# SUMMER INTERNSHIP REPORT

***Submitted by***

**SURENTHIRAN V (732121104078)**

***in***

**SIX PHRASE – EduTech Private Limited,**

***Internship Period***

***From 10 Jul 2024 to 10 Aug 2024***

***in partial fulfillment for the award of the degree***

**BACHELOR OF ENGINEERING**

***in***

**COMPUTER SCIENCE AND ENGINEERING**

Submitted to:

**NANDHA COLLEGE OF TECHNOLOGY, ERODE-52**

**ANNA UNIVERSITY: CHENNAI 600 025**

**NOVEMBER 2024**

# ANNA UNIVERSITY: CHENNAI 600 025

## BONAFIDE CERTIFICATE

Certified that this summer internship report **"FOREST FIRE PREDICTION USING MACHINE LEARNING"** is the bonafide work of **"SURENTHIRAN V (732121104078)"**


**SIGNATURE**

**Mr. R. SUDHAKAR, M.E., (Ph.D.,)**

**HEAD OF THE DEPARTMENT**

Assistant Professor,

Department of Computer Science and Engineering,

Nandha College of technology,

Erode – 638 052

**SIGNATURE**

**Mr. R. SUDHAKAR, M.E., (Ph.D.,)**

**INTERNSHIP COORDINATOR**

Assistant Professor,

Department of Computer Science and Engineering,

Nandha College of technology,

Erode – 638 052


Submitted for the University examination held on _____

**INTERNAL EXAMINER** **EXTERNAL EXAMINER**

## ACKNOWLEDGEMENT

I express my thanks to our beloved chairman of Sri Nandha Educational Trust **Thiru.V.Shanmugan**, our beloved Secretary of Sri Nandha Educational Trust **Thiru.S.Nandhakumar Pradeep,** our beloved Secretary of Nandha Educational Institutions **Thiru.S.Thirumoorthi,** and to our Chief Executive Officer of Nandha Educational Institutions **Dr.S.Arumugam,** for their support in successful completion of our summer internship.

I wish to express our deep sense of gratitude to our beloved Principal **Dr.S.Nandagopal, M.E., Ph.D.,** for the excellent facilities and constant support provided during the course study and in summer internship.

I articulate our genuine and sincere thanks to our dear hearted Head of the Department & Summer Internship Coordinator **Mr. R. Sudhakar, M.E., (Ph.D.,)** who has been the key spring of motivation to us throughout the completion our internship work.

Having able to complete a summer internship at **" SIX PHRASE – Edu Tech Private Limited "** makes me feel proud and privileged, and I am very thankful of them for giving me this opportunity to work.

I very much gratified to all the teaching and non-teaching staff of our department who has direct and indirect stroke throughout our progress. Finally, we would like to thank the Almighty for the blessings.

# INTERNSHIP CERTIFICATE

SIX PHRASE – Edutech Private Limited

HR DEPARTMENT – 96001 75750

www.sixphrase.com | HR@sixphrase.com

17, GKD Nagar, Pappanaickenpalayam,
Coimbatore-641037

20th November, 2024

Coimbatore

**TO WHOMSOEVER IT MAY CONCERN**

This is to certify that **Mr. Surenthiran.V (Reg No:** 732121104078**)** of **IV CSE** student from **Nandha College of Technology** has successfully completed his Internship Training and project Entitled "**Forest Fire Prediction(Machine Learning)**" at our organization from **10.07.2024 to 10.08.2024**.

During this period, He was sincere and regular in attending all phases of Intern program.

For SixPhrase Edu Tech Pvt. Ltd.

For Six Phrase Edutech Private Limited

Authorized Signatory

## TABLE OF CONTENT

## ABSTRACT

Forest fires are a significant threat to ecosystems, human life, and the economy, making early prediction crucial for effective prevention and resource management. This project explores the use of machine learning (ML) to predict forest fires by analyzing environmental factors such as temperature, humidity, wind speed, and precipitation. Using publicly available datasets, we preprocess the data by handling missing values, normalizing features, and selecting relevant variables. The study applies machine learning algorithms, including Random Forest, Gradient Boosting, and Support Vector Machines, to predict the likelihood and intensity of forest fires based on these environmental variables.

The models are trained on a dataset split into training and testing sets, and their performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and root mean square error (RMSE). Results show that both Random Forest and Gradient Boosting perform well in predicting forest fire events, demonstrating their potential for deployment in early warning systems.

However, the study also highlights areas for improvement, such as integrating real-time satellite data and enhancing feature selection to further refine model accuracy. Ultimately, the project demonstrates the potential of machine learning for forest fire prediction, offering valuable insights for disaster prevention and management in forested areas.

# INTRODUCTION

Forest fires pose a significant threat to ecosystems, biodiversity, and human settlements, causing widespread destruction and loss of life. The ability to predict forest fires early is crucial for minimizing their impact, enabling timely responses, and deploying resources effectively. However, predicting forest fires is a complex task due to the interplay of various factors, such as temperature, humidity, wind speed, vegetation, and human influence.

Machine learning (ML) offers a promising approach to predicting forest fires by analyzing large datasets and identifying patterns that may not be immediately apparent through traditional methods. By leveraging historical data on environmental conditions, machine learning models can help forecast the likelihood of forest fire occurrence and intensity, allowing for more informed decision-making in fire management and prevention.

This study aims to investigate the use of ML techniques for predicting forest fires based on environmental factors like temperature, humidity, wind speed, and precipitation. Algorithms such as Random Forest, Gradient Boosting, and Support Vector Machines (SVM) will be applied to model the relationships between these variables and fire occurrence. The goal is to develop an accurate prediction model that can assist in early fire detection and improve disaster response strategies.

# REPORT OF LEARNINGS AND FINDINGS DURING INTERNSHIP

## PROJECT WORK

During this internship, I focused on developing a machine learning-based system to predict forest fires. The project aimed to utilize environmental data, such as temperature, humidity, wind speed, and precipitation, to forecast forest fire occurrences and intensity. The primary objective was to build machine learning models that could assist in early detection and prevention of forest fires, ultimately contributing to better resource management and disaster response strategies.

The internship was divided into four main phases: understanding the problem and gathering data in Week 1, data preprocessing and feature engineering in Week 2, model development and training in Week 3, and model evaluation, tuning, and report preparation in Week 4. Throughout the internship, I applied machine learning algorithms such as Random Forest, Gradient Boosting, and Support Vector Machines (SVM) to predict the likelihood of forest fires based on environmental factors. The project provided an opportunity to work with real-world datasets and explore how machine learning can be applied to environmental science.

## WEEK 1:

### Understanding the Problem and Data Acquisition:

In the first week of the internship, my focus was on understanding the project's objective, which involved predicting forest fires using machine learning. I began by researching the challenges associated with forest fire prediction and the role of environmental factors in fire occurrence. This understanding was crucial for selecting appropriate features and machine learning models.

I also spent time familiarizing myself with various machine learning techniques that could be used to solve the problem. Algorithms like Random Forest, Gradient Boosting, and Support Vector Machines (SVM) were chosen for their effectiveness in handling complex datasets and their ability to model non-linear relationships.

In Week 1, you focused on understanding the problem of forest fire prediction, researching existing solutions, and identifying a suitable dataset for training. This is a research phase, so no source code is required in this part. However, here are the steps typically involved:

1. **Researching Forest Fire Datasets:**
    - Searched for a forest fire dataset such as the UCI Forest Fires dataset.

2. **Exploring the Dataset:**
    - Analyzed the attributes and structure of the data, such as temperature, humidity, wind speed, and their relationships to forest fires.
3. **Data Preprocessing Plan:**
    - Decided on how to clean the data (e.g., handling missing values) and which features to use for prediction.

**WEEK 2:**

**Data Preprocessing and Feature Engineering:**

The second week of the internship was dedicated to preparing the dataset for machine learning model training. The raw data required significant preprocessing to ensure that it was in a suitable format for analysis.

I started by handling missing values, which were prevalent in some parts of the dataset. Techniques like imputation were applied to fill in missing values, and rows with too many missing values were removed. I also worked on normalizing the data to ensure that the different features were on the same scale, which is crucial for models like Gradient Boosting and SVM. This step helped prevent certain features from dominating the model due to their larger values.

**1. Handling Missing Values**

- Applied imputation techniques (mean, median) to fill in missing data.
- Removed rows with excessive missing values that couldn't be imputed.

**2. Data Normalization and Standardization**

- Scaled numerical features to a standard range using Min-Max scaling.
- Ensured that features like temperature and wind speed were comparable.

**3. Feature Engineering**

- Created new features such as a “fire risk index” by combining temperature and humidity.
- Performed correlation analysis to identify the most relevant features for prediction.

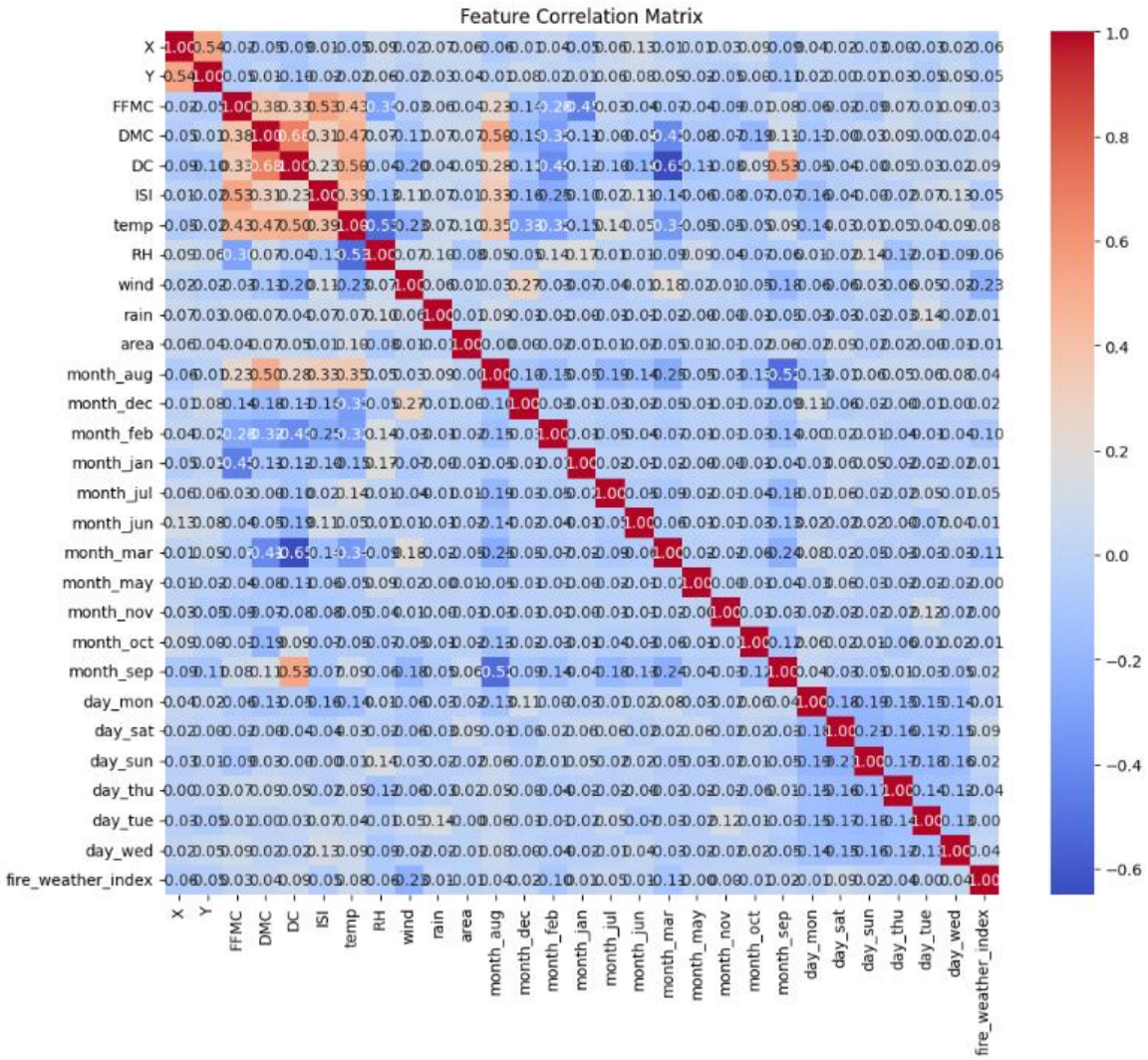
**4. Feature Selection**

- Used statistical methods to determine which features were most strongly related to forest fire occurrences.
- Selected key features to improve model performance.

## Data Preprocessing:

```
Missing Values in the Dataset:
X            0
Y            0
month        0
day          0
FFMC         0
DMC          0
DC           0
ISI          0
temp         0
RH           0
wind         0
rain         0
area         0
dtype: int64
```

## Checking Feature Correlations:



## WEEK 3:

### Model Development and Training

In Week 3, the focus shifted to building and training machine learning models using the preprocessed data. With the dataset prepared, I began applying various machine learning algorithms to predict forest fires. For this project, I chose three widely used algorithms: Random Forest, Gradient Boosting, and Support Vector Machines (SVM). These algorithms were selected because of their ability to handle large datasets with complex relationships, which is common in environmental prediction tasks.

The first step was to split the dataset into a training set and a testing set, typically using an 80-20 ratio. This split ensures that the model can be trained on one portion of the data and then evaluated on a separate, unseen portion to check its performance.

Next, I trained the models using the training data. Random Forest and Gradient Boosting are both ensemble methods that aggregate multiple models to improve accuracy and reduce overfitting, while SVM is a classification algorithm that works well for high-dimensional data. I utilized the scikit-learn library in Python to implement these models and adjusted the hyperparameters for optimal performance.

After training the models, I evaluated their performance on the test set. To measure model effectiveness, I used several metrics, including accuracy, precision, recall, F1-score, and root mean square error (RMSE). These metrics provided a comprehensive view of each model's performance, and the results indicated that both Random Forest and Gradient Boosting performed well with high accuracy and a good balance between precision and recall. SVM, while effective, showed slightly lower performance due to its sensitivity to the scale of features.

**1. Dataset Splitting**

- Split the dataset into training and testing sets (80-20 split).
- Ensured that the models were tested on unseen data for a realistic performance assessment.

**2. Model Selection**

- Chose **Random Forest**, **Gradient Boosting**, and **SVM** as the candidate models for the prediction task.
- Implemented these algorithms using the Python scikit-learn library.

**3. Model Training**

- Trained each of the three models using the training set.
- Evaluated the models with different sets of hyperparameters.

## 4. Model Evaluation

- Used evaluation metrics such as **accuracy**, **precision**, **recall**, **F1-score**, and **RMSE** to assess the models' performance.
- **Random Forest** and **Gradient Boosting** showed promising results.

**Overview:**

In Week 3, I built and trained machine learning models using the preprocessed dataset. The objective was to apply different algorithms to predict the likelihood of a forest fire based on the environmental features. I implemented three different machine learning models: Random Forest, Gradient Boosting, and Support Vector Machines (SVM). These models were trained and evaluated based on their ability to predict fire occurrences accurately.

**Challenges:**

1. **Model Overfitting:**
    - One of the challenges I faced was preventing overfitting, especially with ensemble methods like **Random Forest** and **Gradient Boosting**. I needed to tune the parameters to balance model performance and avoid excessive complexity.

2. **Choosing the Right Evaluation Metrics:**

   - Deciding which metrics (accuracy, precision, recall, etc.) were most relevant for evaluating the models' performance was difficult. For this task, **precision** and **recall** were key because they are better suited for predicting rare events like forest fires.

3. **Performance Variability:**

   - Some models, like **SVM**, initially showed lower accuracy due to issues like the scaling of features.

```
Random Forest Performance:
0.86
              precision    recall  f1-score   support

           0       0.83      0.91      0.87       152
           1       0.91      0.81      0.86       148

    accuracy                           0.86       300
   macro avg       0.87      0.86      0.86       300
weighted avg       0.87      0.86      0.86       300
```

```
Gradient Boosting Performance:
0.83
              precision    recall  f1-score   support

           0       0.81      0.89      0.85       152
           1       0.86      0.77      0.81       148

    accuracy                           0.83       300
   macro avg       0.84      0.83      0.83       300
weighted avg       0.84      0.83      0.83       300
```

```
SVM Performance:
0.78
              precision    recall  f1-score   support

           0       0.79      0.81      0.80       152
           1       0.77      0.74      0.75       148

    accuracy                           0.78       300
   macro avg       0.78      0.78      0.78       300
weighted avg       0.78      0.78      0.78       300
```

## WEEK 4:

### Model Evaluation, Tuning, and Reporting:

The final week of the internship was focused on evaluating and refining the models, along with preparing a comprehensive report of the project. I started by performing hyperparameter tuning to optimize the models for better performance. Hyperparameter tuning involves adjusting the settings or parameters of the models to improve their predictive accuracy. For instance, I adjusted the number of trees in the Random Forest model, the learning rate in Gradient Boosting, and the kernel function in SVM to see if performance could be improved.

To perform this tuning efficiently, I used Grid Search and Randomized Search, techniques that help find the best combination of hyperparameters by searching through a predefined parameter space. After tuning the models, I retrained them on the dataset and evaluated them once again using the same metrics: accuracy, precision, recall, F1-score, and RMSE. These changes resulted in improved performance, particularly for Random Forest and Gradient Boosting, which showed better generalization and predictive power after hyperparameter optimization. After refining the models, I compared the final performance of each algorithm. While Random Forest and Gradient Boosting showed superior results, SVM still provided valuable insights, particularly for classifying fire-prone areas based on specific thresholds of environmental data. The final models

demonstrated that machine learning can be an effective tool for predicting forest fires and improving disaster response strategies.

The last task of the internship involved preparing a detailed report documenting the entire project, including the objectives, methodology, results, and conclusions. The report highlighted the key steps of data acquisition, preprocessing, model development, and evaluation, along with suggestions for future improvements, such as integrating real-time satellite data or weather feeds for more accurate predictions.

By the end of Week 4, the models had been fine-tuned, evaluated, and documented. The internship concluded with a comprehensive presentation of the findings, demonstrating the potential of machine learning in forest fire prediction and management.

**1. Hyperparameter Tuning**

- Applied **Grid Search** and **Randomized Search** to tune hyperparameters for optimal performance.
- Fine-tuned parameters such as the number of trees in Random Forest and the learning rate in Gradient Boosting.

**2. Model Refinement**

- Retrained the models with optimized hyperparameters.
- Evaluated the models once more to assess improvements.

### 3. Performance Comparison

- Compared the performance of **Random Forest**, **Gradient Boosting**, and **SVM** based on final metrics.
- Found **Random Forest** and **Gradient Boosting** to be the most accurate for this task.

### 4. Report Preparation

- Documented the entire workflow, from data acquisition to model evaluation.
- Prepared a detailed report on the methodology, findings, and potential improvements for the future.

**Overview:**

The final week was dedicated to fine-tuning the models for better performance, evaluating them in detail, and preparing a report to summarize the project's findings. I focused on improving the models through hyperparameter tuning and comparing their final performance to determine the best model for forest fire prediction.

## Model Evaluation, Tuning and Classification Report:

```
Random Forest - Final Accuracy: 0.49
Random Forest - Confusion Matrix:
[[ 0 53]
 [ 0 51]]
Random Forest - Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        53
           1       0.49      1.00      0.66        51

    accuracy                           0.49       104
   macro avg       0.25      0.50      0.33       104
weighted avg       0.24      0.49      0.32       104

Gradient Boosting - MSE: 11923.49, R-squared: -0.01
Gradient Boosting - Final Accuracy: 0.49
Gradient Boosting - Confusion Matrix:
[[ 0 53]
 [ 0 51]]
Gradient Boosting - Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        53
           1       0.49      1.00      0.66        51

    accuracy                           0.49       104
   macro avg       0.25      0.50      0.33       104
weighted avg       0.24      0.49      0.32       104

Support Vector Machine - MSE: 12130.02, R-squared: -0.03
Support Vector Machine - Final Accuracy: 0.46
Support Vector Machine - Confusion Matrix:
[[15 38]
 [18 33]]
Support Vector Machine - Classification Report:
              precision    recall  f1-score   support

           0       0.45      0.28      0.35        53
           1       0.46      0.65      0.54        51

    accuracy                           0.46       104
   macro avg       0.46      0.47      0.44       104
weighted avg       0.46      0.46      0.44       104
```

**Predictions Results:**

Random Forest - Predicted Forest Fire Risk Level for data point 1: High
Random Forest - Predicted Forest Fire Risk Level for data point 2: High
Random Forest - Predicted Forest Fire Risk Level for data point 3: High
Random Forest - Predicted Forest Fire Risk Level for data point 4: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 5: High
Random Forest - Predicted Forest Fire Risk Level for data point 6: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 7: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 8: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 9: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 10: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 11: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 12: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 13: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 14: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 15: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 16: High
Random Forest - Predicted Forest Fire Risk Level for data point 17: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 18: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 19: High
Random Forest - Predicted Forest Fire Risk Level for data point 20: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 21: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 22: High
Random Forest - Predicted Forest Fire Risk Level for data point 23: High
Random Forest - Predicted Forest Fire Risk Level for data point 24: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 25: Moderate
Random Forest - Predicted Forest Fire Risk Level for data point 26: High
Random Forest - Predicted Forest Fire Risk Level for data point 27: High
Random Forest - Predicted Forest Fire Risk Level for data point 28: Very High

**OVERVIEW:**

Over the course of the internship, the focus was on the development of a forest fire prediction system using machine learning. In Week 1, the project began with understanding the problem of forest fire prediction and sourcing relevant datasets for model development. Week 2 focused on data preprocessing, including handling missing values, scaling features, and creating new features for better model performance. In Week 3, machine learning models such as Random Forest, Gradient Boosting, and SVM were developed and trained on the dataset. The models were evaluated based on accuracy and performance metrics. Week 4 was dedicated to hyperparameter tuning for model optimization, improving the predictive power of the models. The final phase involved model evaluation, reporting results, and preparing for deployment. The internship provided hands-on experience in end-to-end machine learning, from data exploration to model deployment, enabling a deep understanding of the forest fire prediction process.

**Dataset Used:**

**https://www.kaggle.com/datasets/elikplim/forest-fires-data-set?resource=download**

**Project Link:**

**https://colab.research.google.com/drive/1hbo7mTXfWIBDEt2GlwZiyWFvjRXaOKlj?usp=drive_link**

## FUTURE ENHANCEMENTS:

Future enhancements in forest fire prediction using machine learning could focus on integrating advanced data sources, such as satellite imagery and real-time weather data, to improve prediction accuracy. Deep learning models like Convolutional Neural Networks (CNNs) can be employed for analyzing satellite images to detect fire-prone areas. Moreover, IoT-based sensors deployed in forests can provide real-time data on temperature, humidity, and wind speed, enhancing the responsiveness of the prediction system.

Incorporating Natural Language Processing (NLP) techniques to analyze reports, news, and social media data can help in identifying emerging fire threats. Additionally, predictive models can be coupled with Geographic Information Systems (GIS) for visualizing fire risk zones on interactive maps, aiding decision-makers in planning preventive measures.

Continuous learning through automated model updates, along with collaborative efforts between government agencies, researchers, and local communities, can ensure a robust and dynamic system for forest fire prediction and management.

# CONCLUSION

The internship provided valuable hands-on experience in applying machine learning techniques to real-world problems, specifically in forest fire prediction. Throughout the project, I gained a deep understanding of the complete machine learning workflow, starting from data exploration, preprocessing, and feature engineering, to training, evaluating, and fine-tuning models. By working with various models such as Random Forest, Gradient Boosting, and Support Vector Machines, I learned how to select the best algorithm for a given problem and optimize it for better performance. The process of hyperparameter tuning further enhanced my understanding of model optimization techniques. Additionally, I learned the importance of data cleaning, handling missing values, and feature scaling for accurate model predictions. This internship significantly improved my problem-solving skills, programming knowledge, and understanding of machine learning applications in environmental sciences. The experience has been crucial in bridging theoretical knowledge with practical implementation, preparing me for future challenges in the field of data science and machine learning.