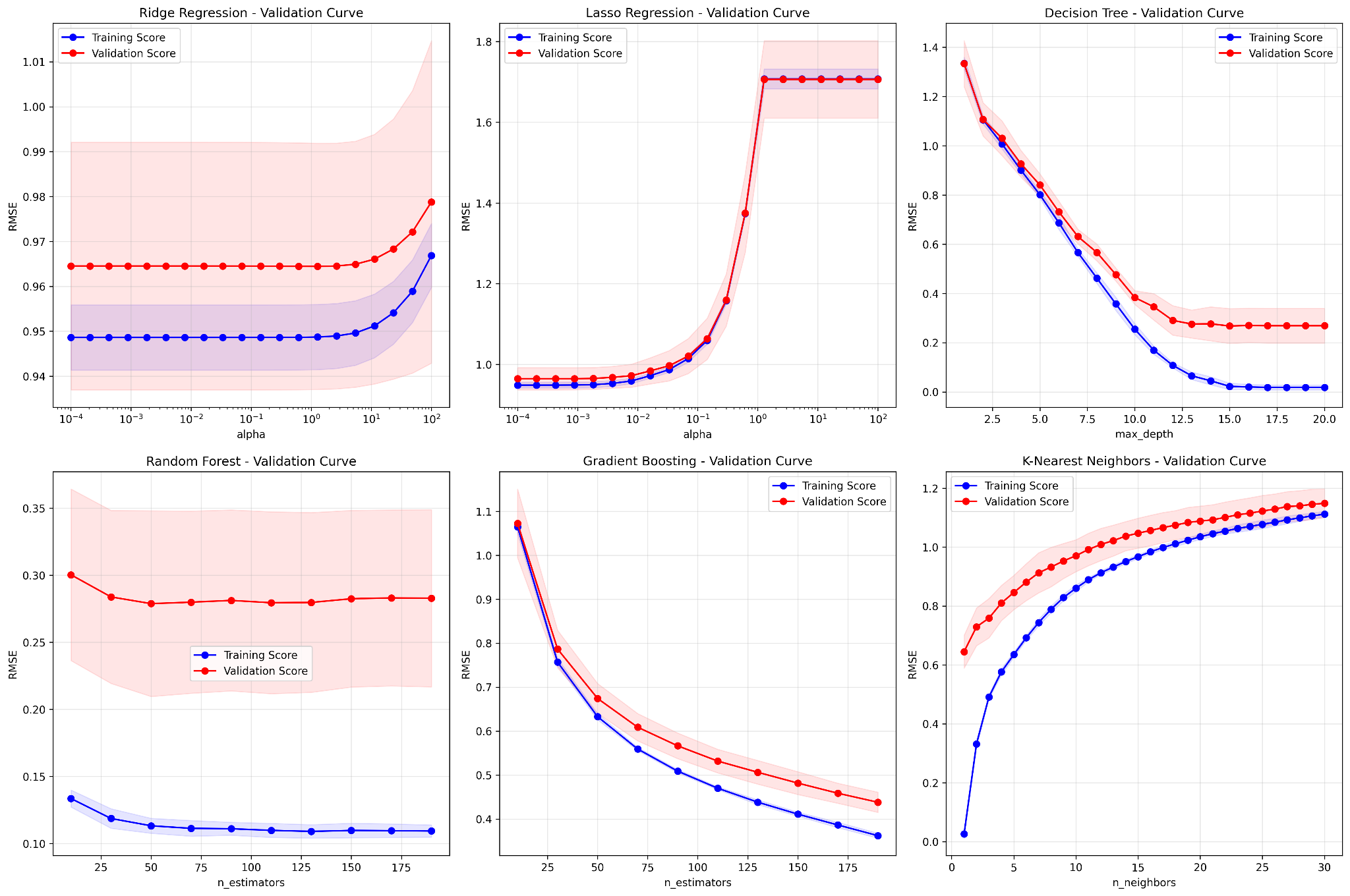
## Experiments

To evaluate each model, we measured performance using Root Mean Square Error (RMSE) and R² score on the validation set. We also varied the size of the training data to examine learning behavior and overfitting tendencies. Figures 1 and 2 show RMSE trends, across the training set size and model’s complexity, respectively.

*Figure 1.*

****

*Figure 2.*

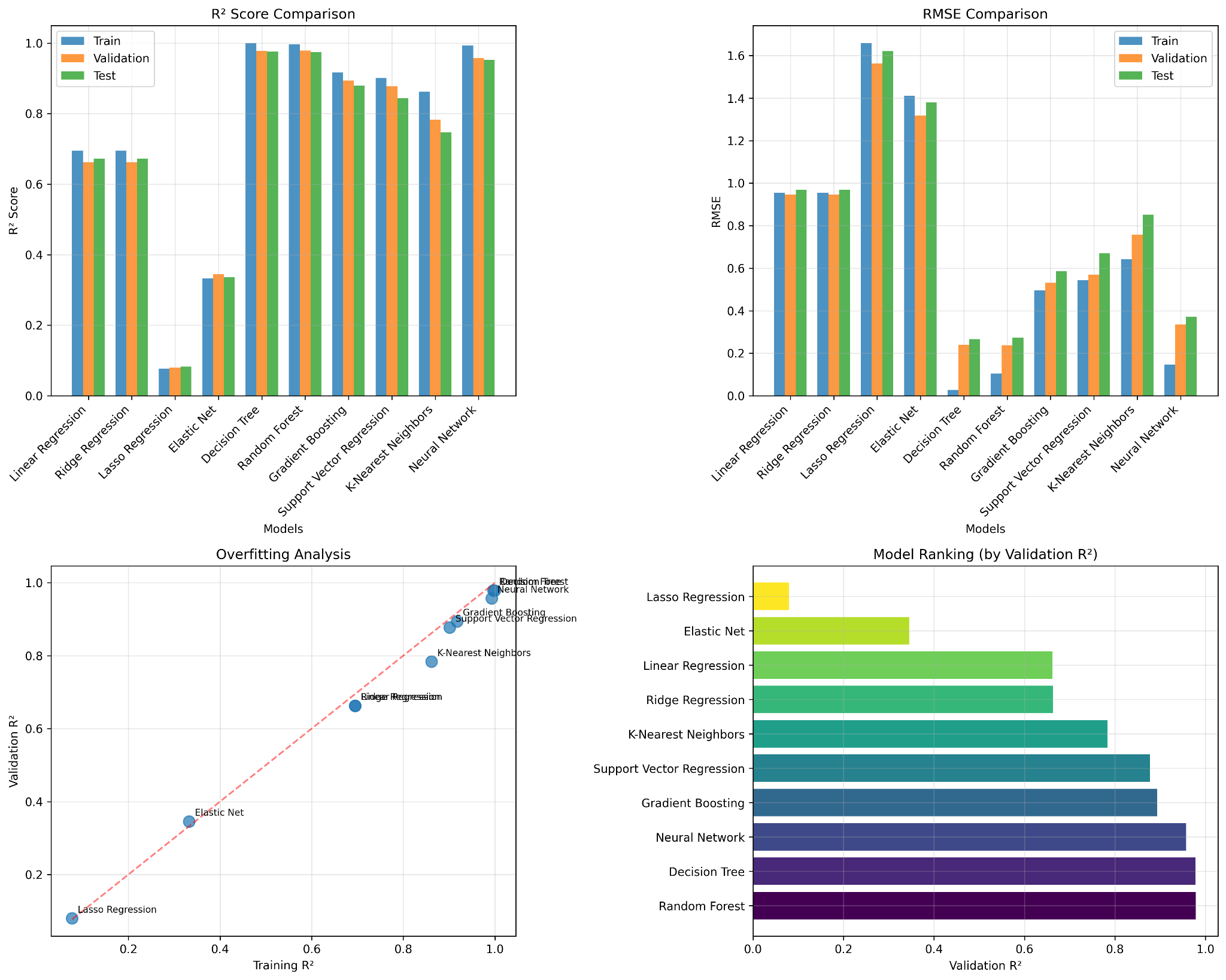
****

In addition to those visualizations, we provide a comprehensive performance table (Table 1) and figure 3 that summarizes validation RMSE and R² for all models. This table and figure serves as key references for identifying the most effective models.

*Table 1.*

| Model | Train R² | Val R² | Test R² | Train RMSE | Val RMSE | Test RMSE |
| --- | --- | --- | --- | --- | --- | --- |
| Linear Regression | 0.6899 | 0.7039 | 0.6414 | 0.9531 | 0.9085 | 1.0006 |
| Ridge Regression | 0.6899 | 0.7044 | 0.6426 | 0.9532 | 0.9076 | 0.9989 |
| Lasso Regression | 0.0709 | 0.0748 | 0.0685 | 1.6499 | 1.6059 |  |
| Elastic Net | 0.3288 | 0.3484 | 0.3302 | 1.4023 | 1.3477 | 1.3675 |
| Decision Tree | 0.9999 | 0.9963 | 0.9983 | 0.0169 | 0.1013 | 0.068 |
| Random Forest | 0.9983 | 0.9918 | 0.9894 | 0.0703 | 0.1511 | 0.1717 |
| Gradient Boosting | 0.9142 | 0.8997 | 0.8738 | 0.5013 | 0.5289 | 0.5936 |
| Support Vector Regression | 0.9072 | 0.8924 | 0.7936 | 0.5216 | 0.5478 | 0.7591 |
| K-Nearest Neighbors | 0.8852 | 0.8207 | 0.6928 | 0.5799 | 0.707 | 0.9262 |
| NeuralNetwork | 0.9988 | 0.9902 | 0.9894 | 0.0583 | 0.1656 | 0.1718 |

*Figure 3.*

****

Tree-based models such as Decision Tree, Random Forest, and Gradient Boosting consistently outperformed linear methods. However, we observed signs of overfitting in some cases, particularly with deeper trees and higher estimator counts. For instance, while Gradient Boosting showed high training accuracy, its validation performance plateaued and eventually declined, indicating over-complexity.

The Random Forest model achieved the best trade-off between accuracy and generalizability. We then focused on fine-tuning this model further. Figure 2 illustrates how model accuracy varied with the number of estimators. Performance gains diminished beyond approximately 50 estimators, and computational cost increased. A similar analysis for maximum tree depth found that values between 10 and 14 yielded the most stable results. We selected 35 estimators and a maximum depth of 12 as the final configuration.

## Conclusion

This project demonstrates that Random Forest Regression is a highly effective model for predicting surf quality using environmental data. While more complex models such as neural networks and gradient boosting offered competitive performance, the simplicity and interpretability of tree-based methods make them attractive for practical deployment. A key insight is that simpler models can often match or exceed the performance of more resource-intensive alternatives when tuned appropriately.

Future work may involve incorporating real-time buoy data and additional geographic features such as seabed topology. These factors are known to affect how swells break and could enhance model accuracy. A visual tool that combines predictions with surf maps could be a valuable application for surfers worldwide.

## References

Li, Z., Xu, R., Hu, J., Peng, Z., Lu, X., Guo, C., & Yang, B. (2024). *Ocean significant wave height estimation with spatio-temporally aware large language models*. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM ’24) (pp. 3892–3896). Association for Computing Machinery. <https://doi-org.libproxy.scu.edu/10.1145/3627673.3679973>

**Link to Presentation:**<https://drive.google.com/file/d/1fQymAOqeBRZ23k_YBdpddhvFtRaMui6E/view?usp=sharing>