Predicting Temperature in CAMANAVA Area using Multiple Regression Model (MLR)

Final Project for

COSC-E4 Elective 4 - Machine Learning

By:

Group 4

Besmonte, Kim Aldrin B.

Gallardo, Matthew

Gellido, Andrei Lois B.

Odias, Marc Lindon B.

Mejia, Juan Paulo S.

Santos, Juan Francisco P.

Submitted to:

Montaigne G. Molejon, MSIT

TABLE OF CONTENTS

ABSTRACT

Accurate temperature prediction is crucial in climatology, impacting agriculture, energy management, disaster preparedness, and environmental conservation. Despite technological advancements, traditional methods often fall short due to the complexity of meteorological factors. This project explores this challenge by employing Multiple Linear Regression (MLR) to predict temperature, incorporating variables such as atmospheric pressure, humidity, cloud cover, weather conditions, and wind speed. Using the Philippine Major Cities Weather dataset from Kaggle, which includes data collected at 3-hour intervals from November 2023 to May 2024, we preprocess the data to handle missing values and outliers. The dataset is split into training and testing sets to train and evaluate the MLR model. The model's performance is assessed using Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Results indicate that incorporating multiple meteorological factors significantly improves prediction accuracy, with the model yielding an MAE of 1.50, MSE of 3.70, and RMSE of 1.92. This study demonstrates MLR's capability to provide a more comprehensive and reliable temperature forecasting tool, with implications for meteorology and environmental science.

Keywords: temperature prediction, multiple linear regression, meteorological factors, climatology, weather forecasting

INTRODUCTION

Every day, the news provides us with weather forecasts for the upcoming days, indicating whether it will be sunny, cloudy, windy, rainy, or cold. In the ever-evolving field of climatology, accurate temperature prediction is crucial for numerous applications, ranging from agriculture and energy management to disaster preparedness and environmental conservation. Understanding and forecasting temperature variations not only aid in day-to-day activities but also enhance our ability to mitigate the impacts of climate change (Holmberg, 2023).

Despite advancements in technology, accurate temperature prediction remains a challenging problem due to the complex interplay of various meteorological factors. Traditional methods often fall short in capturing this complexity, leading to less reliable forecasts.

Temperature forecasting involves various factors, including atmospheric pressure, humidity, cloud cover, weather conditions and wind speed. One way to analyze these factors and predict temperature is linear regression. Linear regression is a key data science tool for predicting continuous outcomes. It predicts the relationship between variables by assuming that they have a straight-line connection. It finds the best line that minimizes the differences between predicted and actual values (Mali, 2024). However, simple linear regression only models a single independent variable. This can be overly simplistic and fail to capture the complexity of real world data where outcomes are often influenced by multiple factors.

To address this issue, Multiple Linear Regression (MLR) can be used. MLR models the relationship between a dependent variable and two or more independent variables. In predicting temperature, it allows us to consider multiple meteorological factors like atmospheric pressure, humidity, cloud cover, weather conditions and wind speed, which will be used in the study, simultaneously to provide a more accurate and comprehensive forecast. A study by Noi et. al. (2021) applied multiple linear regression to forecast daily maximum and minimum temperatures using historical weather data. The study found that incorporating variables such as ast temperature, humidity and wind speed significantly improved prediction accuracy.

The project will employ MLR to predict temperature using a dataset composed of various meteorological variables. The data will be preprocessed to handle missing values and outliers, and then split into training and testing sets. The MLR model will then be trained on the training set and evaluated on the testing set to measure its accuracy.

The aim of the project is to explore multiple linear regression model’s capability to predict the temperature accounting for multiple influencing factors, providing a more comprehensive and reliable forecasting tool.

RELATED WORKS

Understanding Machine Learning in Weather Forecasting

In recent years, the integration of machine learning techniques into various fields has garnered significant interest, particularly in weather forecasting. The ability to predict weather patterns with high accuracy has profound implications for agriculture, disaster management, and daily life.

In a study conducted by Holmstrom, Liu, and Vo (2016) on applying machine learning techniques to weather forecasting, the researchers focused on predicting maximum and minimum temperatures over a seven-day period using weather data from the preceding two days. They compared the efficacy of a linear regression model with a variation of a functional regression model. The findings indicated that although both models were outperformed by professional weather forecasting services, the accuracy gap lessened for forecasts made for later days, suggesting the potential of machine learning models for longer-term forecasts. Interestingly, the linear regression model outperformed the functional regression model, likely due to the limited two-day data period, which was insufficient for capturing significant weather trends in the functional regression model. The authors suggested that using data from a longer period, such as four or five days, might enable the functional regression model to outperform the linear regression model.

Whereas in the same year, Paras and Mathur (2016) introduced a straightforward approach to weather forecasting using Multiple Linear Regression (MLR), aimed at enhancing accessibility for moderately educated farmers. Their study focused on predicting maximum temperature, minimum temperature, and relative humidity based

on time-series weather data. Paras and Mathur demonstrated that MLR effectively forecasts future weather conditions by deriving features from correlation values within the weather data series. Their approach also included predicting relative humidity using maximum and minimum temperatures and rainfall, and categorizing rainfall based on derived features. The study emphasized practicality by utilizing simple data processing tools like MS Excel for model development and validation, catering to users with limited technical expertise. Results indicated that MLR models reliably predicted weather conditions, with optimal data collection periods identified as 15 weeks for temperature forecasts and 45 weeks for humidity predictions. This research highlighted MLR's potential in developing accessible and dependable weather forecasting tools tailored for localized applications.

In a similar study, Fang and Lahdelma (2016) evaluated a Multiple Linear Regression (MLR) model alongside a Seasonal Autoregressive Integrated Moving Average (SARIMA) model for forecasting heat demand in district heating systems. Their study utilized hourly weather data (outdoor temperature and wind speed) and hourly heat consumption data from Espoo, Finland. The MLR model was enhanced by incorporating a weekly rhythm of heat consumption to account for social components, significantly improving accuracy. The SARIMA model combined exogenous weather variables with historical heat consumption data, offering high accuracy for both long-term and short-term forecasts. The evaluation showed that the proposed MLR model (T168h), which considers a 168-hour demand pattern with midweek holidays classified as weekends, achieved the highest accuracy and robustness among all tested models. The study concluded that, due to its simplicity, ease of use, and high accuracy,

the T168h model is the most practical for heat demand forecasting. They suggested that the model could be further refined for individual buildings if automated meter reading data were available, enabling more precise and real-time forecasts.

A study by Menon et al. (2017) titled “Prediction of Temperature using Linear Regression” explores the impact of the urban heat island effect, using temperature as the independent variable and pollution and population as dependent variables. The study assesses the accuracy of the predictions by comparing them to actual values from 2013 to 2016. The researchers concluded that the implemented multiple linear regression analysis was accurate based on current statistics, with measured values of 23.9, 24.2, 23.9, and 24.4, and predicted values of 23.92, 24.20, 23.94, and 24.38, respectively.

Building upon this foundation, Anusha, Chaithanya, and Reddy (2019) investigated the application of the Multi-Linear Regression (MLR) algorithm for weather prediction, highlighting its superiority over traditional statistical approaches such as Support Vector Machine (SVM). They observed that conventional systems often struggle to capture sudden weather changes due to their reliance on generalized equations. In contrast, the MLR approach considers the impact of each individual parameter, resulting in more precise predictions. The study compared the accuracy of various methods, revealing that the MLR approach achieved an 88% accuracy rate, surpassing SVM (75%), Bayesian Enhanced Modified Approach (BEMA) (80%), and other models. The researchers emphasized that their proposed technique optimizes results by accurately assessing independent variables like temperature, wind speed, wind direction, humidity, and atmospheric pressure in predicting rainfall. Their findings

underscored the MLR method's significant enhancement in weather forecast precision compared to traditional models.

Contemporary Approaches to Weather Forecasting

In the review of Cifuentes, Marulanda, Bello, and Reneses (2020) discuss the increasing efforts over the past decade to understand historical climate change's impact on global and regional levels, emphasizing the importance of accurate air temperature estimation for agricultural, ecological, environmental, and industrial sectors. The study reviews various machine learning strategies for temperature forecasting, highlighting their advantages and disadvantages, and identifying research gaps. It demonstrates that machine learning techniques can effectively predict temperatures using inputs like past temperature values, relative humidity, solar radiation, rainfall, and wind speed.

Trieu, Huynh, Rodin, and Pottier (2021) explored the application of interpretable machine learning techniques to meteorological data, with an emphasis on explaining the characteristics and relationships within the data used for weather forecasting. Traditionally, weather prediction has relied on physical models of atmospheric dynamics, but the computational intensity of solving complex fluid dynamics equations has driven interest in machine learning approaches. Their research highlighted the importance of understanding how different features influence weather predictions. For instance, they found that features such as relative humidity, cloud amount, and the height of the cloud base were significant predictors. Using Shapley values, they demonstrated the contributions of individual features to specific predictions, which helps in interpreting and improving the model's performance. The study concluded that

leveraging machine learning models for interpreting meteorological data can enhance the understanding of feature importance and interactions, ultimately leading to more accurate weather forecasts.

Meanwhile in the same year, Karna et al. (2021), conducted a study using statistical linear regression and linear regression with elastic net and hyperparameters. They observed the Mean Absolute Difference for training and testing data across various cases. The correlation between predicted and actual data was assessed using Mean and Standard Deviation. The study found that regression models for predicting long-term maximum temperatures achieved better accuracy with lower Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). It concluded that these regression models performed well for long-term (monthly) temperature predictions.

The paper authored by Gupta et al. (2022), "MLRM: A Multiple Linear Regression Based Model for Average Temperature Prediction of A Day", focuses on using multiple linear regression (MLR) to predict the average daily temperature. The researchers utilized past meteorological data to build the model, which aims to provide accurate temperature predictions. The performance of the model was evaluated, and it was found that the MLR model could predict the average temperature of a day with an error margin of 2.8 °C, demonstrating its efficacy in this application.

In the study of Ismaila, Muhammed and Adamu (2022) modeling land surface temperature in urban areas using spatial regression models, employs spatial lag models (SLM) and spatial error models (SEM) using Landsat-8 OLI/TIRS data and digital elevation models to generate the dependent and explanatory variables. Initially, the study assesses the correlation between these variables and LST, along with the

presence of spatial autocorrelation, followed by validation of the modeled LST. Findings indicate that built-up areas, green areas, and water bodies exhibit lower LST compared to non-urbanized areas. The SLM predicted LST ranges from 20 to 42.9 °C, slightly below the original data range. SEM predictions range from 20.4 to 42.2 °C, closely matching the original data. At a 0.01 significance level, all variables except elevation are significant predictors of LST. Although both models perform well, SEM demonstrates superior performance. The study's outcomes provide urban planners with insights for interventions to mitigate surface temperature in urban areas.

Furthermore, in the study of Guermoui, Abdelaziz, Gairaa, Djemoui, and Benkaciali (2020) introduce a novel model for predicting daily horizontal global solar radiation using only air temperature as an input, leveraging support vector regression (SVR). Their method combines the outputs of two SVR models: the first estimates the horizontal global solar radiation, while the second estimates the error from the first model. These estimates are then combined using a weighted sum to produce the final prediction, termed the Corrected-SVM model. This approach was evaluated using three years of data (2013–2015) from Ghardaïa, Algeria, a region with semi-arid climate conditions. The Corrected-SVM model demonstrated improved performance over the conventional SVM model, with better statistical parameters, including mean absolute bias error (MABE) of 1.623 MJ/m², root mean square error (RMSE) of 2.286 MJ/m², relative square error of 11.35%, and a correlation coefficient (r) of 94.20%, compared to the conventional SVM model's MABE of 1.713 MJ/m², RMSE of 2.575 MJ/m², relative square error of 12.61%, and a correlation coefficient of 93.06%.

Similarly, Nazeri-Tahroudi and Ramezani (2020) present a study on estimating dew point temperature (DPT) across different climates in Iran using support vector regression (SVR) optimized by the ant colony algorithm. They utilized meteorological data from six stations (Ahvaz, Urmia, Kerman, Gorgan, Rasht, and Babolsar) and evaluated four different input patterns for the SVR model. These patterns included various combinations of monthly meteorological data: Pattern I with seven inputs (minimum, maximum, and average air temperatures; precipitation; saturation vapor pressure; actual vapor pressure; and relative humidity), Pattern II with three inputs (average air temperature, saturation vapor pressure, and actual vapor pressure), Pattern III with two inputs (minimum and maximum air temperatures), and Pattern IV with one input (average air temperature). The study concluded that Pattern III, which uses monthly minimum and maximum air temperatures, is the most suitable for estimating DPT values based on root mean square error (RMSE), Nash–Sutcliffe model efficiency coefficient (NSE), and the coefficient of determination (R^2). This pattern improved model accuracy by up to 24% compared to the conventional model.

Hsu et al. (2020) in their study “New land use regression model to estimate atmospheric temperature and heat island intensity in Taiwan” explore the spatial-temporal variability of atmospheric temperature across Taiwan, an island with diverse local emission sources influenced by its Asian cultural characteristics. The study developed a new land use regression (LUR) model using temperature data from the Taiwan Central Weather Bureau collected between 2000 and 2016, with 2017 data used for external verification to ensure model reliability. Recognizing the cultural-specific emission sources such as incense and joss money burning, the study included the

locations of temples, cemeteries, and crematoriums as potential predictors. The overall model performance demonstrated high predictive capability with an R^2 of 0.88 and a tenfold cross-validated R^2 of 0.87. Using this LUR model, the researchers estimated urban heat island intensity (UHII) for six metropolises in Taiwan, finding Taichung City to have the highest UHII value of 4.60 °C. These findings provide significant insights into the application of remote sensing for studying the spatial-temporal variation of atmospheric temperature and its impact on UHI effects.

OBJECTIVES

The general objective of this project is to develop a multiple linear regression model to predict temperature based on various meteorological features such as atmospheric pressure, humidity, cloudiness, weather condition, and wind speed.

Specifically, this project aims to answer the following questions:

1. How does temperature vary with atmospheric pressure, humidity, cloudiness, weather condition, and wind speed?

   This can help identify the relationship between temperature and these independent variables.

2. Can the model predict the temperature based on these independent variables?

   The regression model can be used to predict the temperature given values of atmospheric pressure, humidity, cloudiness, weather condition, and wind speed.

SYSTEM DESIGN AND METHODOLOGY

Source of Data

The dataset is the Philippine Major Cities Weather data from Kaggle.com. The data itself was collected from the website https://openweathermap.org/, starting from the month of November 2023 until May 2024. It contains various meteorological and geographical features collected at 3-hour intervals like temperature, atmospheric pressure, humidity, cloud cover, wind speed, weather conditions, latitude, longitude, ground level, sea level, etc.

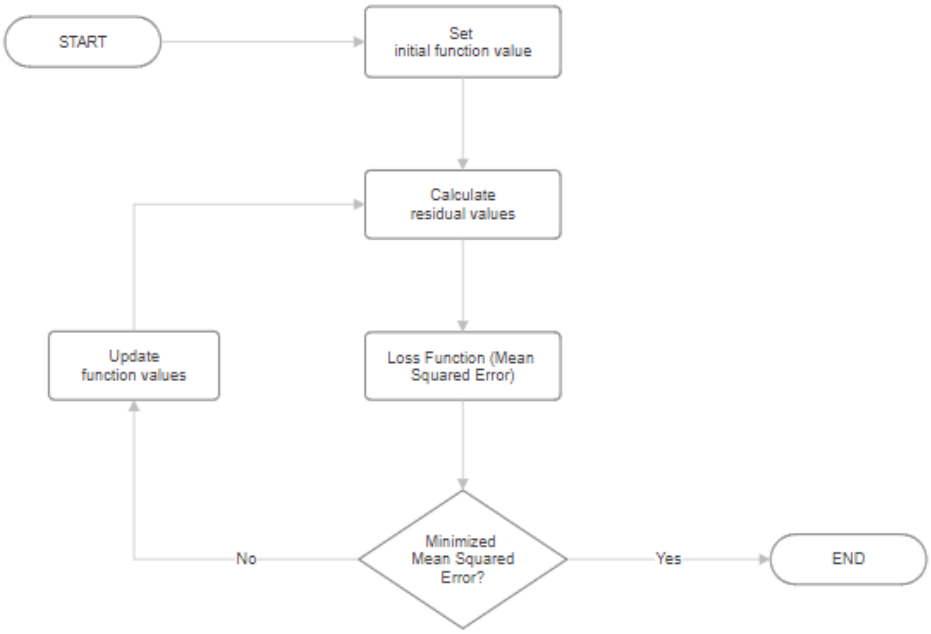
Research Instrument

The research instrument utilized in this study includes a dataset sourced from Kaggle, providing comprehensive meteorological data, including temperature, atmospheric pressure, humidity, cloud cover, wind speed, and weather conditions, collected at regular intervals. Leveraging the Jupyter Notebook environment and the Python programming language, we conducted data analysis, preprocessing, model training, and evaluation. Python libraries such as pandas, scikit-learn, matplotlib, seaborn, and joblib were instrumental in efficiently handling the dataset, implementing the Multiple Linear Regression (MLR) model, visualizing the model's performance, and saving the trained model for future use.

This integrated approach enabled a thorough analysis of meteorological data and the development of an accurate temperature prediction model, with implications for diverse fields such as meteorology, environmental science, and beyond.

Algorithm

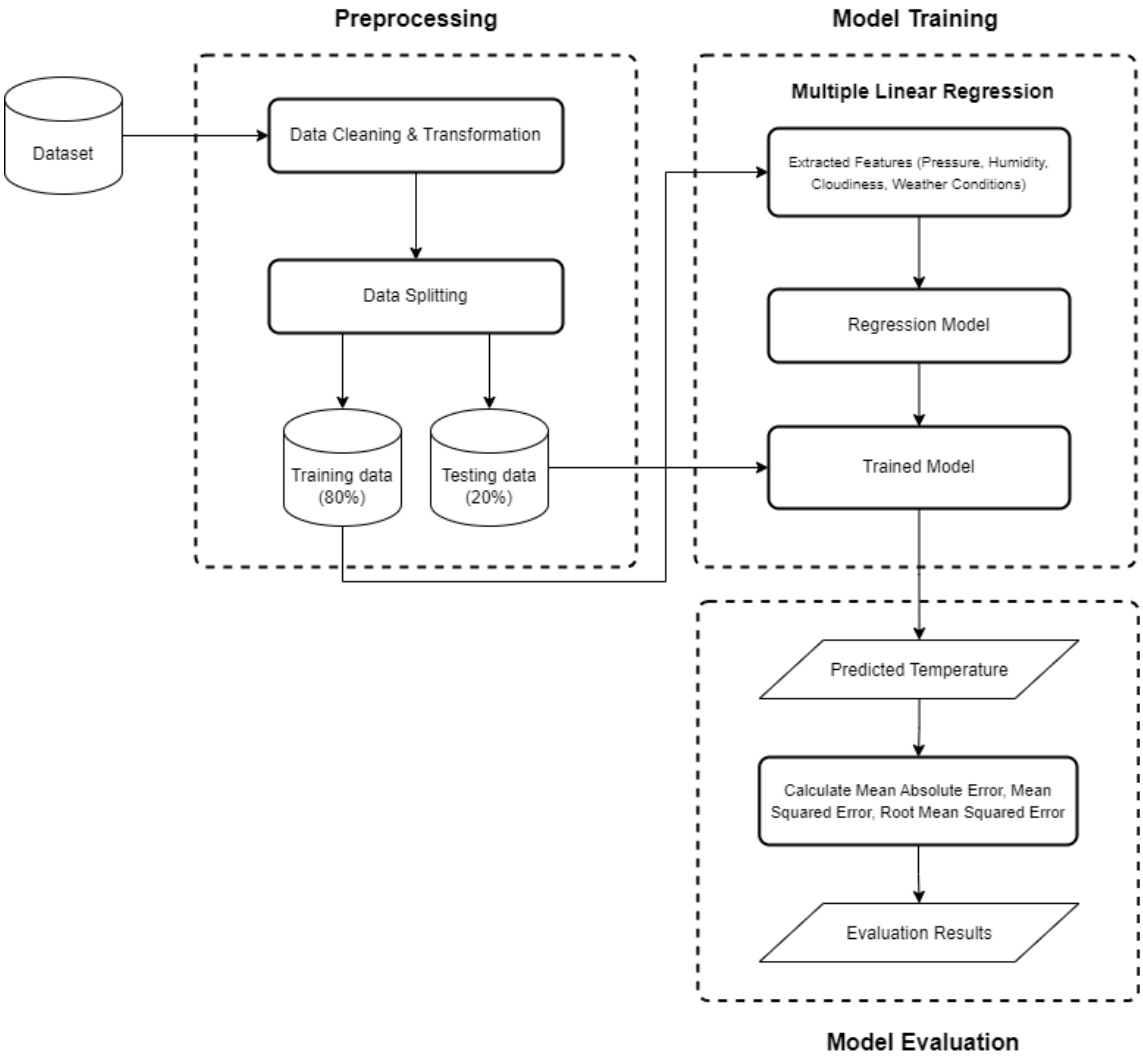
Figure 1. Regression Learning Algorithm



The figure above describes the iterative process of optimizing a loss function done by a regression algorithm. First, it sets initial values or parameters for the linear regression model. These parameters are used to compute predicted values which are then compared to the actual values observed in the dataset, the resulting difference is also known as that function's residuals. The loss function—Mean Squared Error (MSE) afterwards is calculated. In the case that the MSE is not yet at its most minimal or most optimal value the model iterates over various parameters until the MSE is minimized, after which the model is finished learning from the dataset and the algorithm ends.

System Architecture

Figure 2. System Architecture



Preprocessing

A total of 138 cities are available in the dataset but for the purpose of our project, only the sub-region of CAMANAVA in Metro Manila was considered, otherwise known as the cities of Caloocan, Malabon, Navotas, and Valenzuela.

To ensure uniformity and a cleaned dataset, various preprocessing steps were taken. First, to maintain uniformity in the dataset, the data was ensured to be devoid of null values by leveraging the Python library numpy. Second, to fulfill only considering the sub-regions of CAMANAVA in the dataset, the rows in the dataset were filtered according to the city name field. Third, to optimize data access and analysis, the datetime field was used as the index. Lastly, to make use of the available categorical data such as weather status, they were subsequently converted into numerical data using One-Hot encoding.

Model Training

In our study, we utilized a Multiple Linear Regression (MLR) model to predict temperature based on various meteorological factors. The dataset, consisting of 5539 data points, was obtained from the Kaggle Dataset. We preprocessed the dataset by separating the features, excluding temperature, into camanava_x (9 Independent Variables), and the temperature column into camanava_y (1 Dependent Variable).

To ensure a robust evaluation, we employed temporal splitting, allocating 80% (4431) of the data to training (X_train, y_train) and the remaining 20% (1108) for testing (X_test, y_test). This temporal splitting technique ensures that the model is trained on historical data and evaluated on unseen future data, mimicking real-world scenarios where the model is deployed for forecasting.

We initialized the MLR model using the scikit-learn library and trained it on the training data. The model was then utilized to make predictions on the testing

set (y_pred). To visualize the model's performance, we plotted the actual temperature values against the predicted values for the testing period. Additionally, we calculated various evaluation metrics to assess the model's accuracy.

Model Evaluation

In evaluating the performance of our Multiple Linear Regression (MLR) model for temperature prediction, we rely on three fundamental metrics:

Mean Absolute Error (MAE):

MAE measures the average absolute difference between the predicted and actual temperature values. It provides insights into the magnitude of errors in the predictions.

Mean Squared Error (MSE):

MSE measures the average squared difference between the predicted and actual temperature values. A lower MSE indicates better model performance.

Root Mean Squared Error (RMSE):

RMSE is the square root of the MSE and provides a measure of the model's performance in the original scale of the data. Like MSE, lower RMSE values signify better predictive accuracy.

Our model yielded the following results:

Mean Absolute Error (MAE): 1.50

Mean Squared Error (MSE): 3.70

Root Mean Squared Error (RMSE): 1.92

These results give us valuable insights into how well our model predicts temperature. The MAE shows us that, on average, our predictions are off by about 1.50 °C. The MSE tells us that, on average, our predictions miss the actual temperatures by about 3.70 °C, with larger errors getting more emphasis. Finally, the RMSE gives us a sense of the typical error in our predictions, which is around 1.92 °C in terms of our original temperature scale.

In simpler terms, these metrics help us understand how close our model's predictions are to the real temperatures. While they each focus on slightly different aspects, together, they provide a comprehensive picture of our model's performance, guiding us in making informed decisions for various applications, like weather forecasting.

REFERENCES

Anusha, N., Sai Chaithanya, M., & Jithendranath Reddy, G. (2019). Weather Prediction Using Multi Linear Regression Algorithm. IOP Conference Series: Materials Science and Engineering, 590, 012034. doi:10.1088/1757-899x/590/1/012034

BwandoWando. (2023). Philippine Major Cities Weather Data [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DS/3990689

Cifuentes, J., Marulanda, G., Bello, A., & Reneses, J. (2020). Air temperature forecasting using Machine Learning Techniques: A Review. Energies, 13(16), 4215. https://doi.org/10.3390/en13164215

Fang, T., & Lahdelma, R. (2016). Evaluation of a multiple linear regression model and SARIMA model in forecasting heat demand for district heating system. Applied Energy, 179, 544–552. doi:10.1016/j.apenergy.2016.06.133

Guermoui, M., Abdelaziz, R., Gairaa, K., Djemoui, L., & Benkaciali, S. (2020). New temperature-based predicting model for global solar radiation using support vector regression. International Journal of Ambient Energy, 43(1), 1397–1407. https://doi.org/10.1080/01430750.2019.1708792

Gupta, I., Mittal, H., Rikhari, D., & Singh, A. K. (2022). MLRM: A Multiple Linear Regression based Model for Average Temperature Prediction of A Day. arXiv.org. https://arxiv.org/abs/2203.05835

Holmstrom, M., Liu, D., & Vo, C. (2016). Machine learning applied to weather forecasting. Stanford University. https://cs229.stanford.edu/proj2016/report/HolmstromLiuVo-MachineLearningAppliedToWeatherForecasting-report.pdf

Hsu, C. Y., Ng, U. C., Chen, C. Y., Chen, Y. C., Chen, M. J., Chen, N. T., ... & Wu, C. D. (2020). New land use regression model to estimate atmospheric temperature and heat island intensity in Taiwan. Theoretical and Applied Climatology, 141, 1451-1459.

Ismaila, A.-R. B., Muhammed, I., & Adamu, B. (2022). Modelling land surface temperature in urban areas using spatial regression models. *Urban Climate*, *44*, 101213. https://doi.org/10.1016/j.uclim.2022.101213

Karna, N., Roy, P. C., & Shakya, S. (2021). Long Term Temperature Forecasting using Regression Model. International Journal of Advanced Engineering, 4(2). https://ictaes.org/wp-content/uploads/2021/09/IJAE-Vol.04-No.02/IJAE_V4_No2_9.pdf

Menon, S. P., Bharadwaj, R., Shetty, P., Sanu, P., & Nagendra, S. (2017). Prediction of temperature using linear regression. 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT). doi:10.1109/iceeccot.2017.8284588

Nazeri-Tahroudi, M., & Ramezani, Y. (2020). Estimation of dew point temperature in different climates of Iran using support vector regression. *IDŐJÁRÁS / Quarterly Journal of the Hungarian Meteorological Service, 124*(4), 521-539.

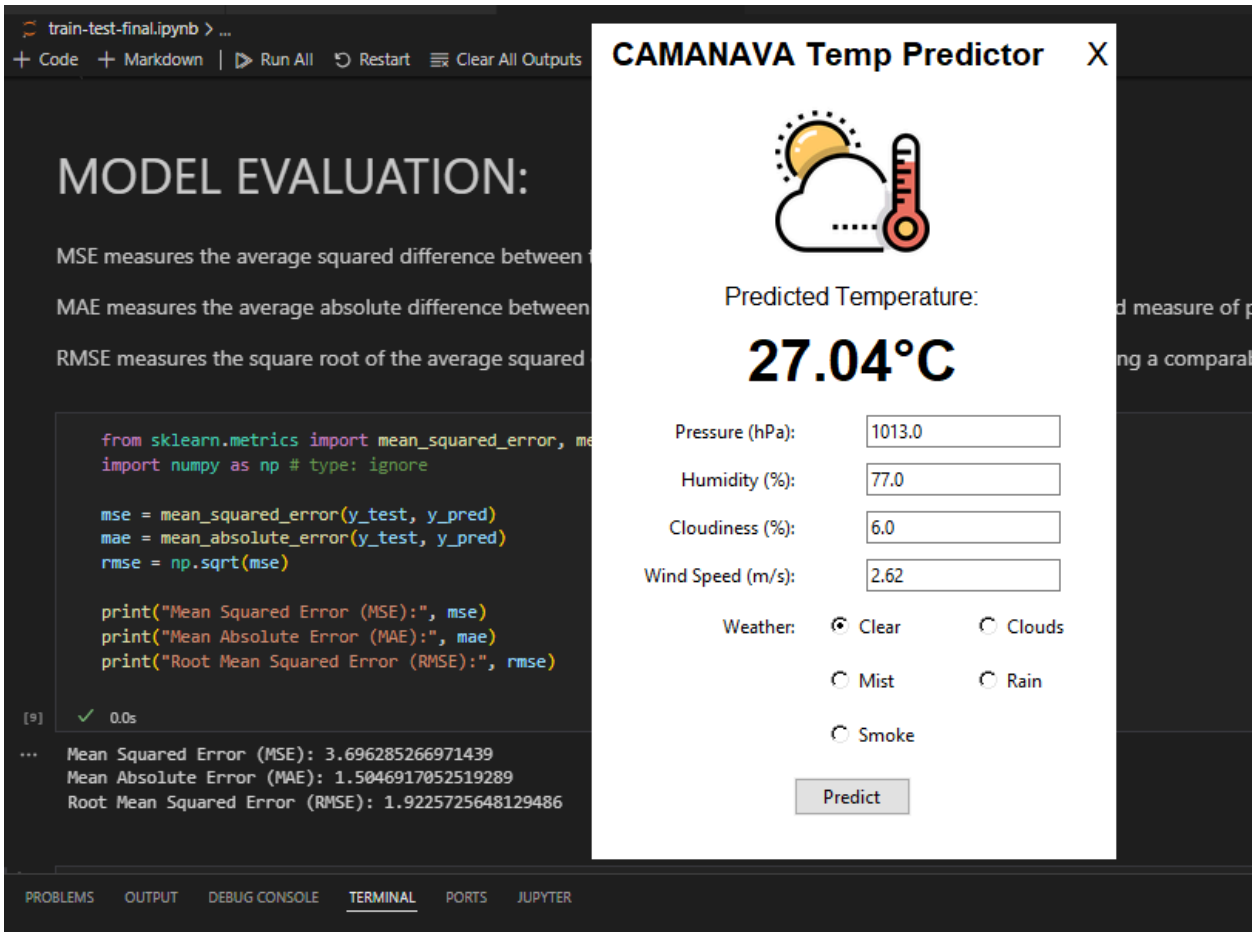
Paras, S.M. (2016). A Simple Weather Forecasting Model Using Mathematical Regression. Indian Research Journal of Extension Education, 12, 161-168.

Thanh Trieu, N., Pottier, B., Rodin, V., & Xuan Huynh, H. (2021). Interpretable Machine Learning for Meteorological Data. 2021 The 5th International Conference on Machine Learning and Soft Computing. doi:10.1145/3453800.3453803

APPENDICES

Appendix 1. Screenshot of the System

Figure 3. The CAMANAVA Temperature Predictor System



Github Repository:

https://github.com/Matthew-Gallardo/Temperature-Forecast-Prediction-in-CAMANAVA-using-Regression

Appendix 2. Data Visualizations

Figure 3. Matrix: Comparison of Actual and Predicted Temperature Values

Actual vs Predicted:

datetime	Actual	Predicted	Difference
2024-04-16 03:06:53	27.87	27.036809	0.833191
2024-04-16 03:08:25	27.00	27.779929	-0.779929
2024-04-16 03:08:47	27.87	26.989466	0.880534
2024-04-16 03:10:13	27.84	27.074057	0.765943
2024-04-16 06:06:56	26.93	26.993479	-0.063479
...	...	...	...
2024-05-31 18:10:02	32.04	27.928364	4.111636
2024-05-31 21:06:54	30.73	27.196296	3.533704
2024-05-31 21:08:22	28.21	25.824120	2.385880
2024-05-31 21:08:43	30.73	27.012250	3.717750
2024-05-31 21:10:07	30.69	26.828205	3.861795

Figure 4. Graph: Comparison of Actual and Predicted Temperature Values