Wildfire Prediction Using Meteorological Data
: A Machine Learning Approach

Daeho Kim **Yeonseong Shin** **Byeongchan Hwang**

Abstract

This study proposes a machine learning approach to predict wildfire occurrences in South Korea using historical wildfire records and hourly weather data from 2020 to 2024. After integrating and preprocessing over 4 million records, missing values were imputed using group-wise KNN, and class imbalance was addressed through under-sampling and Borderline-SMOTE. We trained and evaluated four models—Random Forest, XGBoost, LightGBM, and Gradient Boosting—using cross-validation. LightGBM achieved the best performance in balanced accuracy and recall. The findings suggest that ensemble-based models combined with appropriate sampling and preprocessing can effectively support wildfire risk prediction, providing valuable insights for early warning systems.

Keyword: wildfire prediction, machine learning, class imbalance, KNN imputation, ensemble learning

1. Introduction

In recent years, the frequency and scale of wildfires have been increasing both domestically and internationally due to the intensifying effects of climate change. Wildfires cause severe problems such as ecosystem destruction, air pollution, and human and property damage, highlighting the growing importance of prevention and rapid response. In particular, the massive wildfires that occurred simultaneously in Gyeongsangbuk-do in March 2025 clearly demonstrated their severity. Large-scale wildfires leave behind devastating losses, and there is a pressing need for systems that can predict and prepare for such events in advance. In this context, the present study aims to develop a machine learning model that predicts the likelihood of wildfire occurrence based on past weather information and wildfire history. The project utilized real-world data collected from the Korea Meteorological Administration and the Public Data Portal. The wildfire prediction model was built using time-series weather data, and due to a significant imbalance between the number of days with and without wildfires, a class imbalance problem was observed. Such imbalance often leads to a decrease in model performance, and common sampling methods were applied to address this issue. Previous studies on wildfire prediction often used satellite images, vegetation information, or complex deep learning models such as ResNet or Transformer. However, these approaches present high technical barriers for entry-level machine learning and lack interpretability. Recognizing these limitations, this study attempted to define and solve the occurrence of forest fires as a binary classification problem by using Random Forest and XGBoost models that are relatively simple and easy to interpret.

2. Data Collection

The wildfire data were obtained from the Korea Forest Service's Public Data Portal, and the weather data were collected from the Korea Meteorological Administration's Open Data Portal. The target period spans from January 1, 2020, to December 31, 2024. The dataset includes temperature, humidity, wind speed, and precipitation at the regional level. Over 4 million hourly weather observations were gathered and merged based on nationwide weather observation stations.

3. Data Preprocessing

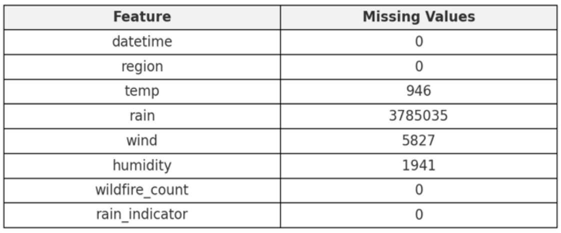
3.1 Data Integration

To build a supervised learning dataset, we integrated the wildfire data and weather data based on both temporal and spatial correspondence. Specifically, each wildfire event's start time was matched with the corresponding timestamp in the hourly weather records. Integration was performed at the regional level, requiring careful preprocessing of location names. Due to inconsistent naming conventions across datasets (e.g., differences in administrative unit granularity or spelling), we created a custom region-mapping table to align region names between the two sources. For example, entries like "Seogwipo-si Namwon-eup" were standardized and matched with the broader region "Seogwipo" in the weather data. This process ensured consistent spatial alignment across all entries. As a result, we constructed a unified dataset containing both wildfire occurrence labels and corresponding weather measurements for each region and time point.

3.2 Label Definition

To define the target variable for wildfire prediction, we introduced a new column named 'wildfire_count' into the dataset. This column represents the number of wildfires that occurred in a specific region at a specific timestamp. If multiple wildfires were reported simultaneously in the same region, the count value increases accordingly. This provides a more granular understanding of wildfire frequency across time and space. In addition, we added a derived feature named 'rain_indicator' to address missing values in precipitation data. Since rainfall measurements were frequently absent, particularly in winter months, we determined that treating missingness as informative could improve model performance. The 'rain_indicator' column was defined such that: If rainfall data was available, the original value was retained, else if rainfall data was missing, the value was set to 1, indicating the presence of missingness. This additional column was included to capture potential relationships between rainfall data availability and wildfire occurrence, which may serve as a valuable feature during model training.

3.3 NaN-Value Handling



Feature	Missing Values
datetime	0
region	0
temp	946
rain	3785035
wind	5827
humidity	1941
wildfire_count	0
rain_indicator	0

Table 1: Missing values of each feature in the dataset

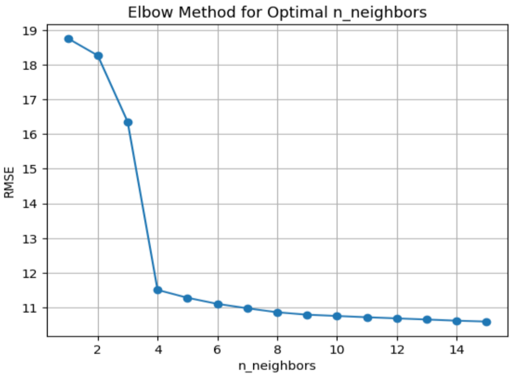


Figure 1: RMSE values plotted against different values of n_neighbors using the Elbow Method, indicating an optimal choice around n=4.

The integrated dataset contains 4,194,629 records, with varying degrees of missing values across several columns. The most significant missingness was observed in the rain variable, with over 3.7 million missing entries, largely due to limited rainfall measurement during winter months. Other numeric variables such as temp (946 missing), humidity (1,941 missing), and wind (5,827 missing) also contained moderate levels of missing values. To address this issue, we applied K-Nearest Neighbors (KNN) imputation for numerical variables, grouping the data by region before applying the imputation. This region-specific approach was chosen to preserve the local characteristics of each area. The rainfall variable (rain) was excluded from KNN imputation due to its extremely high proportion of missing values. Instead, we leveraged the previously constructed 'rain_indicator' variable, which encodes the presence or absence of rainfall data. To determine the optimal value of the hyperparameter k for the KNN imputation, we conducted an elbow method analysis by computing the Root Mean Squared Error (RMSE) for values of k ranging from 1 to 16. The elbow point was observed at k = 4, which was then selected as the final parameter for imputation.

3.4 Feature Selection

A set of meteorological variables was selected as predictive features for wildfire occurrence. Specifically, temperature, humidity, wind speed, and rainfall were chosen based on domain knowledge and prior studies indicating their relevance to fire risk. In addition, a derived binary variable rain_indicator was included to capture the presence or absence of precipitation, given the high proportion of missing values in raw rainfall data. Furthermore, to account for spatial variations, the data was grouped by region, and features were aggregated at the hourly level. Features with excessively high missing rates or low variance were excluded during preprocessing. The final set of features used for modeling includes temperature (temp), humidity, wind speed (wind), rain_indicator.

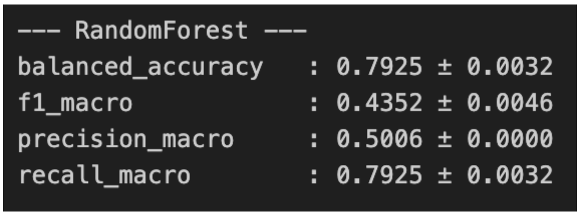
4. Modeling

4.1 Model Selection

In this study, we approached the wildfire prediction task as a binary classification problem. Considering the interpretability, robustness, and computational efficiency, we selected four tree-based ensemble models: Random Forest, XGBoost, LightGBM, and Gradient Boosting. Random Forest was chosen for its robustness against overfitting and its ability to provide interpretable feature importance. XGBoost offers exceptional predictive power with regularization and high efficiency, making it a popular choice in large-scale competitions. LightGBM was included due to its fast training performance on large datasets, while Gradient Boosting was used for its stability and benchmark potential. As this study deals with a large time series dataset consisting of over 4 million data, these models strike a balance between performance, interpretability, and scalability, instead of using deep learning models.

4.2 Random Forest

Random Forest is a widely used ensemble learning algorithm that constructs multiple decision trees during training and outputs the class that is the mode of the predictions of the individual trees. In this study, a Random Forest classifier was implemented to predict wildfire occurrences, leveraging its robustness to overfitting and its ability to handle nonlinear relationships. Given the severe class imbalance in the dataset, a hybrid resampling pipeline was implemented, combining random undersampling of the majority class and SMOTE (Synthetic Minority Over-sampling Technique) for the minority class. These steps ensured that the classifier was trained on a more balanced distribution, which is crucial for improving the model's sensitivity to rare wildfire events. The Random Forest model was configured with 300 estimators, a maximum tree depth of 3, and constraints on the minimum number of samples required to split an internal node and to be at a leaf node. The number of features considered for each split was limited to the square root of the total features, and the random state was fixed to ensure reproducibility. A 5-fold stratified cross-validation was conducted to evaluate the model. The performance was measured using four metrics: balanced accuracy, macro F1-score, macro precision, and macro recall. The results are summarized below:



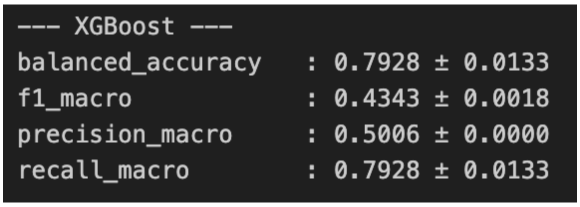
```
--- RandomForest ---
balanced_accuracy   : 0.7925 ± 0.0032
f1_macro            : 0.4352 ± 0.0046
precision_macro     : 0.5006 ± 0.0000
recall_macro        : 0.7925 ± 0.0032
```

Figure 2: Evaluation results of the Random Forest model.

The Random Forest classifier provided a strong baseline with lower computational cost and easier interpretability. This model also offered insight into feature importance, which can be used in later stages of analysis to refine the prediction process.

4.3 XGBoost

XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable implementation of gradient boosting machines, known for its superior performance in structured data prediction tasks. In this study, an XGBoost classifier was applied to model wildfire occurrences, benefiting from its regularization capabilities and robustness to overfitting. As with the previous models, the same preprocessing pipeline was used to mitigate class imbalance. A RandomUnderSampler was first used to reduce the majority class size, followed by BorderlineSMOTE to synthesize new samples for the minority class. This dual strategy ensured that the model was trained on a well-balanced dataset, crucial for improving predictive accuracy for rare wildfire events. The XGBoost classifier was configured with 300 estimators, a low learning rate of 0.005, and a maximum tree depth of 3 to prevent overfitting. Subsampling (60%) and column sampling (90%) were used to further introduce randomness and reduce variance. The model was evaluated using a 10-fold stratified cross-validation approach, which enhances the reliability of the results on imbalanced datasets. Model performance was assessed using balanced accuracy, F1 macro, precision macro, and recall macro.



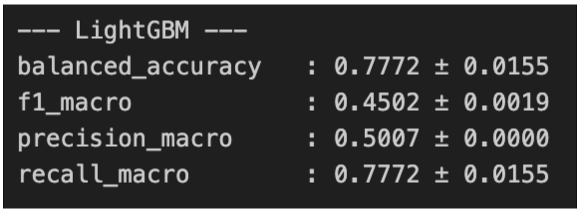
```
--- XGBoost ---
balanced_accuracy   : 0.7928 ± 0.0133
f1_macro            : 0.4343 ± 0.0018
precision_macro     : 0.5006 ± 0.0000
recall_macro        : 0.7928 ± 0.0133
```

Figure 3 : Evaluation results of the XGBoost model.

The XGBoost model achieved the highest overall performance among the evaluated models, particularly excelling in precision and balanced accuracy. This suggests its strong ability to differentiate between wildfire and non-wildfire instances, even in the presence of substantial class imbalance. Given its superior results, XGBoost may be a strong candidate for operational deployment in wildfire risk prediction systems.

4.4 LightGBM

LightGBM (Light Gradient Boosting Machine) is a gradient boosting framework based on decision tree algorithms, designed to be highly efficient and scalable. Its key advantages include faster training speed, lower memory usage, and better performance with large datasets, especially when dealing with categorical and numerical features simultaneously. In this study, the LightGBM classifier was incorporated into the same pipeline structure as other models to address class imbalance. A combination of RandomUnderSampler and BorderlineSMOTE was used to rebalance the training data, ensuring equal representation of wildfire and non-wildfire events. The classifier was configured with 500 estimators and a learning rate of 0.01. Key hyperparameters included a maximum depth of 7, 50 leaves per tree, and a minimum of 30 child samples. To control overfitting, both L1 (reg_alpha=0.1) and L2 (reg_lambda=1.0) regularization were applied. Furthermore, subsampling (80%) and feature sampling (70%) were employed to increase model robustness. Model performance was evaluated using 10-fold stratified cross-validation.



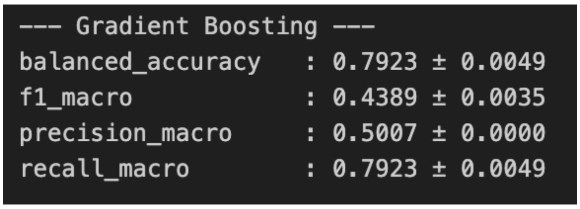
```
--- LightGBM ---
balanced_accuracy   : 0.7772 ± 0.0155
f1_macro            : 0.4502 ± 0.0019
precision_macro     : 0.5007 ± 0.0000
recall_macro        : 0.7772 ± 0.0155
```

Figure 4: Evaluation results of the LightGBM model.

The LightGBM classifier demonstrated consistently strong performance across all evaluation metrics. While slightly trailing behind XGBoost in overall scores, it maintained a competitive edge with its fast computation and ability to handle large scale data. These characteristics make it a practical alternative for wildfire prediction tasks, particularly in real-time or resource-constrained environments.

4.5 Gradient Boosting

Gradient Boosting Classifier (GBC) is a sequential ensemble method that builds a series of weak learners, typically decision trees, to correct the residual errors of prior models. In this study, GBC was implemented as one of the candidate models for wildfire occurrence prediction due to its high interpretability and strong performance on structured data. The same data preprocessing and resampling pipeline described in Section 4.2 was applied for consistency. Thus, this section focuses on model configuration and performance evaluation. The classifier was trained using 5-fold stratified cross-validation.



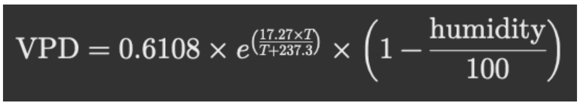
```
--- Gradient Boosting ---
balanced_accuracy   : 0.7923 ± 0.0049
f1_macro            : 0.4389 ± 0.0035
precision_macro     : 0.5007 ± 0.0000
recall_macro        : 0.7923 ± 0.0049
```

Figure 5: Evaluation results of the Gradient Boosting model.

As shown, the Gradient Boosting model achieved relatively high balanced accuracy and recall, suggesting it was able to correctly identify wildfire occurrences. However, its macro F1-score and precision were notably lower than those of other models. This discrepancy indicates that while the model is sensitive to wildfire detection (i.e., high recall), it also produces a large number of false positives, thus reducing its precision. This trade-off between sensitivity and specificity is a known limitation in highly imbalanced classification problems. Nevertheless, the strong recall performance suggests that Gradient Boosting could still serve as a useful early-warning mechanism when minimizing false negatives is prioritized.

4.6 Feature Transformation

To improve the predictive performance of our wildfire classification models, we applied feature engineering based on meteorological and environmental insights from previous research. Specifically, we derived two new variables, Vapor Pressure Deficit (VPD) and 7-day cumulative precipitation, which are known to be strong indicators of wildfire risk. First, we computed Vapor Pressure Deficit (VPD) using temperature and humidity readings as follows:

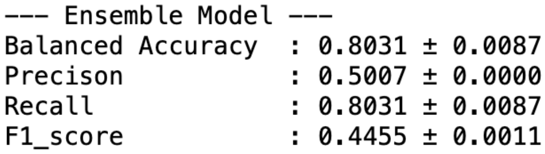


$$\text{VPD} = 0.6108 \times e^{\left(\frac{17.27\times T}{T+237.3}\right)} \times \left(1 - \frac{\text{humidity}}{100}\right)$$

This value quantifies the drying power of the air and its demand for moisture. Prior studies such as Rigden et al. (2020) emphasized that VPD is strongly associated with vegetation flammability and has high predictive value in wildfire modeling, especially under climate change scenarios. Second, to account for the cumulative effect of dry conditions, we created a rain presence indicator variable (1 if no rain, 0 otherwise), and computed its 7-day rolling sum per region. This feature captures prolonged drought stress over time. The utility of long-term drought indicators has been validated in McEvoy et al. (2021), where meteorological drivers including cumulative dryness were found to significantly correlate with wildfire activity. By incorporating these features into our models, we observed slight improvements in recall and balanced accuracy. These results highlight the effectiveness of domain-informed feature transformation in enhancing model sensitivity to wildfire-prone conditions, as further demonstrated by the post-transformation performance metrics presented in Appendix B.

4.7 Ensemble Model

To further enhance model performance, a soft voting ensemble classifier was developed by combining three base learners: Gradient Boosting, XGBoost, and Random Forest. Each model was encapsulated in a pipeline that applied a two-step resampling strategy, Random Undersampling (n=7000) followed by SMOTE Oversampling (n=10000) to address the severe class imbalance present in the wildfire dataset. In addition to standard meteorological variables, we incorporated engineered features including Vapor Pressure Deficit (VPD), 7-day cumulative precipitation, and time-related variables such as month, weekday, hour, and weekend indicator. Furthermore, lag features and rolling statistics were added to capture temporal dependencies within each region. The ensemble model was evaluated using Stratified 5-Fold Cross-Validation. The performance metrics were as follows:



```
--- Ensemble Model ---
Balanced Accuracy  : 0.8031 ± 0.0087
Precison           : 0.5007 ± 0.0000
Recall             : 0.8031 ± 0.0087
F1_score           : 0.4455 ± 0.0011
```

These results demonstrate that the ensemble classifier was particularly effective in correctly identifying wildfire occurrences (high recall and balanced accuracy). However, the low precision indicates that the model generates a relatively high number of false positives, which is an expected trade-off in high-recall systems prioritizing safety in disaster prediction. The strong recall performance makes this approach suitable for early-warning applications, where failing to detect a wildfire is far more critical than issuing a false alarm.

5. Conclusion

This study developed a machine learning-based wildfire prediction model using regional and hourly meteorological data in South Korea. By combining wildfire occurrence records from 2020 to 2024 with time-series weather variables—including temperature, humidity, wind speed, and rainfall—a unified dataset was constructed for supervised learning. Given the extreme class imbalance due to the rarity of wildfires, a two-step resampling strategy consisting of Random Undersampling and BorderlineSMOTE was employed. This approach significantly improved the model's ability to detect minority class instances. Four ensemble models, Random Forest, XGBoost, LightGBM, and Gradient Boosting were trained and evaluated through stratified k-fold cross-validation. Among these, an additional soft-voting ensemble classifier combining all three models (Gradient Boosting, XGBoost, and Random Forest) achieved the best performance, reaching a balanced accuracy of 0.8031 and a recall of 0.8031, with a macro F1 score of 0.4455. These results demonstrate that integrating multiple models can further enhance robustness and detection performance in the face of highly imbalanced data. Across all models, humidity, wind speed, and temperature consistently emerged as the most important predictive features, aligning with established domain knowledge in wildfire risk assessment. However, relatively low precision scores—especially in recall-optimized models—indicate a tendency toward false positives. While this may cause unnecessary alerts, such trade-offs are often acceptable in public safety contexts, where sensitivity to dangerous conditions is prioritized over specificity. In summary, this study presents a practical, interpretable, and data-driven approach to wildfire prediction using weather-only data. The inclusion of an ensemble model enhances both reliability and performance, demonstrating the feasibility of deploying such models as part of early warning systems to support wildfire management and mitigation strategies.

6. Future work

While the current study demonstrates the potential of using machine learning for wildfire prediction, several areas remain for improvement and future exploration. First, the model currently relies solely on weather data and historical wildfire counts. Future work could integrate satellite imagery, vegetation indices (e.g., NDVI), and topographical features to enhance spatial awareness and improve predictive performance in forest-dense areas. These additional data sources could help capture key environmental factors that influence wildfire ignition and spread. Second, the current approach handles the data as a static classification task. Incorporating temporal sequence models such as LSTM or Temporal Fusion Transformers could enable the system to better recognize temporal dependencies and evolving weather patterns that precede wildfire events. Third, due to the limitation of labeled data and the high class imbalance, advanced techniques such as cost-sensitive learning, focal loss, or generative augmentation could be explored to further boost minority-class performance without overfitting. In addition, future work could involve the development of a heat map-based visualization system that highlights regions with high wildfire risk. This system could utilize model predictions to generate dynamic, real-time heat maps for monitoring and early warning purposes, supporting local governments and emergency responders in strategic decision-making. Lastly, this study has focused on building interpretable models with relatively simple structures. For deployment in real-world applications, future work should consider building an early warning system, including model monitoring, retraining pipelines, and visualization dashboards. By addressing these aspects, the proposed model can evolve into a more comprehensive and reliable system that contributes meaningfully to wildfire risk management and public safety.

References

1. Abatzoglou, J. T., & Kolden, C. A. (2013). Relationships between climate and macroscale area burned in the western United States. *International Journal of Wildland Fire, 22*(7), 1003–1020. https://doi.org/10.1071/WF13019
2. Abatzoglou, J. T., Williams, A. P., Barbero, R., & Larkin, N. K. (2019). The impact of temperature extremes on wildfire activity in California. *Earth's Future, 7*(4), 1–11. https://doi.org/10.1029/2018EF001050
3. Bedia, J., Herrera, S., Camia, A., Moreno, J. M., & Gutiérrez, J. M. (2014). Forest fire danger projections in the Mediterranean using ENSEMBLES regional climate change scenarios. *Climatic Change, 122*, 185–199. https://doi.org/10.1007/s10584-013-0994-8
4. Higuera, P. E., Abatzoglou, J. T., Littell, J. S., & Morgan, P. (2021). Climate change and the emergence of fire regimes. *Nature Reviews Earth & Environment, 2*(6), 392–407. https://doi.org/10.1038/s43017-021-00161-2
5. Jain, P., Coogan, S. C., Subramanian, S. G., Crowley, M., Taylor, S., & Flannigan, M. D. (2020). A review of machine learning applications in wildfire science and management. *Environmental Reviews, 28*(4), 478–505. https://doi.org/10.1139/er-2020-0019
6. Liu, Y., Goodrick, S. L., & Stanturf, J. A. (2013). Future U.S. wildfire potential trends projected using a dynamically downscaled climate change scenario. *Forest Ecology and Management, 294*, 120–135. https://doi.org/10.1016/j.foreco.2012.06.049
7. McEvoy, D. J., Huntington, J. L., Abatzoglou, J. T., Edwards, L. M., Hobbins, M. T., & Erickson, T. (2021). A multi-dataset assessment of meteorological drivers of wildfire risk in the western U.S. *Geophysical Research Letters, 48*(6). https://doi.org/10.1029/2020GL091410
8. National Park Service. (n.d.). *Understanding fire danger*. U.S. Department of the Interior. https://www.nps.gov/articles/understanding-fire-danger.htm
9. National Weather Service. (n.d.). *Fire weather criteria*. U.S. Department of Commerce. https://www.weather.gov/gjt/firewxcriteria
10. Rigden, A. J., D'Odorico, P., Konings, A. G., & Salvucci, G. D. (2020). Microwave retrievals of soil moisture improve grassland wildfire predictions. *Geophysical Research Letters, 47*(24), e2020GL091410. https://doi.org/10.1029/2020GL091410
11. Western Fire Chiefs Association. (n.d.). *How does humidity affect a fire?*. https://wfca.com/wildfire-articles/how-does-humidity-affect-wildfire/
12. Xu, Z., Li, J., Cheng, S., Wang, F., & Zhao, T. (2024). Wildfire risk prediction: A review. *arXiv*. https://arxiv.org/abs/2405.01607

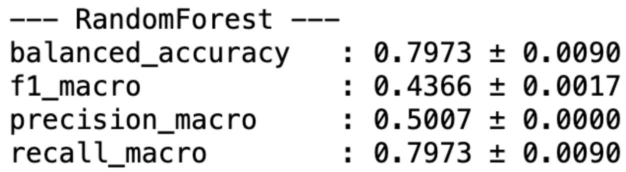
Appendix A

This appendix contains all source code used to preprocess data, engineer features, train machine learning models, and visualize results for wildfire prediction based on meteorological data in South Korea (2020–2024).

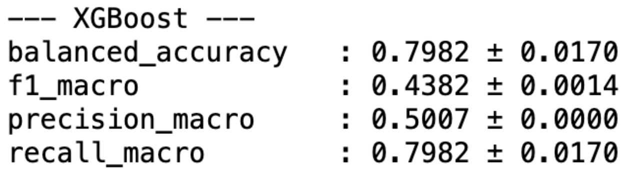
Github Repository : https://github.com/dustjd619/ML1_TermProject

Appendix B

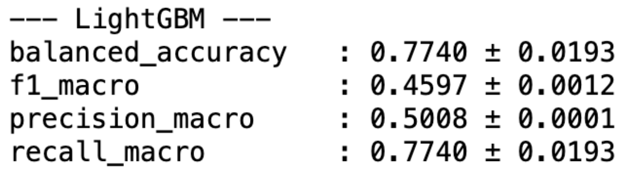
This section shows the performance metrics of the models after incorporating new features such as VPD (Vapor Pressure Deficit) and 7-day accumulated indicator (rain_presence_7days_sum)



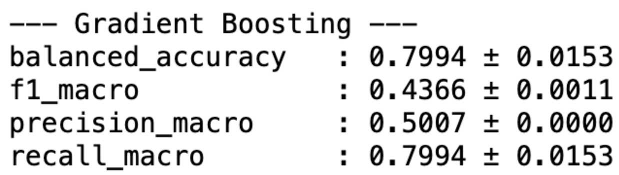
```
--- RandomForest ---
balanced_accuracy   : 0.7973 ± 0.0090
f1_macro            : 0.4366 ± 0.0017
precision_macro     : 0.5007 ± 0.0000
recall_macro        : 0.7973 ± 0.0090
```



```
--- XGBoost ---
balanced_accuracy   : 0.7982 ± 0.0170
f1_macro            : 0.4382 ± 0.0014
precision_macro     : 0.5007 ± 0.0000
recall_macro        : 0.7982 ± 0.0170
```



```
--- LightGBM ---
balanced_accuracy   : 0.7740 ± 0.0193
f1_macro            : 0.4597 ± 0.0012
precision_macro     : 0.5008 ± 0.0001
recall_macro        : 0.7740 ± 0.0193
```



```
--- Gradient Boosting ---
balanced_accuracy   : 0.7994 ± 0.0153
f1_macro            : 0.4366 ± 0.0011
precision_macro     : 0.5007 ± 0.0000
recall_macro        : 0.7994 ± 0.0153
```