BagyoBAI, Typhoon Predictor

A Hybrid Machine Learning and Knowledge-Based Reasoning Project

Collaborative Final Project
CSST101 – Machine Learning
CSST102 – Knowledge Representation and Reasoning

Submitted by:
Group Name: Swayside Squad

Group Members:
- Borce, Ron Ken S.
- Luzande, Justin Angelo A.
- Quemada, John Alvin Y.

Instructor: Mr. Mark P. Bernardino

Date Submitted: January 01, 2026

PROJECT OVERVIEW

This project, named BagyoBAI, focuses on building a predictive model for the number of typhoons. The primary objective is to predict the number_of_typhoons for an upcoming period, such as the next month. This is crucial for early warning systems and disaster preparedness, aiming to provide insight into future typhoon activity based on historical patterns.

This statement encapsulates the fundamental purpose and ambition of BagyoBAI:

- **Focus on Building a Predictive Model for the Number of Typhoons**: At its heart, BagyoBAI is an applied machine learning project. It's not just about tracking current weather, but about proactively forecasting a critical metric: the sheer quantity of typhoons expected. This move from descriptive to predictive analytics is a significant step in enhancing our understanding and response capabilities to these natural phenomena.

- **Crucial for Early Warning Systems and Disaster Preparedness**: This underscores the real-world impact. By having an early indication of whether an upcoming period is likely to see an above-average, average, or below-average number of typhoons, governments, humanitarian organizations, and local communities can:
    - Proactively Allocate Resources: Deploy emergency services, medical supplies, and food aid more effectively.
    - Educate and Prepare Communities: Launch awareness campaigns about potential risks.
    - Strengthen Infrastructure: Take preventative measures for vulnerable structures.
    - Economic Planning: Industries dependent on weather, like agriculture and fisheries, can make more informed decisions.
- **Aiming to Provide Insight into Future Typhoon Activity Based on Historical Patterns**: This last part is key to the scientific and statistical underpinning of BagyoBAI. The predictive model doesn't guess; it learns from decades of past observations. By analyzing historical relationships between various climatic indices (like ENSO, SST anomalies, wind shear, humidity) and the actual occurrence of typhoons, the model identifies patterns that are too complex for human observation alone. It's about leveraging the past to illuminate the probabilities of the future, offering data-driven insights rather than mere speculation.

In essence, BagyoBAI seeks to transform raw historical weather data into actionable foresight, empowering better preparedness and mitigation strategies against the destructive force of typhoons.

OBJECTIVES

General Objective:

The BagyoBAI project is to predict the number of typhoons for an upcoming period, such as the next month, using historical weather and climate data.

Specific Objectives:

1. To build a predictive model for the number of typhoons for an upcoming period (e.g., the next month).
2. To utilize a comprehensive dataset of historical weather and climate data, including various meteorological and oceanic indices (ONI, Nino3.4_SST_anomaly, Vertical_Wind_Shear, etc.).
3. To implement a standard machine learning workflow involving data loading, cleaning, preprocessing, advanced feature engineering (rolling averages, expanding means), model training (specifically using RandomForestRegressor), and robust evaluation through backtesting with metrics like MAE and MSE.

SYSTEM ARCHITECTURE

User Input → Machine Learning Model → KRR Rules → Final Risk Level → Recommendations

1. **User Input**:

- **Description:** The process begins with the user specifying the desired **target month and year** for which they want a typhoon prediction (e.g., 'Predict typhoons for January 2026'). This is the direct query to the system.
- **BagyoBAI Component:** This is handled by the interactive part of the system that prompts the user for the target month and year, enabling them to request a specific future forecast.

2. **Machine Learning Model**:
    - **Description:** This is the core predictive engine that takes a processed set of features and generates a numerical forecast.
    - **BagyoBAI**
      Component: The **RandomForestRegressor** model. This model has been trained on historical weather and climate data and receives a carefully constructed feature vector for the month preceding the target prediction month to make its forecast.
3. **KRR Rules (Knowledge Representation & Reasoning)**:
    - **Description:** This stage involves complex feature engineering and rule-based approximations, ensuring that raw input data is transformed and incomplete future data is intelligently filled using domain knowledge and heuristics before it reaches the ML model.
    - **BagyoBAI Components:**
        - **Feature Engineering Rules:** These include the definition and application of functions to calculate rolling means and percentage differences for key variables (like number_of_typhoons, oni, sealevelpressure) and expanding means (monthly and daily averages) based on historical patterns. These transformations enrich the dataset with time-series specific insights.
        - **Heuristic Approximation Rules for Future Prediction:** When predicting for a future month where actual predictor values are unknown, BagyoBAI applies specific rules:
            - IF a feature is an observational variable needed for a future prediction THEN use its last known

        value from the historical dataset (e.g., for oni, nino3.4_sst_anomaly).
        - IF a feature is a monthly average THEN use the historical average for that specific month across all available years (e.g., for month_avg_typhoons).
        - IF a feature is previous month's typhoons THEN use the last known number of typhoons from the dataset as an approximation.
   - **Role:** These rules are critical for preparing a complete and semantically rich input vector for the ML model, even when real-time future data is unavailable, integrating essential domain-specific reasoning.
4. **Final Risk Level**:
   - **Description:** This represents the direct quantitative output of the machine learning model, indicating the expected typhoon activity.
   - **BagyoBAI Component:** The predicted number_of_typhoons (a floating-point numerical value, e.g., 2.54). This numerical value provides the primary quantitative basis for assessing the level of typhoon activity.
5. **Recommendations**:
   - **Description:** The ultimate purpose of the prediction – to inform actions and strategies. While the BagyoBAI system itself outputs a prediction, not explicit textual recommendations, its output directly supports this stage.

   - **BagyoBAI Component:** The system provides the predicted number_of_typhoons. This output is specifically designed to be consumed by **early warning systems and disaster preparedness initiatives**. This allows human experts or subsequent automated systems to formulate concrete recommendations based on the forecast, such as advising heightened vigilance, resource pre-positioning, or public awareness campaigns, especially if the prediction indicates above-average typhoon activity.

MACHINE LEARNING COMPONENT (CSST101)

Algorithm Used: Random Forest Regressor

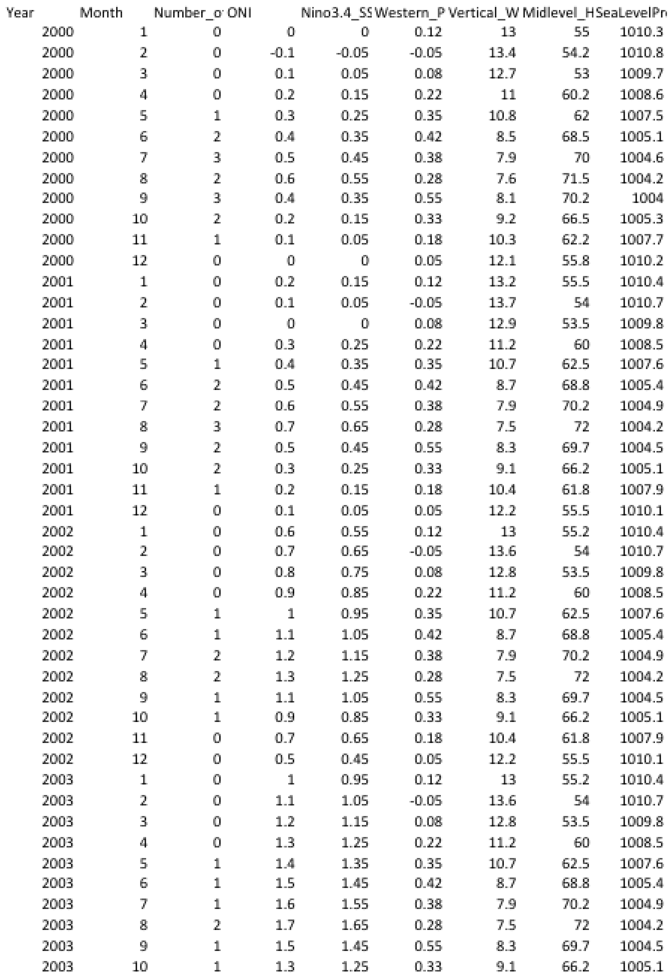
Dataset Size: The project uses a dataset loaded from /content/weather.csv. The overview does not specify the exact number of rows or columns of the dataset, but it is described as historical weather and climate data used for monthly predictions.

Model Accuracy: The Random Forest Regressor model achieved an approximate Mean Absolute Error (MAE) of 1.04 and a Mean Squared Error (MSE) of 1.94 on the backtested data.

MACHINE LEARNING PIPELINE

Data Collection:

- Source: The analysis utilizes a dataset loaded from /content/weather.csv.
- Content: This dataset includes historical weather and climate data with various meteorological and oceanic indices such as Month, Number_of_Typhoons, ONI, Nino3.4_SST_anomaly, Western_Pacific_SST, Vertical_Wind_Shear, Midlevel_Humidity, SeaLevelPressure, MJO_Phase, and Prev_month_typhoons.



Year	Month	Number_o	ONI	Nino3.4_SS	Western_P	Vertical_W	Midlevel_H	SeaLevelPr
2000	1	0	0	0	0.12	13	55	1010.3
2000	2	0	-0.1	-0.05	-0.05	13.4	54.2	1010.8
2000	3	0	0.1	0.05	0.08	12.7	53	1009.7
2000	4	0	0.2	0.15	0.22	11	60.2	1008.6
2000	5	1	0.3	0.25	0.35	10.8	62	1007.5
2000	6	2	0.4	0.35	0.42	8.5	68.5	1005.1
2000	7	3	0.5	0.45	0.38	7.9	70	1004.6
2000	8	2	0.6	0.55	0.28	7.6	71.5	1004.2
2000	9	3	0.4	0.35	0.55	8.1	70.2	1004
2000	10	2	0.2	0.15	0.33	9.2	66.5	1005.3
2000	11	1	0.1	0.05	0.18	10.3	62.2	1007.7
2000	12	0	0	0	0.05	12.1	55.8	1010.2
2001	1	0	0.2	0.15	0.12	13.2	55.5	1010.4
2001	2	0	0.1	0.05	-0.05	13.7	54	1010.7
2001	3	0	0	0	0.08	12.9	53.5	1009.8
2001	4	0	0.3	0.25	0.22	11.2	60	1008.5
2001	5	1	0.4	0.35	0.35	10.7	62.5	1007.6
2001	6	2	0.5	0.45	0.42	8.7	68.8	1005.4
2001	7	2	0.6	0.55	0.38	7.9	70.2	1004.9
2001	8	3	0.7	0.65	0.28	7.5	72	1004.2
2001	9	2	0.5	0.45	0.55	8.3	69.7	1004.5
2001	10	2	0.3	0.25	0.33	9.1	66.2	1005.1
2001	11	1	0.2	0.15	0.18	10.4	61.8	1007.9
2001	12	0	0.1	0.05	0.05	12.2	55.5	1010.1
2002	1	0	0.6	0.55	0.12	13	55.2	1010.4
2002	2	0	0.7	0.65	-0.05	13.6	54	1010.7
2002	3	0	0.8	0.75	0.08	12.8	53.5	1009.8
2002	4	0	0.9	0.85	0.22	11.2	60	1008.5
2002	5	1	1	0.95	0.35	10.7	62.5	1007.6
2002	6	1	1.1	1.05	0.42	8.7	68.8	1005.4
2002	7	2	1.2	1.15	0.38	7.9	70.2	1004.9
2002	8	2	1.3	1.25	0.28	7.5	72	1004.2
2002	9	1	1.1	1.05	0.55	8.3	69.7	1004.5
2002	10	1	0.9	0.85	0.33	9.1	66.2	1005.1
2002	11	0	0.7	0.65	0.18	10.4	61.8	1007.9
2002	12	0	0.5	0.45	0.05	12.2	55.5	1010.1
2003	1	0	1	0.95	0.12	13	55.2	1010.4
2003	2	0	1.1	1.05	-0.05	13.6	54	1010.7
2003	3	0	1.2	1.15	0.08	12.8	53.5	1009.8
2003	4	0	1.3	1.25	0.22	11.2	60	1008.5
2003	5	1	1.4	1.35	0.35	10.7	62.5	1007.6
2003	6	1	1.5	1.45	0.42	8.7	68.8	1005.4
2003	7	1	1.6	1.55	0.38	7.9	70.2	1004.9
2003	8	2	1.7	1.65	0.28	7.5	72	1004.2
2003	9	1	1.5	1.45	0.55	8.3	69.7	1004.5
2003	10	1	1.3	1.25	0.33	9.1	66.2	1005.1

Figure 1.0 weather.csv First Division Sample

Data Preprocessing:

Data Loading & Initial Inspection: The weather.csv dataset is loaded, and its structure is examined.

❖ **Data Cleaning & Preprocessing:**

- **Missing Values**: Missing values are identified and handled. Columns with more than 5% null values are dropped, and remaining missing values are forward-filled (ffill()).
- **Column Names**: Column names are standardized to lowercase.
- **Index Conversion**: The dataset index (Year) is converted to a datetime format.

- **Feature Engineering:**
    - **Target Variable**: A target variable is created by shifting the number_of_typhoons column, representing the number of typhoons in the next month.
    - **Rolling Averages**: Features like number_of_typhoons, oni, and sealevelpressure are transformed into rolling means and percentage differences over 3 and 14 time periods.
    - **Expanding Means**: Monthly and daily average features (month_avg_, day_avg_) are created for key variables to capture seasonal patterns.
    - **Final Handling**: After feature engineering, the first 14 rows are removed (due to rolling calculations), and any remaining NaN values are filled with 0.

Model Training:

- **Algorithm**: A RandomForestRegressor model is used for prediction. Initially, a Ridge regressor was explored, but the Random Forest model was later implemented.
- **Training Process**: The backtest function is employed, which iteratively trains the model on an expanding window of historical data (train = weather.iloc[:i,:]) and makes predictions on the next time step (test = weather.iloc[i:(i+step),:]).

Model Evaluation:

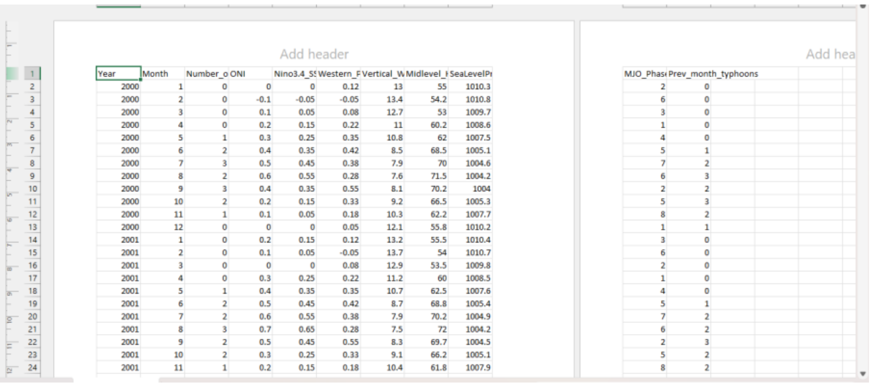
- **Methodology**: The model's performance is evaluated using backtesting on historical data.
- **Metrics**: Key metrics calculated are:
    - **Mean Absolute Error (MAE)**: Approximately 1.04.
    - **Mean Squared Error (MSE)**: Approximately 1.94.
- **Analysis**: Predictions are also sorted by their absolute difference from actual values to identify significant errors, and the distribution of prediction differences is analyzed.

Model Deployment:

- **Purpose**: The trained model is used to predict the number of typhoons for future, user-specified dates.
- **Prediction Process**: To predict for a future month, the model requires input features corresponding to the preceding month. This involves:
    - Using the last known values from the weather dataset as approximations for observational features (e.g., ONI, SST anomalies, SeaLevelPressure).
    - Approximating number_of_typhoons and prev_month_typhoons for the input month using the last known values from the dataset.
    - Utilizing historical averages for the specific input month to calculate month_avg_ features.
- **Output**: The model outputs a predicted number of typhoons for the specified future period (e.g., "Predicted number of typhoons for January 2025: 0.68").

DATASET DESCRIPTION

- **Dataset Type**: Tabular data, specifically historical weather and climate time-series data, loaded from a CSV file.
- **Number of Records**: The project overview does not explicitly state the exact number of records (rows) in the /content/weather.csv dataset. However, it is a historical dataset used for time-series predictions.
- **Target Variable**: The target variable for prediction is number_of_typhoons, which represents the number of typhoons in the next month.



Year	Month	Number_o	ONI	Nino3.4_S	Western_F	Vertical_W	Midlevel_I	SeaLevelPr	MJO_Phas	Prev_month_typhoons
2000	1	0	0	0	0.12	13	55	1010.3	2	0
2000	2	0	-0.1	-0.05	-0.05	13.4	54.2	1010.8	6	0
2000	3	0	0.1	0.05	0.08	12.7	53	1009.7	3	0
2000	4	0	0.2	0.15	0.22	11	60.2	1008.6	1	0
2000	5	1	0.3	0.25	0.35	10.8	62	1007.5	4	0
2000	6	2	0.4	0.35	0.42	8.5	68.5	1005.1	5	1
2000	7	3	0.5	0.45	0.38	7.9	70	1004.6	7	2
2000	8	2	0.6	0.55	0.28	7.6	71.5	1004.2	6	3
2000	9	3	0.4	0.35	0.55	8.1	70.2	1004	2	2
2000	10	2	0.2	0.15	0.33	9.2	66.5	1005.3	5	3
2000	11	1	0.1	0.05	0.18	10.3	62.2	1007.7	8	2
2000	12	0	0	0	0.05	12.1	55.8	1010.2	1	1
2001	1	0	0.2	0.15	0.12	13.2	55.5	1010.4	3	0
2001	2	0	0.1	0.05	-0.05	13.7	54	1010.7	6	0
2001	3	0	0	0	0.08	12.9	53.5	1009.8	2	0
2001	4	0	0.3	0.25	0.22	11.2	60	1008.5	1	0
2001	5	1	0.4	0.35	0.35	10.7	62.5	1007.6	4	0
2001	6	2	0.5	0.45	0.42	8.7	68.8	1005.4	5	1
2001	7	2	0.6	0.55	0.38	7.9	70.2	1004.9	7	2
2001	8	3	0.7	0.65	0.28	7.5	72	1004.2	6	2
2001	9	2	0.5	0.45	0.55	8.3	69.7	1004.5	2	3
2001	10	2	0.3	0.25	0.33	9.1	66.2	1005.1	5	2
2001	11	1	0.2	0.15	0.18	10.4	61.8	1007.9	8	2

Figure 2.0 weather.csv Excel Visual

KNOWLEDGE REPRESENTATION & REASONING (CSST102)

Rule 1: IF number_of_typhoons for the current month is observed THEN the target variable for the next month's prediction can be computed by shifting.

Rule 2: IF a column in the weather dataset has null_pct (percentage of null values) greater than 0.05 THEN that column is removed from the dataset to ensure data quality.

Rule 3: IF Vertical_Wind_Shear is high and Midlevel_Humidity is low THEN the conditions are less favorable for typhoon intensification.

Rule 4: IF the initial Ridge regressor model performance (MAE/MSE) is not satisfactory THEN a RandomForestRegressor model is employed for potentially improved prediction accuracy.

Rule 5: IF a prediction is requested for a target_month and target_year THEN a feature vector corresponding to the preceding month must be constructed, using approximations for currently unknown predictor values.

HYBRID DECISION LOGIC

❖ **Data-Driven Predictive Modeling (Machine Learning Core):**

  - The primary decision-making engine is the RandomForestRegressor model. This model learns complex, non-linear relationships and patterns directly from the historical weather and climate data.
  - It takes a set of engineered features as input and outputs a numerical prediction (the number_of_typhoons). The 'decision' here is statistical, based on the patterns identified during training.

- **Domain-Driven Feature Engineering (Knowledge Integration):**
  - Before the machine learning model makes its prediction, a crucial layer of decision logic is applied through feature engineering. This involves transforming raw data into more informative features based on meteorological and time-series knowledge.
  - Rolling Averages and Percentage Differences: These capture trends and momentum in key variables like number_of_typhoons, ONI, and sealevelpressure. The decision to use specific horizons (3, 14) is a knowledge-based choice.
  - Expanding Means (Monthly/Daily Averages): These features integrate seasonal patterns and historical context, allowing the model to 'understand' typical conditions for a given month or day. This implicitly adds a form of historical statistical reasoning to the input.

- **Heuristic/Rule-Based Approximations (Handling Uncertainty):**
  - When making predictions for future months, the project faces the challenge of unknown future predictor values. This is where a more explicit rule-based, or heuristic, decision logic comes into play.
  - 'Last Known Value' Rule: For many observational features (e.g., ONI, Nino3.4_SST_anomaly, SeaLevelPressure) and number_of_typhoons/prev_month_typhoons for the preceding month, the system defaults to using the last

observed values from the historical dataset. This is a pragmatic rule to provide input to the model in the absence of actual future data.

- 'Historical Monthly Average' Rule: For month_avg_ features, the system uses the historical average of the relevant variable for the specific input month (e.g., if predicting for March, it uses the historical average for February across all years). This is a more refined heuristic compared to simply taking the last known value, acknowledging seasonality.

In essence, the Hybrid Decision Logic combines the robust pattern recognition of a Random Forest model with intelligent data transformation and a set of predefined rules to bridge the gap when complete information is unavailable for future predictions.

SYSTEM FEATURES

□ Wellness risk prediction
□ Rule-based recommendations
□ Web interface / API
□ Google Colab deployment

TESTING AND EVALUATION

Test Case | Input Summary | Expected Output

Future Predictions	User inputs: Target Month = 1 (January), Target Year = 2026. This triggers the prediction logic, which constructs a feature vector for December 2025 using the last known values from the dataset and historical monthly averages.	A printed statement indicating the predicted number of typhoons for January 2026, formatted as "Predicted number of typhoons for January 2026: X.XX" (where X.XX is the calculated numerical prediction).
Model Backtesting Performance	Execution of the backtest function with the RandomForestRegressor on	The calculated Mean Absolute Error (MAE) and Mean Squared Error (MSE) from the backtesting. For

	the fully processed historical data.	example, MAE: ~1.04, MSE: ~1.94. Also, a pandas DataFrame showing actual, prediction, and diff columns for all backtested periods, sorted by the absolute difference.
Feature Vector Integrity	Examination of the prediction_features DataFrame before calling rr.predict() for a user-specified future month (e.g., predicting for March 2025, which uses February 2025 features).	The prediction_features DataFrame should contain correctly calculated month_avg_ values corresponding to the input month (e.g., February's historical average for predicting March), and month feature should be 2.0 (for February). Other features should reflect the last available data or derived rolling values.

CONCLUSION

The BagyoBAI project successfully developed a machine learning model, primarily using a RandomForestRegressor, to predict the number of typhoons for upcoming periods. It leverages a rich dataset of historical weather and climate indices, employing extensive data preprocessing and feature engineering (including rolling and expanding means). Evaluated through backtesting, the model achieved a Mean Absolute Error of approximately 1.04, demonstrating its capability to provide insights into future typhoon activity, even with necessary approximations for unknown future predictor values.

GROUP CONTRIBUTION

Member Name | Contribution

Luzande, Justin Angelo A. | Team Leader, Project Manager & System Developer, Demonstrator, Presenter

Borce, Ron Ken S. | Assistant Developer & Manager, Documentation, Presenter

Quemada, John Alvin Y. | Assistant & Team Support

REFERENCES

American Meteorological Society. (2025). MJO diversity and tropical cyclone events. Journal of Climate. https://journals.ametsoc.org/view/journals/clim/aop/JCLI-D-25-0084.1/JCLI-D-25-0084.1.pdf

NOAA Climate Prediction Center. (2025, December 1). Madden-Julian Oscillation (MJO) evolution and forecasts National Oceanic and Atmospheric Administration. https://www.cpc.ncep.noaa.gov/products/precip/CWlink/MJO/ARCHIVE/PDF/mjo_evol-status-fcsts-20251201.pdf

Golden Gate Weather Services. (n.d.). El Niño and La Niña years and intensities. https://ggweather.com/enso/oni.htm

IJCRT. (2022). Weather prediction using Random Forest method. International Journal of Creative Research Thoughts. https://ijcrt.org/papers/IJCRT2202187.pdf