EY Biodiversity Challenge

Project By:

Samarth Verma | Haeun Kim | Jayasree Lakshmi Narayanan

TABLE OF CONTENT

01 Overview

02 Project Process

03 Project Result

04 How We Improved Model

05 Value Case

06 What We Have Learned

Overview

- Our project aimed to predict frog presence in southeastern Australia using climate data from the TerraClimate dataset.
- Frogs serve as key indicators of ecosystem health, and accurate predictions can guide conservation, farming, and ESG initiatives.
- After advanced preprocessing and model testing, the ExtraTreesClassifier achieved the best performance with 83.10% test accuracy.
- The project shows how machine learning can turn ecological data into actionable insights for environmental planning.

Project Process

Dataset

- **Train Dataset (Training_Data.csv)**
  - 3792 frog presence (Occurrence Status □1)
  - 2520 frog absence (Occurrence Status □0)
  - Includes latitude and longitude for model training

- **TerraClimate Dataset (TerraClimate_output.tiff)**
  - Monthly climate data since 1958 at a 4 km spatial resolution
  - Contains 14 key climate variables impacting frog populations

- **Validation Data (Validation_Template.csv)**
  - 2000 new locations (latitude and longitude)
  - Used for validating model predictions

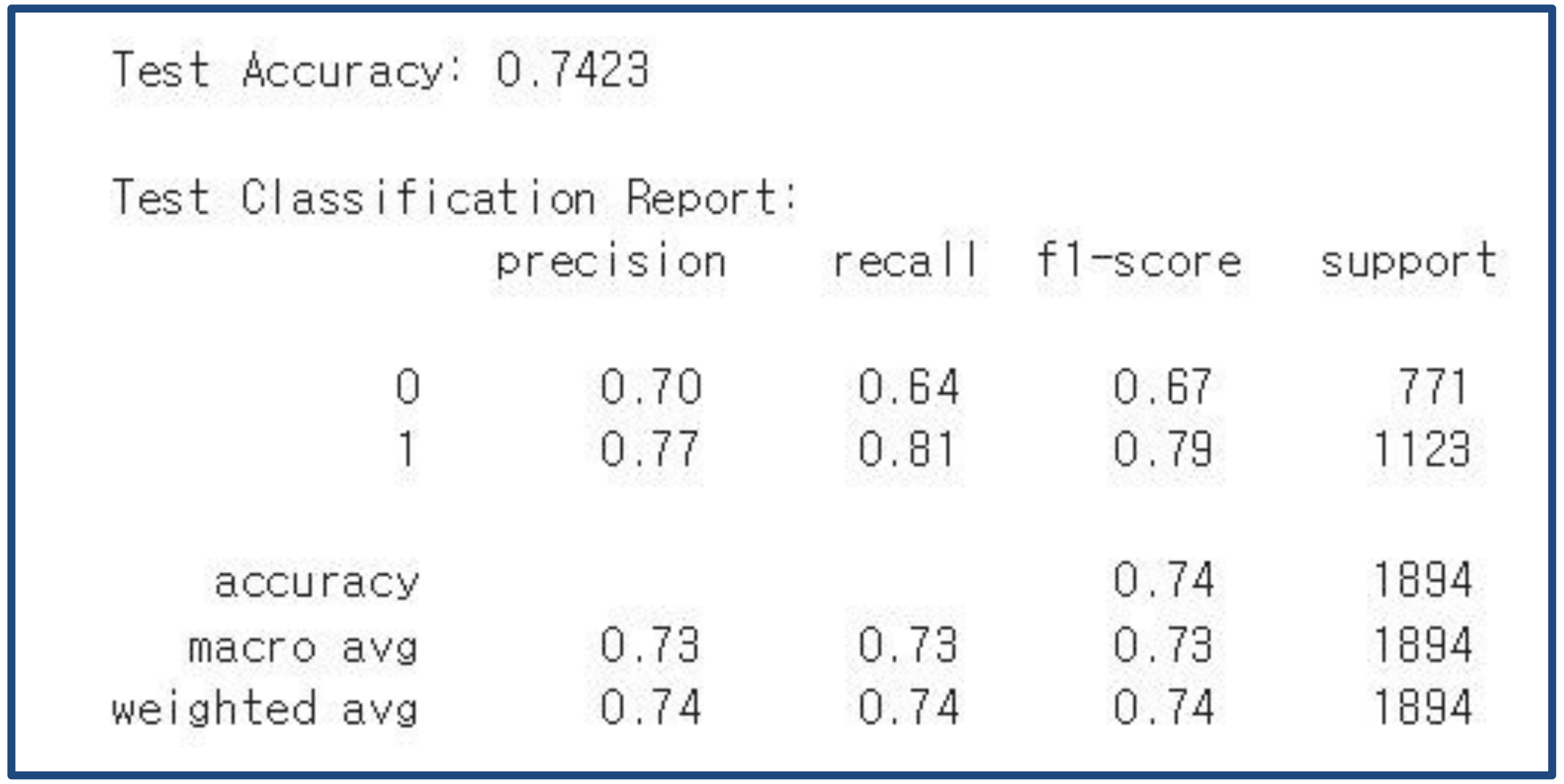
Data Pre-Processing Steps

- **Scope Narrowing**
  - Focused on Southeastern Australia (Nov 2017 - Nov 2019)
  - Predicting frog presence or absence at given coordinates
- **Data Integration**
  - Merged TerraClimate data with training data using latitude and longitude
- **Data Cleaning**
  - Removed all null values
  - Eliminated outliers using Z-score method
- **Feature Selection**
  - Dropped low-impact variable: Snow Water Equivalent (SWE)
  - Selected features based on correlation analysis and conceptual relevance
  - Conducted multiple trials with different feature combinations
- **Class Balancing**
  - Addressed imbalance with RandomOverSampler

Removing Outliers

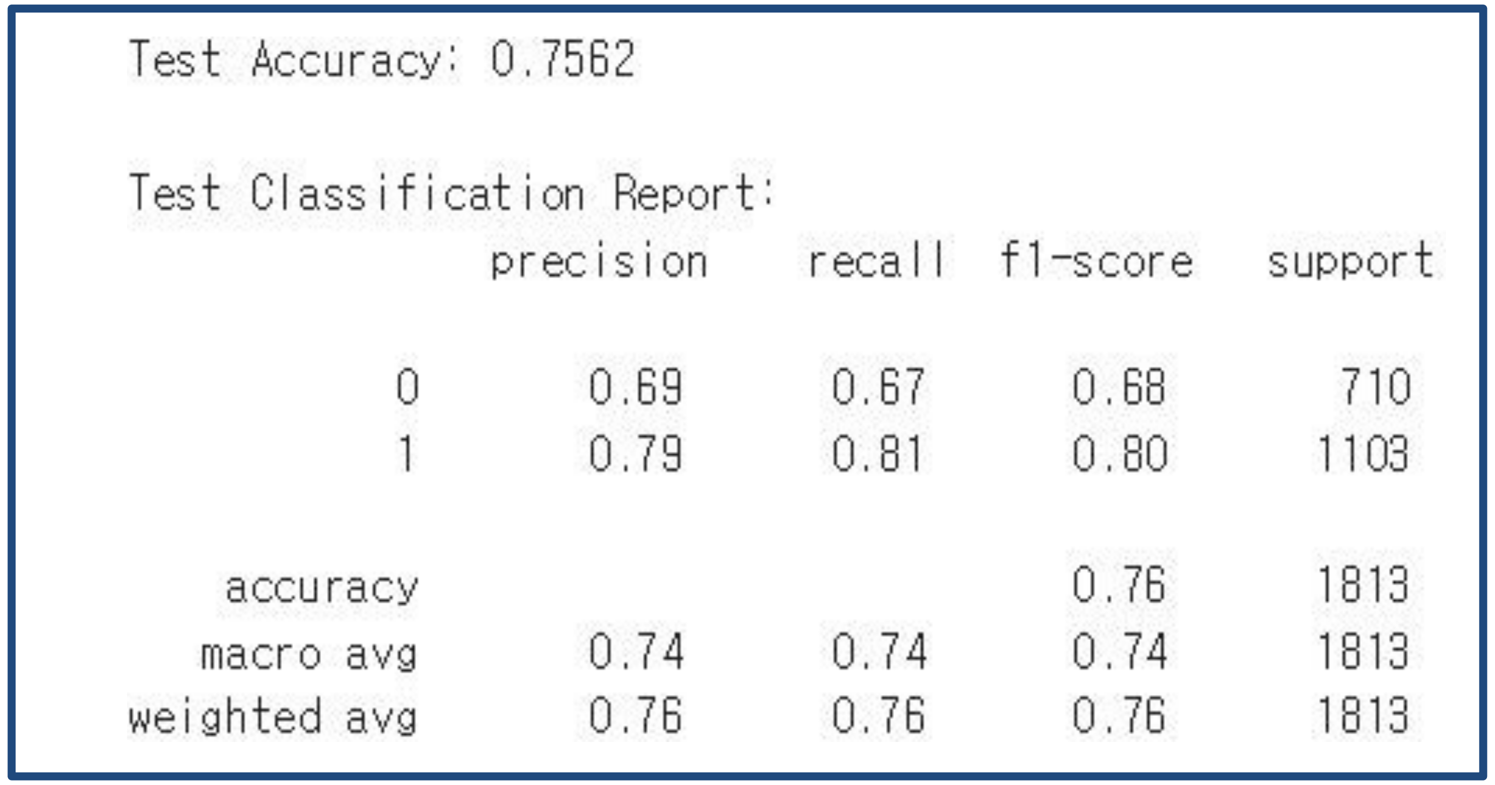
- Before removing outliers, the F1-score was 0.74.
- After applying Z-score outlier removal, the F1-score improved to 0.76.

→ ***Outlier handling improved model robustness and predictive perfiormance***



```
Test Accuracy: 0.7423

Test Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.64      0.67       771
           1       0.77      0.81      0.79      1123

    accuracy                           0.74      1894
   macro avg       0.73      0.73      0.73      1894
weighted avg       0.74      0.74      0.74      1894
```

<Before Removing Outliers>



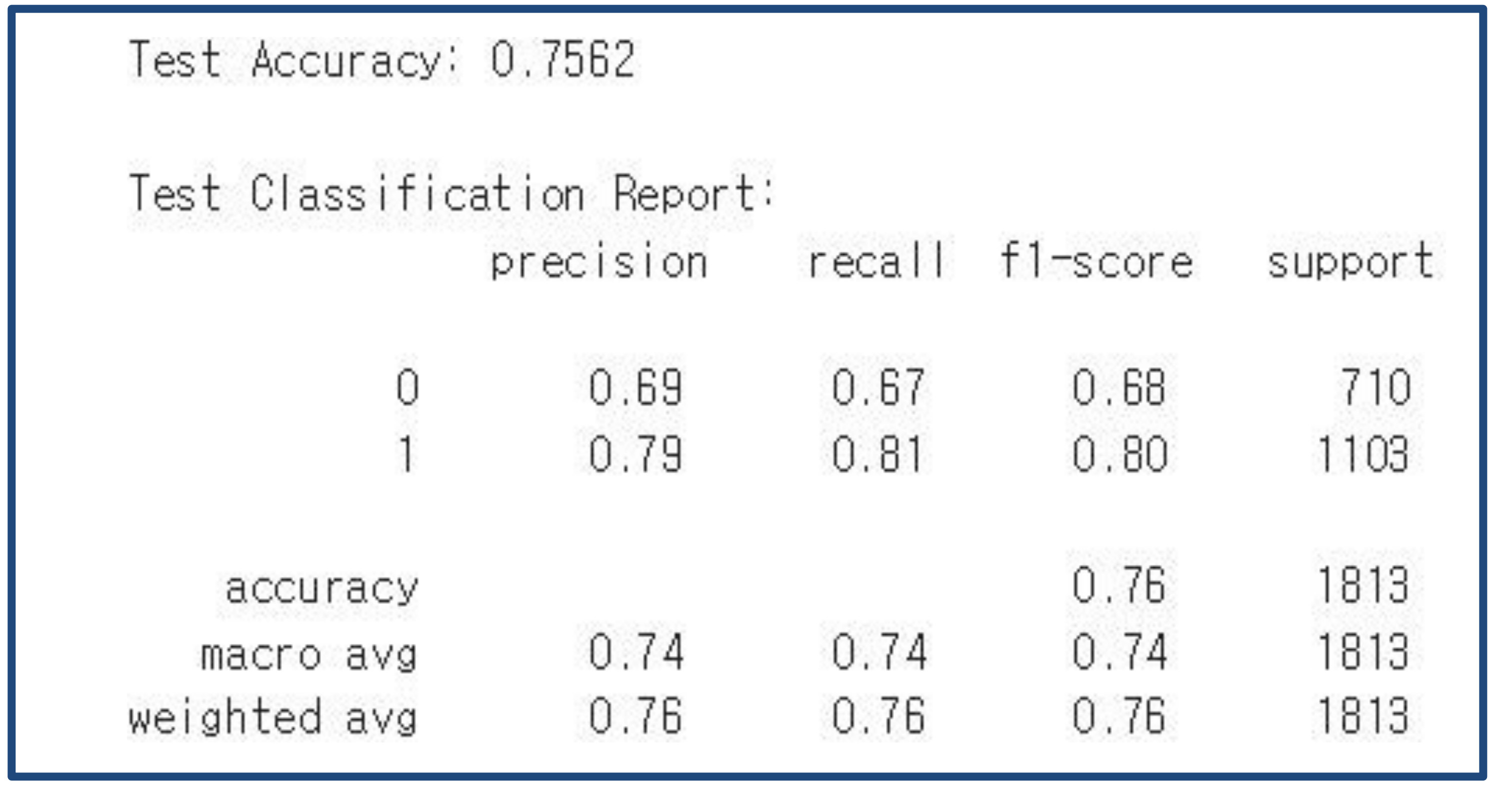
```
Test Accuracy: 0.7562

Test Classification Report:
              precision    recall  f1-score   support

           0       0.69      0.67      0.68       710
           1       0.79      0.81      0.80      1103

    accuracy                           0.76      1813
   macro avg       0.74      0.74      0.74      1813
weighted avg       0.76      0.76      0.76      1813
```

<After Removing Outliers>

Class Balancing

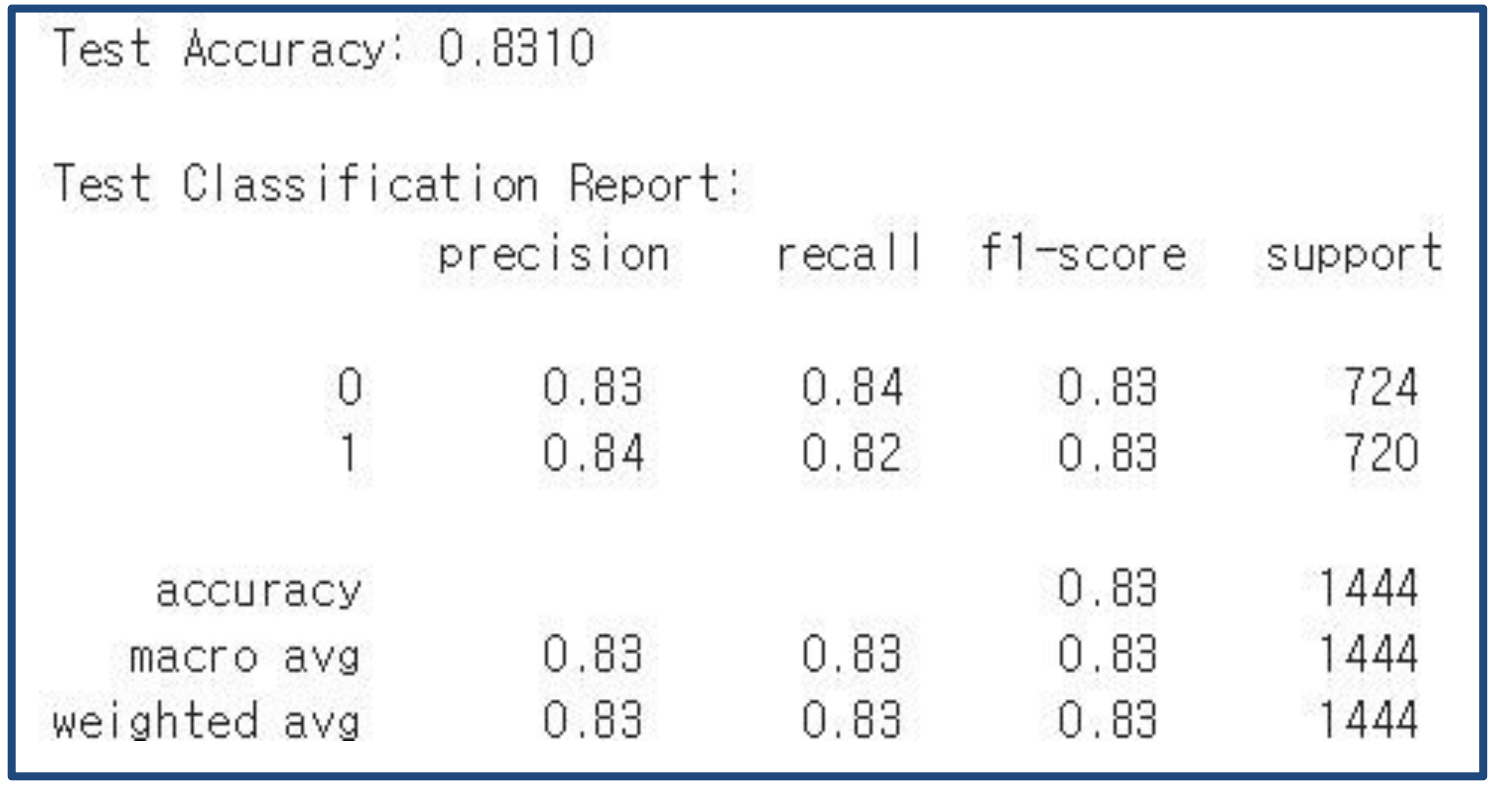
- Before balancing the classes, the F1-score was 0.76.
- After applying **RandomOverSampler** , the F1-score increased to 0.83.

→ ***Class balancing signifiicantly improved model robustness and predictive accuracy.***



```
Test Accuracy: 0.7562

Test Classification Report:
              precision    recall  f1-score   support

           0       0.69      0.67      0.68       710
           1       0.79      0.81      0.80      1103

    accuracy                           0.76      1813
   macro avg       0.74      0.74      0.74      1813
weighted avg       0.76      0.76      0.76      1813
```

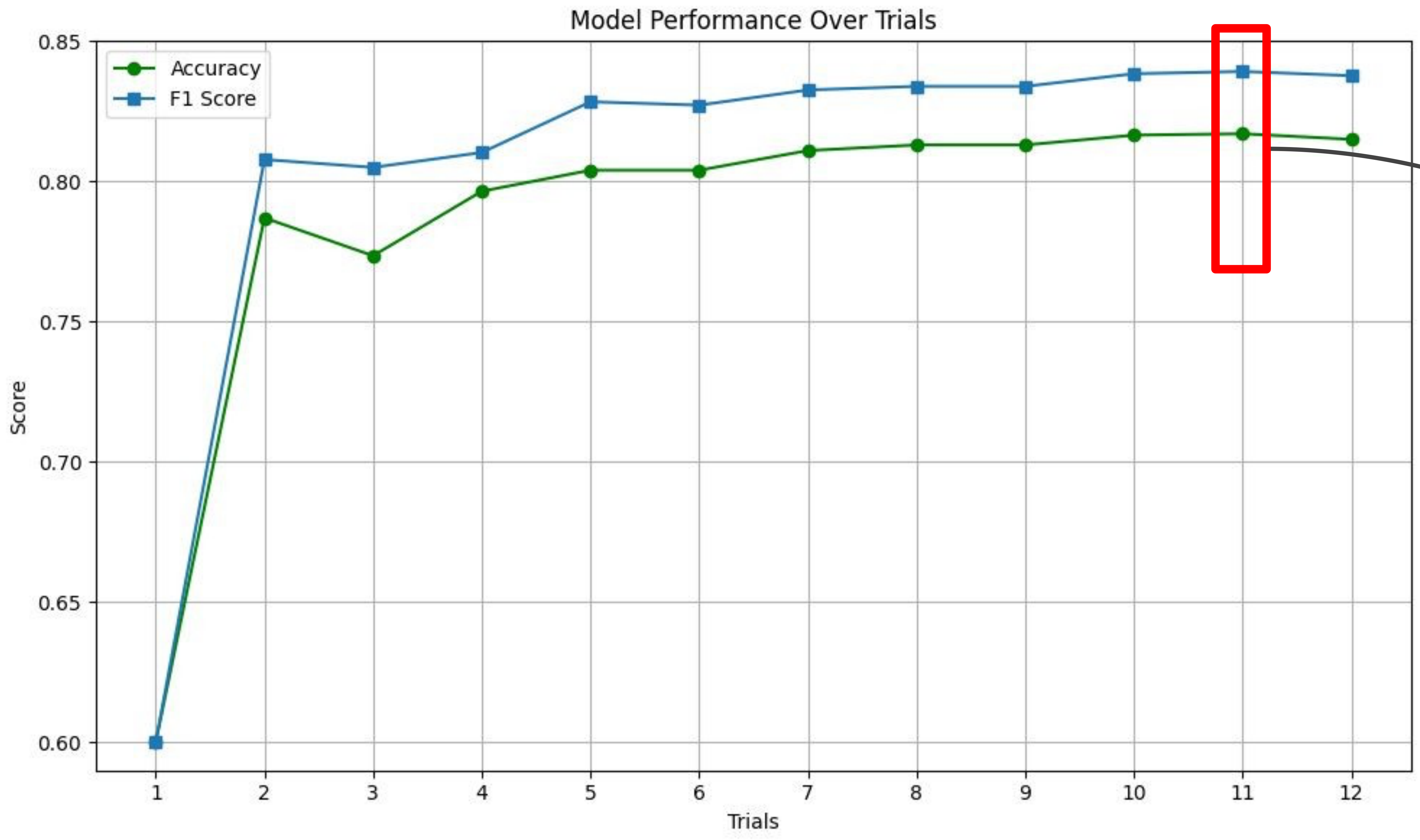
<Before Balancing the Class>



```
Test Accuracy: 0.8310

Test Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.84      0.83       724
           1       0.84      0.82      0.83       720

    accuracy                           0.83      1444
   macro avg       0.83      0.83      0.83      1444
weighted avg       0.83      0.83      0.83      1444
```

<After Balancing the Class>

Trend of Model Scores by Trial



Best Model
Accuracy: 0.8170
Precision: 0.7520
Recall: 0.9493
F1 Score: 0.8392

Project Result

Final Model

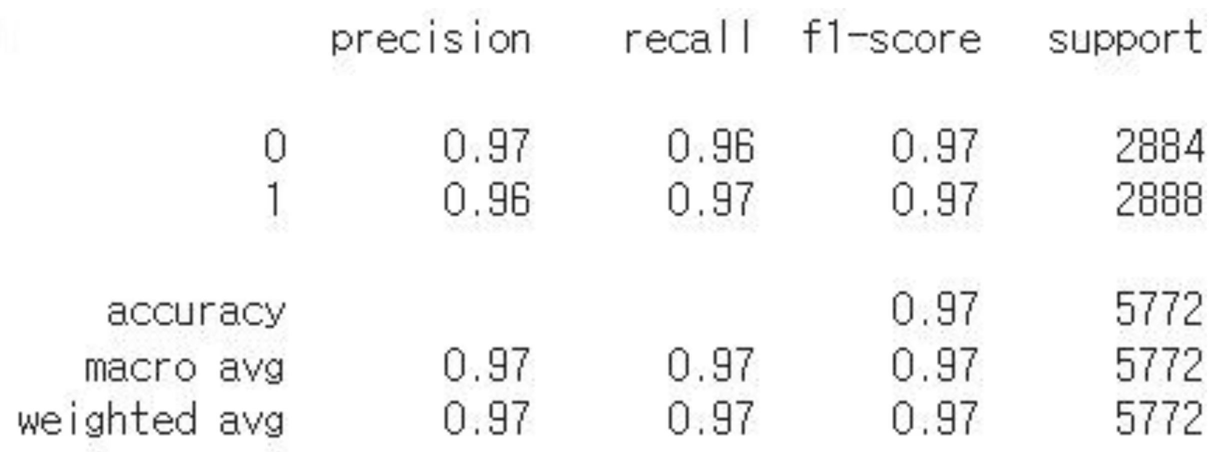
- Feature Selection: ['tmax', 'def', 'ppt', 'ws', 'q', 'soil', 'vpd', 'pet']
  - tmax: Maximum 2m temperature
  - def: Climatic water deficit
  - ppt: Accumulated precipitation
  - we: 10m wind speed
  - q: Runoff
  - soil: Soil moisture at end of month
  - vpd: Vapor pressure deficit
  - pet: Reference evapotranspiration

- Result:
  - Accuracy: 0.8170
  - Precision: 0.7520
  - Recall: 0.9493
  - F1 Score: 0.8392

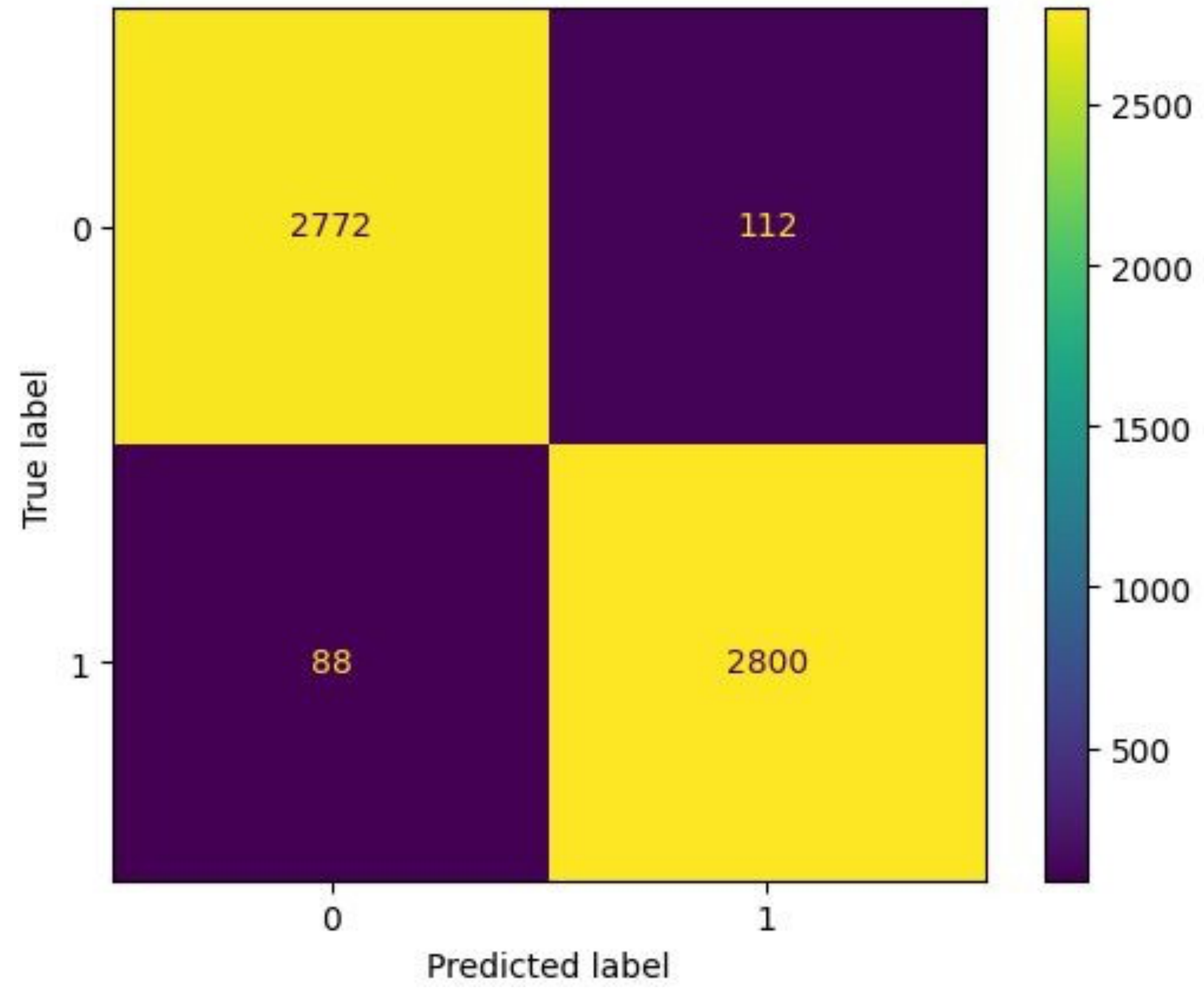
- Model: Extra Tree Classification
  - n_estimators=210
  - criterion='entropy'
  - max_depth = 27
  - bootstrap=True
  - min_samples_split = 2
  - class_weight='balanced'

Complex Matrix

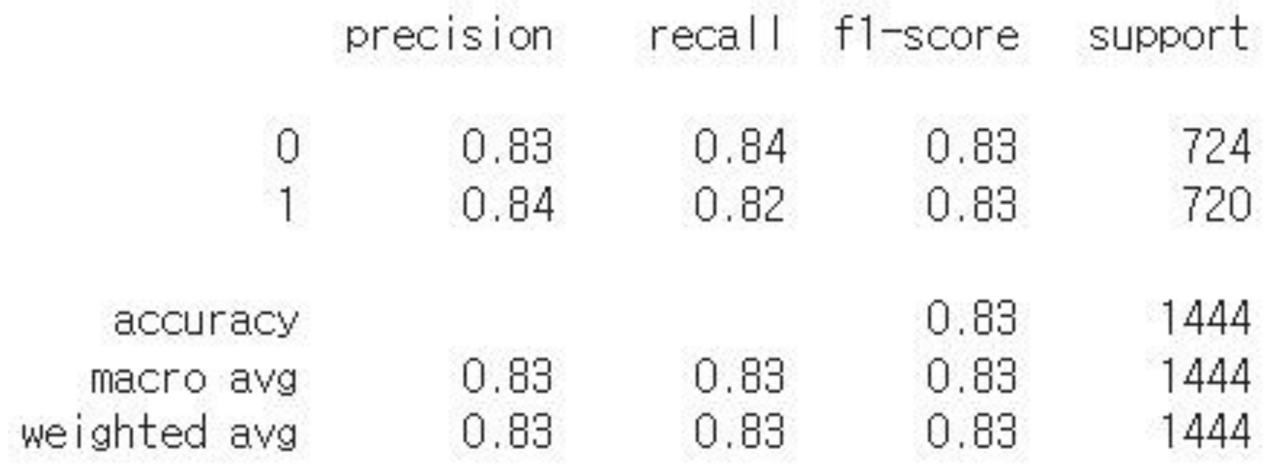
<Train Result>



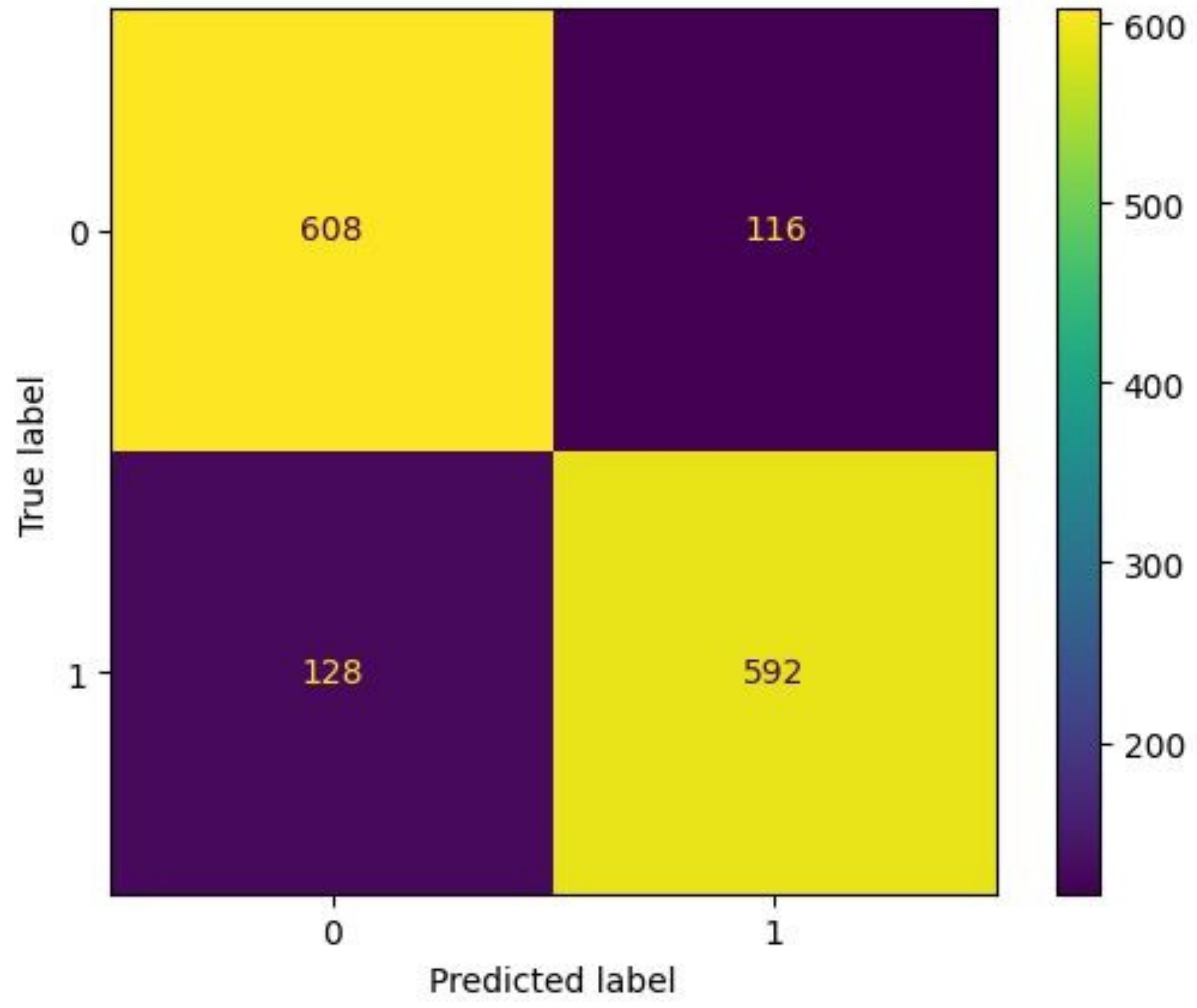
```
              precision    recall  f1-score   support

           0       0.97      0.96      0.97      2884
           1       0.96      0.97      0.97      2888

    accuracy                           0.97      5772
   macro avg       0.97      0.97      0.97      5772
weighted avg       0.97      0.97      0.97      5772
```



<Test Result>



```
              precision    recall  f1-score   support

           0       0.83      0.84      0.83       724
           1       0.84      0.82      0.83       720

    accuracy                           0.83      1444
   macro avg       0.83      0.83      0.83      1444
weighted avg       0.83      0.83      0.83      1444
```



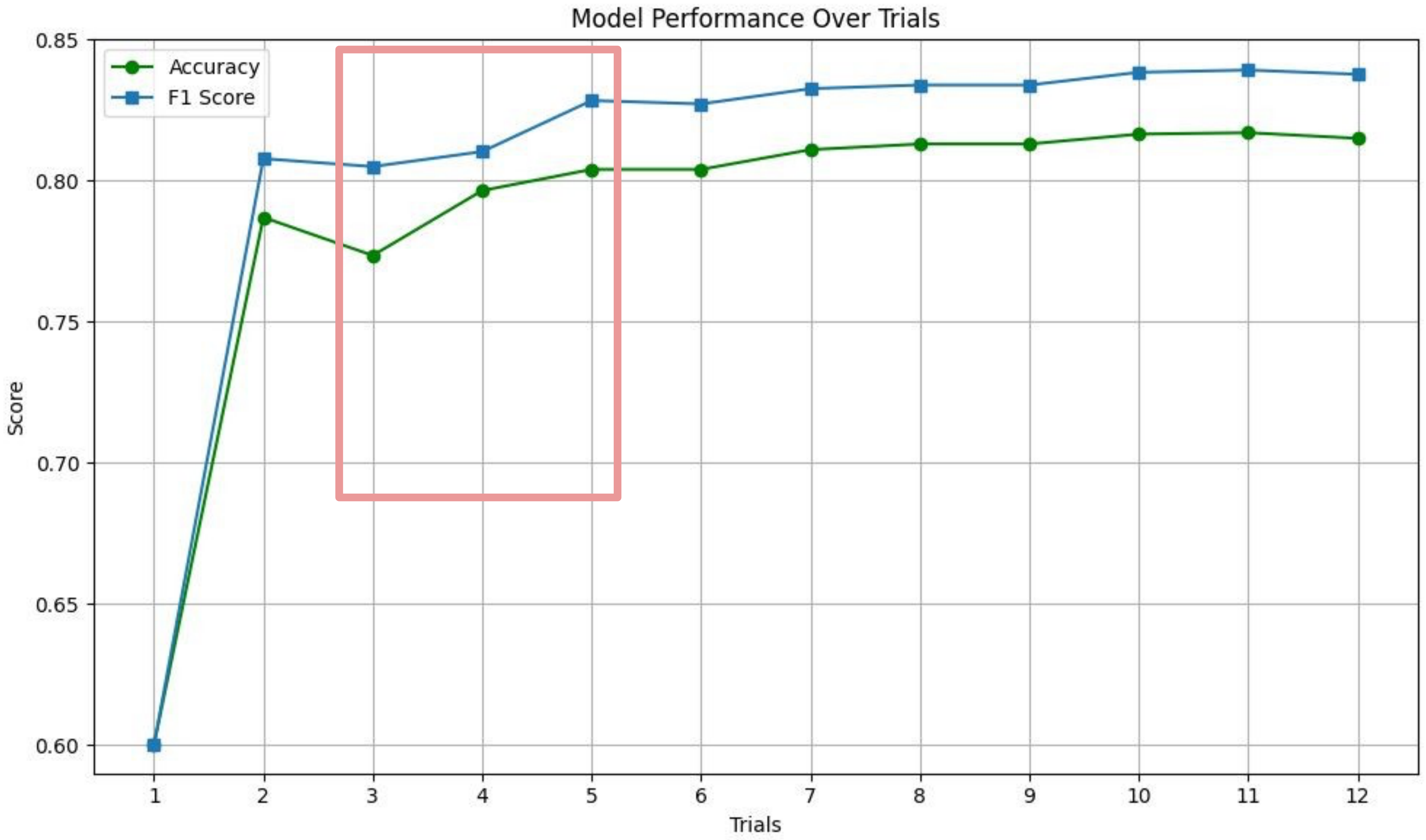
Model Result

- **Overfitting Observation**
  - The model achieved an F1-score of 0.97 on the training set, but only 0.83 on the test set.
  - This performance gap shows overfitting, but generalization remained acceptable.

- **Why We Selected This Model**
  - Despite the gap, this model achieved the highest F1-score (0.8392) on the validation set, compared to other candidates.
  - Showed consistent performance across both classes, with precision and recall balanced at 0.83.
  - Chosen for its strong real-world prediction capability

How We Improved Model

How to Improve?



How to Improve?

Various Machine Learning

- Tested multiple models including **Random Forest, XGBoost, LightGBM, Extra Trees, and Neural Network** to identify the best-performing classifier.

Finding Proper Pre-Processing

- Replaced IQR method with **Z-score** to remove extreme values without major data loss.
- Switched from SMOTE (less effective for spatial data) to **RandomOverSampler.**

Feature Selection

- Selected features based on correlation analysis and conceptual relevance.
- Conducted multiple trials with different combinations of variables.

Hyper-Parameter Tuning

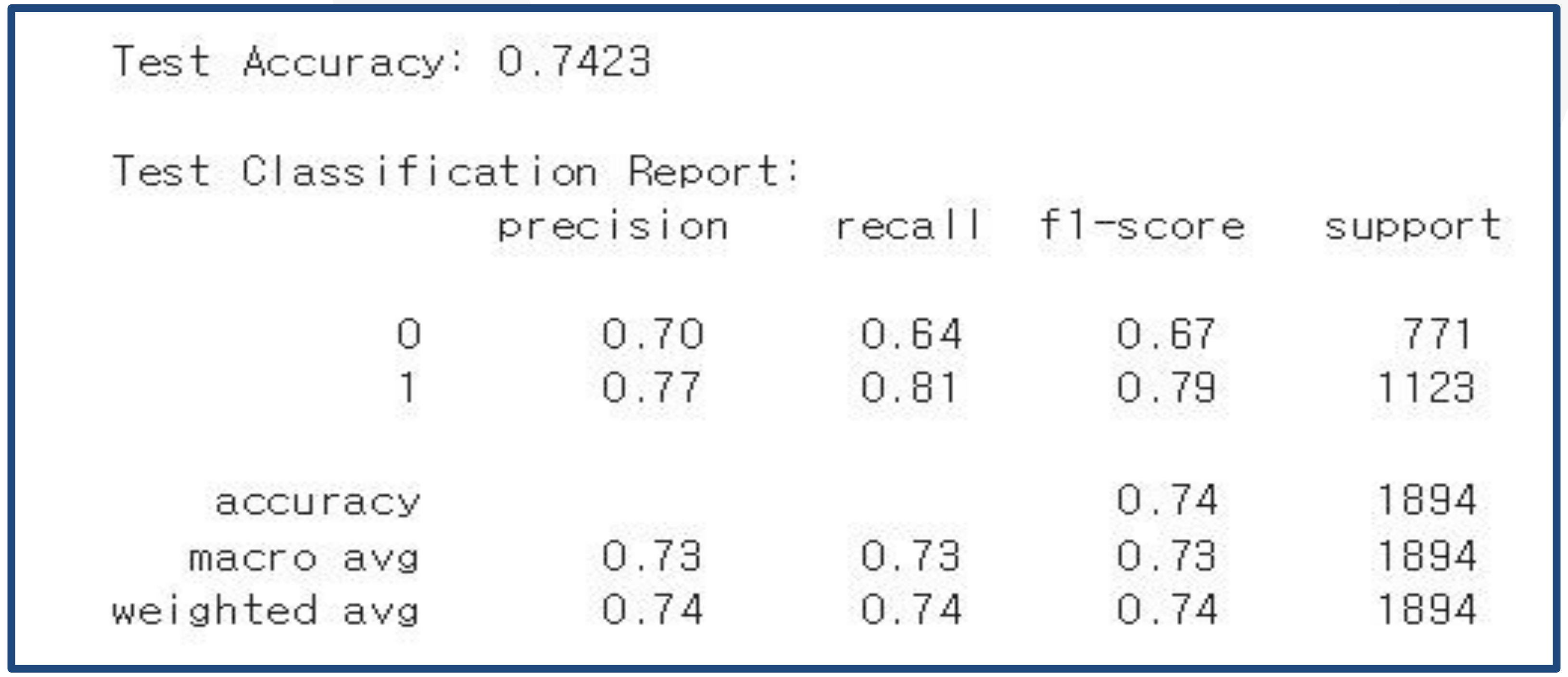
- Performed **Grid Search** to identify optimal parameter ranges for each model.
- Applied **fine-tuning** to further improve model performance and generalization

Various Machine Learning

- Tested Random Forest, XGBoost, LightGBM, and Extra Trees Classifier.
- Overall performance was similar across models.
- Extra Trees achieved the highest Test Accuracy (83.10%)

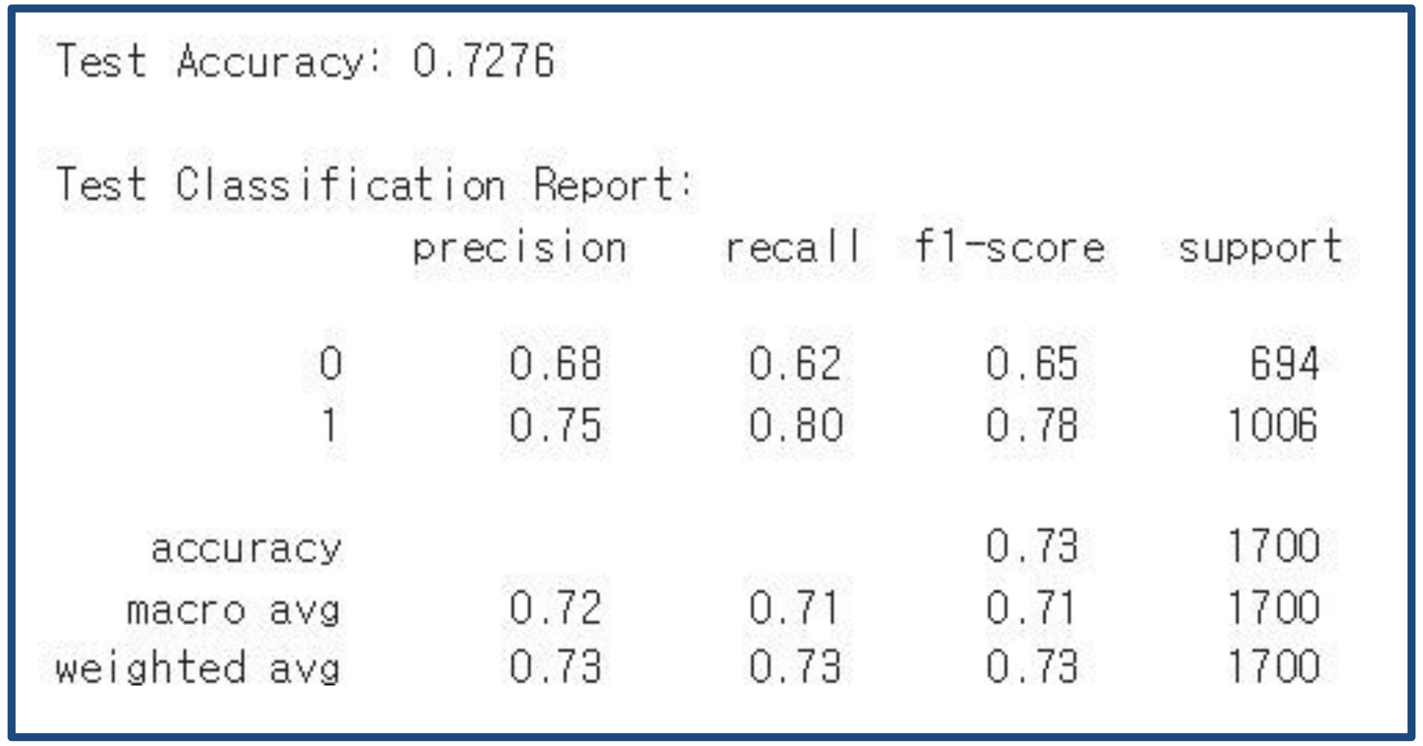
Model	F1- Score	Test Accuracy
Random Forest	0.83	0.8289
XGBoost	0.83	0.8296
LightGBM	0.82	0.8284
Extra Trees	0.83	0.8310

Finding Proper Pre-Processing - Outliers



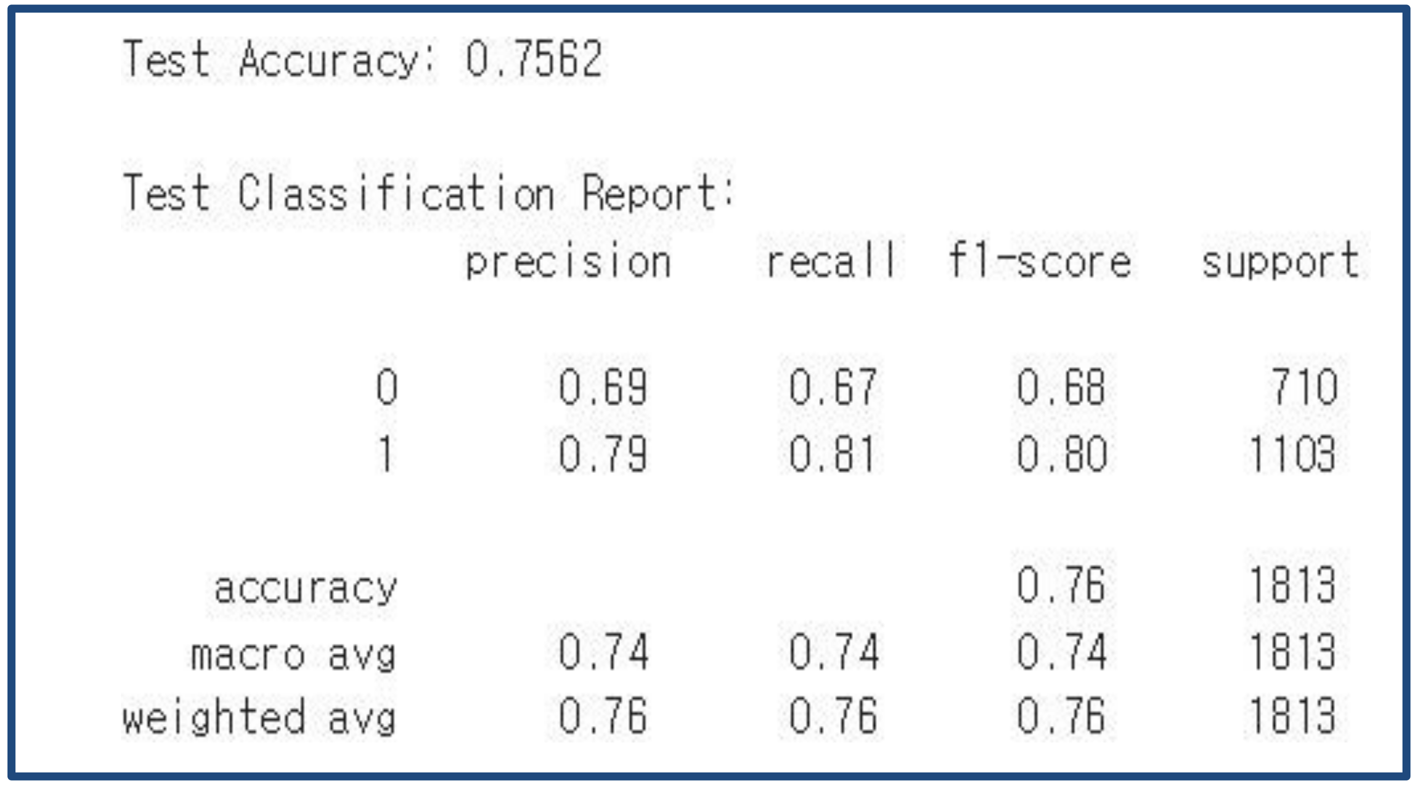
```
Test Accuracy: 0.7423

Test Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.64      0.67       771
           1       0.77      0.81      0.79      1123

    accuracy                           0.74      1894
   macro avg       0.73      0.73      0.73      1894
weighted avg       0.74      0.74      0.74      1894
```

□Before Removing Outliers>



```
Test Accuracy: 0.7276

Test Classification Report:
              precision    recall  f1-score   support

           0       0.68      0.62      0.65       694
           1       0.75      0.80      0.78      1006

    accuracy                           0.73      1700
   macro avg       0.72      0.71      0.71      1700
weighted avg       0.73      0.73      0.73      1700
```

□Removing Outliers with IQR□



```
Test Accuracy: 0.7562

Test Classification Report:
              precision    recall  f1-score   support

           0       0.69      0.67      0.68       710
           1       0.79      0.81      0.80      1103

    accuracy                           0.76      1813
   macro avg       0.74      0.74      0.74      1813
weighted avg       0.76      0.76      0.76      1813
```

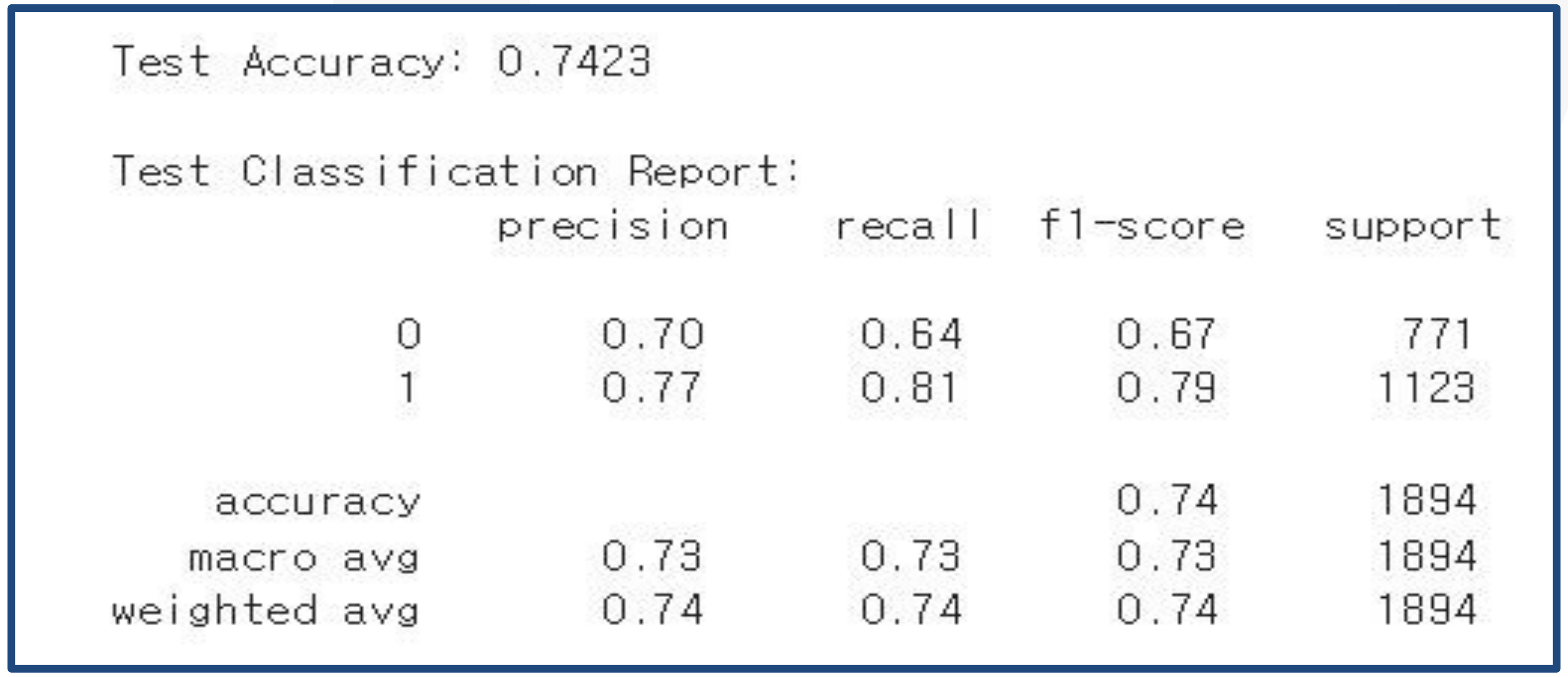
□Removing Outliers with Z-Score>

- Removing the outliers using IQR led to a decline in model performance.

- Removing the outliers using Z-score improved model performance.

→ ***Selecting the appropriate pre-processing method is essential to optimize model accuracy.***

Finding Proper Pre-Processing - OverSampling

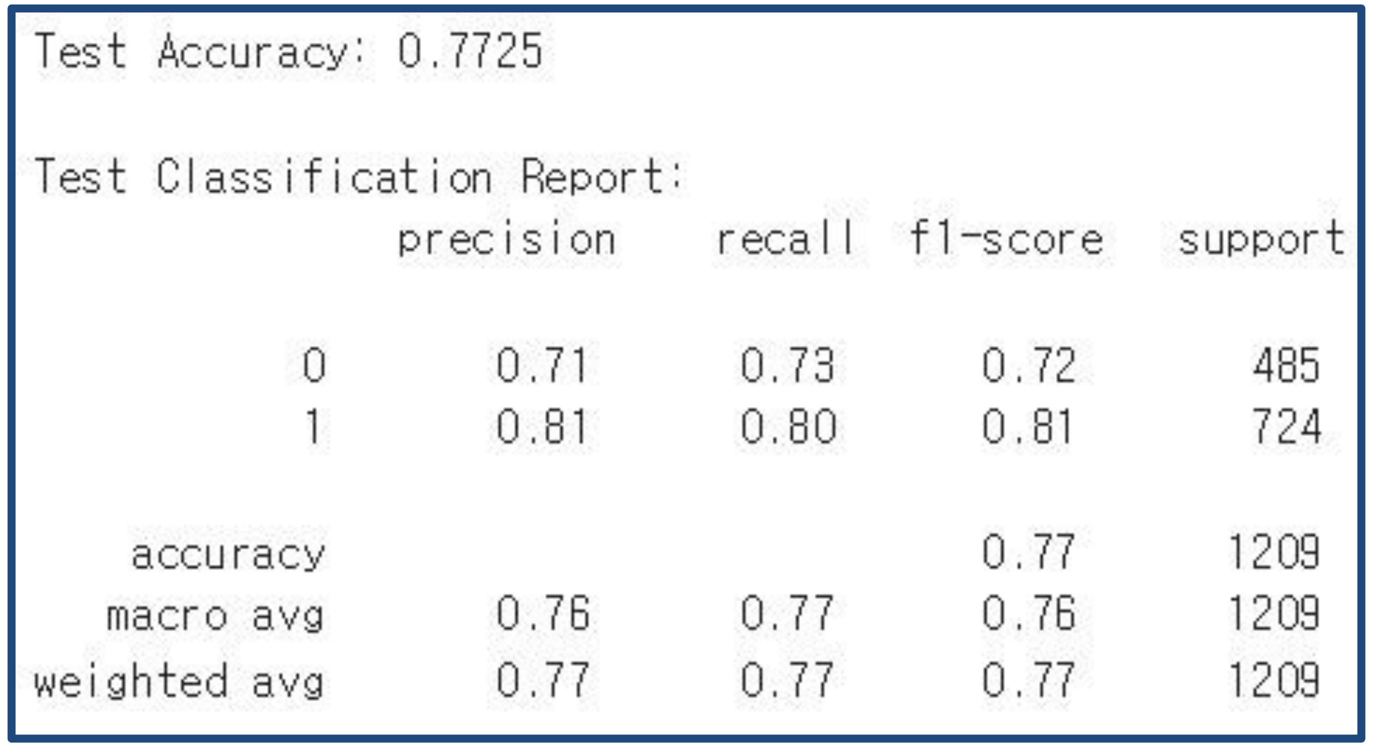


```
Test Accuracy: 0.7423

Test Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.64      0.67       771
           1       0.77      0.81      0.79      1123

    accuracy                           0.74      1894
   macro avg       0.73      0.73      0.73      1894
weighted avg       0.74      0.74      0.74      1894
```

□Before OverSampling>

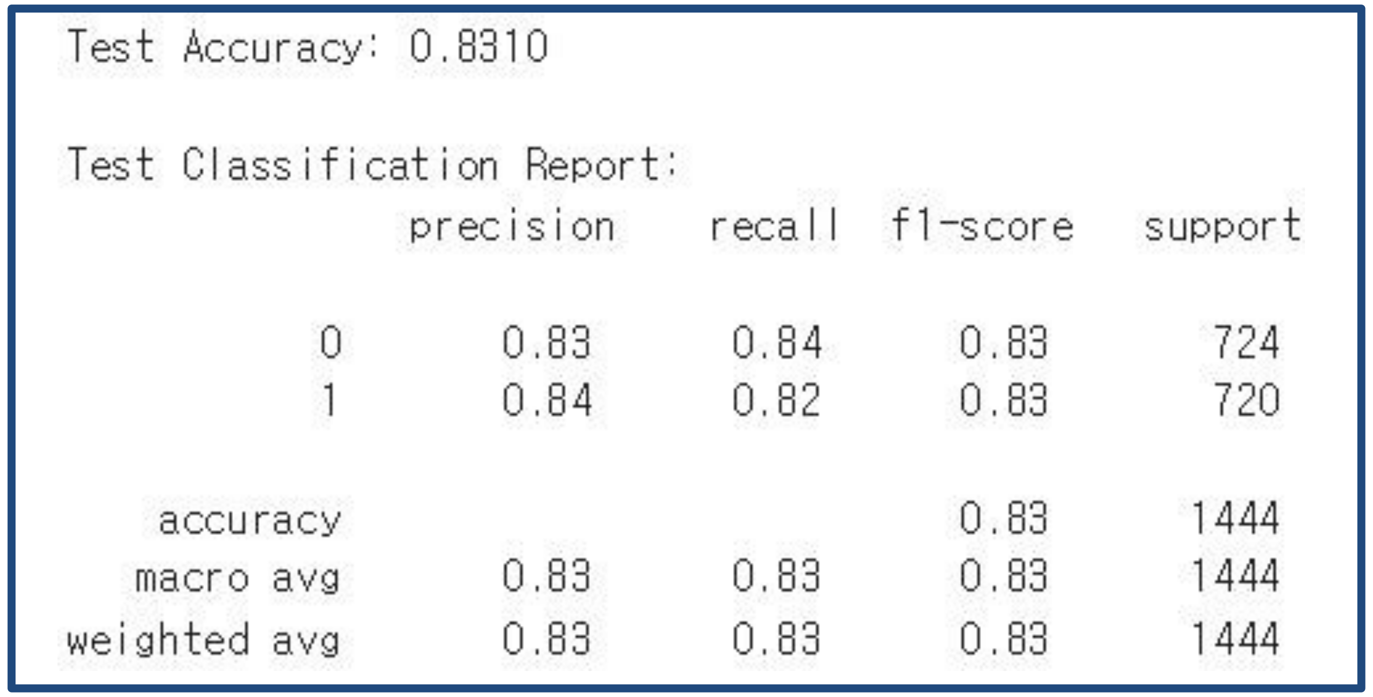
- SMOTE:
  - Slight improvement in accuracy (0.7423 * 0.7725)
  - Synthetic samples generation results in spatial distortion limited performance gain
- RandomOverSampler
  - Significant improvement in accuracy (0.8310)
  - Simple duplication without spatial distortion
  - Better suited for spatial and climate-based data

→ ***Selecting the appropriate pre-processing method is essential to optimize model accuracy.***



```
Test Accuracy: 0.7725

Test Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.73      0.72       485
           1       0.81      0.80      0.81       724

    accuracy                           0.77      1209
   macro avg       0.76      0.77      0.76      1209
weighted avg       0.77      0.77      0.77      1209
```

□Oversampling with SMOTE□



```
Test Accuracy: 0.8310

Test Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.84      0.83       724
           1       0.84      0.82      0.83       720

    accuracy                           0.83      1444
   macro avg       0.83      0.83      0.83      1444
weighted avg       0.83      0.83      0.83      1444
```

□Oversampling with RandomOverSampler>

Feature Selection

- Conducted multiple trials with different combinations of variables.

- Observed that even small differences in variable combinations led to noticeable changes in model accuracy.

→ ***Feature selection improved model accuracy and demonstrated the importance ofi fieature engineering.***

Combination of Variables	**Test Accuracy**
tmax | tmin | vap | ppt | srad | ws | pet | q | def | soil | pdsi | vpd	0.8174
tmax | tmin | ppt | ws | q | soil | vpd | pet	0.8220
def | tmin | ppt | ws | q | soil | vpd | pet	0.8269
tmax | def | ppt | ws | q | soil | vpd | pet	0.8310

Value Case

Business Application

01 Predicting Frog Trace From the Global Warming

- **Conservation Planning:** Frog habitat models help predict future biodiversity hotspots, enabling zoning, restoration, and protected area expansion to adapt to climate change.
- **Research and Policy Development:** Frog migration analysis offers insights into climate impacts, guiding ecological studies and adaptation strategies.
- **Protecting Vulnerable Ecosystems:** By translating climate data into actionable plans, governments and researchers can proactively safeguard biodiversity.

02 Predicting Frog to Protect From Batrachochytrium Under Climate Change

- **Environmental Sensitivity and Conservation:** Frogs habitat modeling helps identify regions for climate-responsive conservation and cure planning.
- **Frog Populations Under Threat:** Bd fungus, driven by climate and land use changes, poses a major threat to amphibians, causing population decline and habitat loss.
- **Actionable Insights for Protection:** The prediction model aids pharmaceutical companies, veterinarians, and researchers in locating stable frog habitats for proactive disease prevention and species preservation.

Business Application

03 Frog Occurrence Predictions to Identify Ecologically Farmland

- **Farmland Identification:** Frog occurrence models help identify farmland with ecological stability, essential for agriculture and biodiversity preservation.
- **Benefits for Farmers and Food Companies:** Frog presence indicates stable environmental conditions, aiding farmers in land selection and supporting organic certifications like USDA Organic and Rainforest Alliance.
- **Economic and Environmental Value:** Frogs contribute to farming by consuming insects and promoting sustainable agriculture, while their habitat predictions guide eco-friendly farming practices.

04 Identifying Ecological Restoration Zones for ESG Strategy

- **Ecological Restoration for ESG Strategies:** Frog occurrence analysis identifies restoration zones where biodiversity recovery is feasible, supporting environmental efforts in corporate ESG plans.
- **Real-World Example:** LG Uplus in South Korea demonstrates biodiversity conservation through initiatives like frog ladder programs in endangered habitats.
- **Corporate Benefits:** Companies can enhance ESG performance, measure biodiversity recovery, and boost their reputation with stakeholders via actionable restoration projects.

Business Case

🌿 1. Real-Time Biodiversity Monitoring Platform

- Transform the status analysis into a web-based and app-based platform with real-time API integration.
- Make continuous updates and live predictions of frog habitat changes for researchers and agencies.

🦠 2. Ecological Risk Mapping

- Predict not only frog presence but also regions vulnerable to Bd fungal infection under changing climate conditions.
- Provide actionable data for food companies and farmers selecting eco-friendly cultivation sites.

🌾 3. Biodiversity Restoration for ESG Corporations

- Helps corporations strategically to plan ecological restoration projects contributing to ESG narrative.
- Helps to increase the brand value of ESG company as an eco-friendly company contributing to long-term financial and social value.

What We Learned

What We Learned

01 The importance of Pre-processing

- Proper handling of missing values, outliers, and class imbalance significantly impacts model performance.

02 Application of Machine Learning Methods

- We could gain the knowledge about the new machine learning techniques and have a chance to apply with real data.
- Trying XGBoost, LightGBM, Random Forest, Extra Trees helped benchmark the best solution.

03 Thinking About the Business Application

- Beyond technical performance thinking about business strengthened the projects practical value.

Thank you