Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi
International Journal of Engineering Research and Development

UMAGD, (2023) 15(1), 312-322.

10.29137/umagd.1232020

Cilt/Volume:15 Sayı/Issue:1 Ocak/January 2023

Araştırma Makalesi / Research Article

Comparison of Regression Algorithms to Predict Average Air Temperature

Berke Ogulcan Parlak[1], Huseyin Ayhan Yavasoglu[1]

[1]Department of Mechatronics Engineering, Yildiz Technical University, 34349 Istanbul, TURKEY

***Başvuru/Received**: 02/01/2023* ***Kabul / Accepted**: 30/01/2023* ***Çevrimiçi Basım / Published Online:** 31/01/2023*

***Son Versiyon/Final Version**: 31/01/2023*

Abstract

Regression algorithms are statistical techniques used to predict the value of a dependent variable, based on one or more independent variables. These algorithms are commonly used in fields such as economics, finance, and engineering. Temperature prediction is a specific application of regression analysis. In this case, the dependent variable is temperature and the independent variables include factors such as humidity, speed of the wind, direction of the wind, and precipitation. There are many different types of regression algorithms, each with its strengths and weaknesses. The study compares the performance of multiple regression models in predicting the average air temperature, using one month's weather data for the Besiktas district of Istanbul. A total of six different regression models, including ridge, lasso, linear, polynomial, random forest (RF), and support vector (SV) regressions, were included in the study. Among the regression models trained and tested on two different data sets, the three most successful models in predicting average air temperature were lasso, RF, and polynomial regressions (PRs), respectively.

Key Words

"Air temperature forecast, linear, nonlinear, regression"

***Corresponding Author:** boparlak@yildiz.edu.tr*

1. Introduction

Temperature is one of the most important weather parameters (Abdel-Aal, 2004), and the prediction of air temperature is a complex process that is a challenging task for researchers (Riordan & Hansen, 2002). In recent years, there has been an increasing interest in the development of regression algorithms for temperature prediction. A variety of approaches have been proposed, including linear and nonlinear methods. In general, linear methods are more efficient and easier to implement (Bastien et al., 2005), but nonlinear methods may be more accurate. Temperature prediction is a difficult problem due to the complex nature of the atmosphere. However, significant progress has been made in recent years thanks to advances in computing power and data availability. Regression algorithms have become very popular (Benyahya et al., 2007) in the literature as they are a promising approach for further improvement in temperature prediction accuracy.

Alaruri and Amer (Alaruri & Amer, 1993) performed least-squares linear regression (LR) analysis on the weather data recorded in Kuwait in 1985, 1986, and 1987 and determined the inputs that should be selected based on the performances. Houthuys et al. (Houthuys et al., 2017) developed a novel model called multi-view least squares SV machines regression. The developed model was tested on a weather dataset. The model performed well on test data with a minimum of 1.14 mean absolute error (MAE). Karna et al. (Karna et al., 2018) trained a simple LR model for a similar weather dataset. The model achieved a minimum of 1.78 MAE on test data. Duan et al. (Duan et al., 2019) trained SV regression models with different kernel parameters for air temperature prediction. The radial basis function (RBF) core showed the best performance on test data. Holmström et al. (Holmstrom et al., 2016) compared the performance of linear and functional regression models in predicting weather temperature. The linear model was clearly more successful. Shafin (Shafin, 2019) trained three different regression models, linear, polynomial, and SV, to predict mean air temperature. PR and SV regression models on the test data showed close success and the results were better than the LR model. Chevalier (Chevalier, 2008) compared the performance of artificial neural networks (ANNs) and SV regression on a weather dataset. SV regression performed better than ANN in predicting air temperature. Jakaria et al. (Jakaria et al., 2020) trained three different regression models, ridge, RF, and SV, to predict air temperature. The results showed that the RF regression model performed best in predicting air temperature. Zhang et al. (Zhang et al., 2021) compared the air temperature prediction performance of the RF regression model with numerical-based weather predicting methods. The RF regression model made observably superior predictions. Vicente-Serrano et al. (Vicente-Serrano et al., 2003) compared the success of linear-nonlinear regression models in predicting air temperature with the success of numerical-based weather predicting methods. The results showed that the regression models produced more accurate predictions.

Although regression algorithms are widely used in air temperature predicting, few studies train multiple regression algorithms on actual weather datasets and compare their performance. This study will therefore be useful for filling this gap in the literature. The main contributions of this manuscript are as follows:

- First, the paper compares the performance of linear, polynomial, lasso, ridge, SV (different kernels), and RF regressions in air temperature prediction.
- Second, a case study for Besiktas, Istanbul is conducted with the dataset that includes 1-month weather data provided by Istanbul Metropolitan Municipality.
- Third, a superior air temperature prediction model is discussed.

During the case study inputs are effectively assigned by the backward elimination method. The data that is thought to be corrupt is filtered out from the data set. Since there is no certainty that the data is corrupt, models are trained and tested on two data sets. The average of the R^2 scores obtained in the two data sets is taken and the relevant regression models are ranked according to their performance.

This paper is organized as follows. Section 2 describes the materials and methods used in the study, including the dataset used and a discussion of the various regression algorithms used. Section 3 presents the results of the analysis for both raw and processed data. Finally, the conclusions are presented in Section 4.

Terminology

a	*Coefficient of the polynomial function*
b	*Model coefficient*
d	*Degree of the polynomial function*
n	*Number of samples*
p	*Number of features*
w	*Normal vector to the surface*
x	*Independent variable*
y	*Dependent variable*
ϵ	*Maximum error*
λ	*Adjustable penalty*

2. Materials and Methods

2.1. Data set

The data set used in the study was obtained from the open data portal of Istanbul Metropolitan Municipality. The data source contains minimum temperature, maximum temperature, average temperature, minimum humidity, maximum humidity, average humidity, minimum wind speed, maximum wind speed, average wind speed, minimum wind direction, maximum wind direction, average wind direction, minimum precipitation, maximum precipitation, average precipitation, minimum road temperature, maximum road temperature, average road temperature, minimum felt temperature, maximum felt temperature and average felt temperature data collected in April 2021 from sensors located in various locations in Istanbul. These data were filtered to be Besiktas district based on location, and only average values based on data. The inputs and outputs of the final model are given in Figure 1.

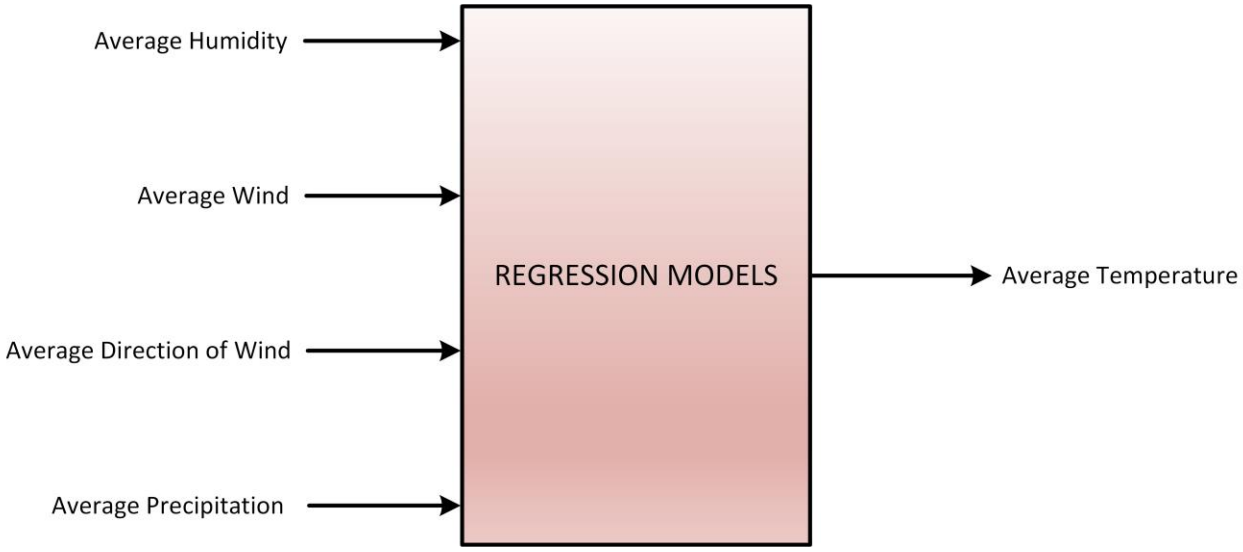

Figure 1. Inputs and output of regression models to be trained

Whether the determined inputs were assigned correctly or not was questioned by the backward elimination method. The method begins by fitting a full model with all possible independent variables, and then iteratively removes the least significant independent variables one by one until the remaining independent variables are all significant. The significance level for removal is chosen by the user (selected as 0.05) and serves as a threshold for determining statistical significance. Some statistical data of the backward elimination method are presented in Table 1. Since the P-value for each feature is below the significance level, it can be said that the inputs are assigned correctly.

Table 1. Statistical outputs of the backward elimination method

Variable	Coefficient	Standard Error	t	P > | t |	[0.025	0.975]
Humidity	0.0587	0.005	12.467	0.000	0.049	0.068
Wind	0.8521	0.147	5.793	0.000	0.563	1.141
Direction of wind	0.0419	0.002	20.868	0.000	0.038	0.046
Precipitation	-2.6179	0.519	-5.040	0.000	-3.638	-1.598

The station collected hourly data from April 1 to April 30. The visualization of the collected average temperature data is presented in Figure 2.

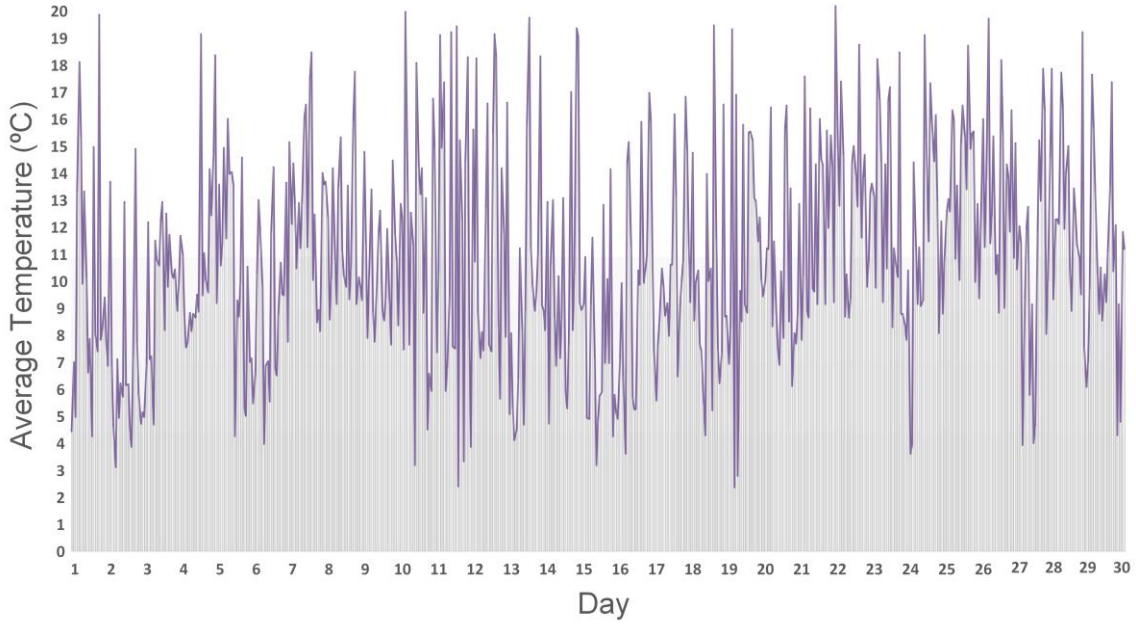

Figure 2. Average temperature data collected from April 1 to April 30

It is important to define the final model parameters given in Figure 1 to determine the regression model to be established and to train the model efficiently. Here, the average relative humidity (RH) is the average humidity (%RH) measured from the related sensor, the average wind is the average wind speed measured from the related sensor (km/h), the average direction of the wind is the average wind direction measured from the related sensor (km/h), the average precipitation is the average precipitation amount measured from the related sensor (kg/m²) and the average temperature are defined as the average temperature measured from the relevant sensor (ºC). The relationship between the variables can be visualized in Figure 3.

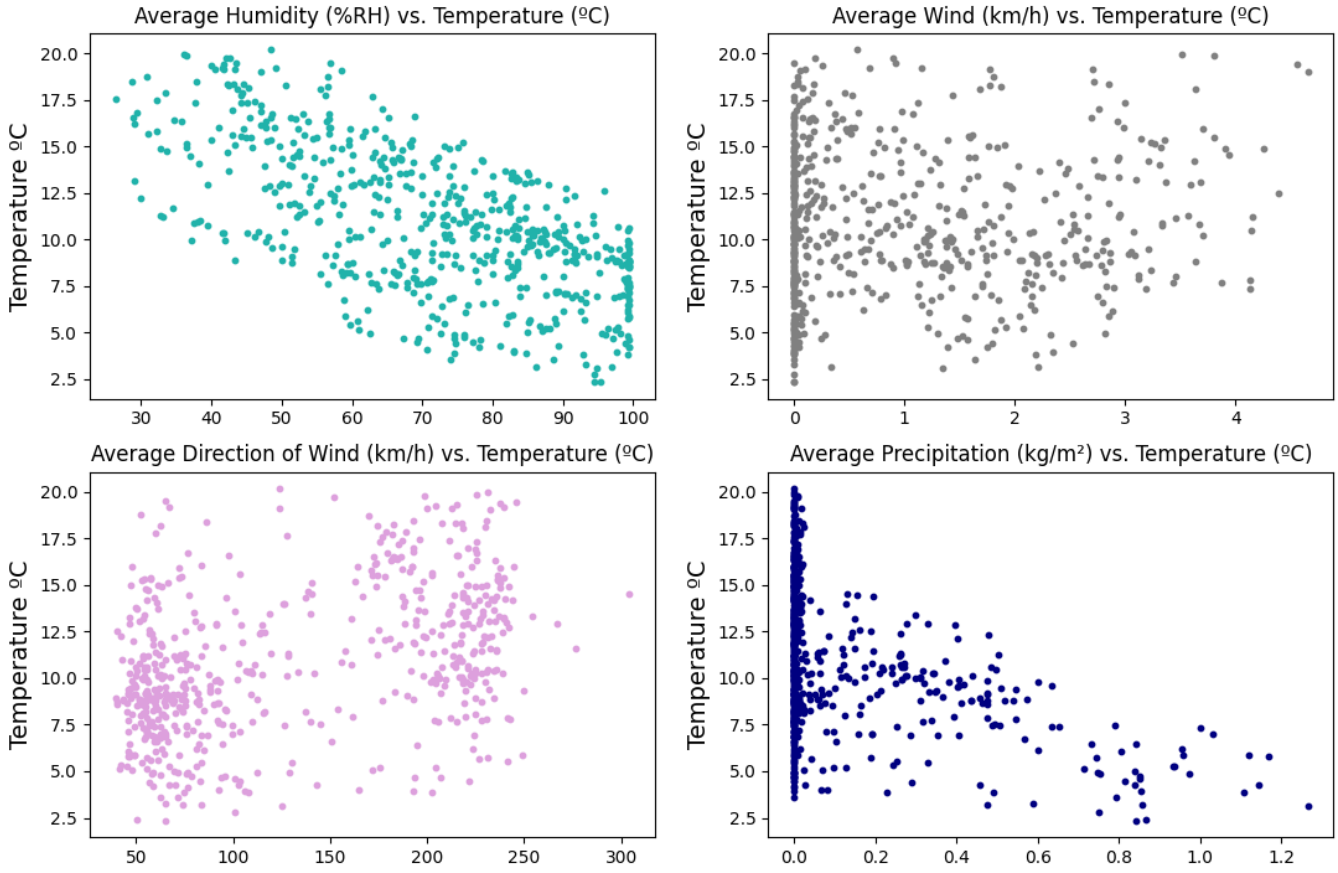

Figure 3. Visualization of relationships between inputs and output

Looking at the data set, it is seen that some values are stacked to zero in the input features other than the average humidity. Although it is not certain, values that are exactly zero can be considered as the sensor not working properly at that time. For this very reason, both the data set visualized in Figure 3 and the data set from which the sensor data, which is thought to be corrupt, is deleted, will be used in the study. This data set is shown in Figure 4. In this data set, each feature was evaluated separately and rows that were exactly equal to zero were deleted. The average precipitation values still seem to be stacked to zero, but this is because the values are very close to zero. Since these values are not exactly zero, they were not removed from the data set.

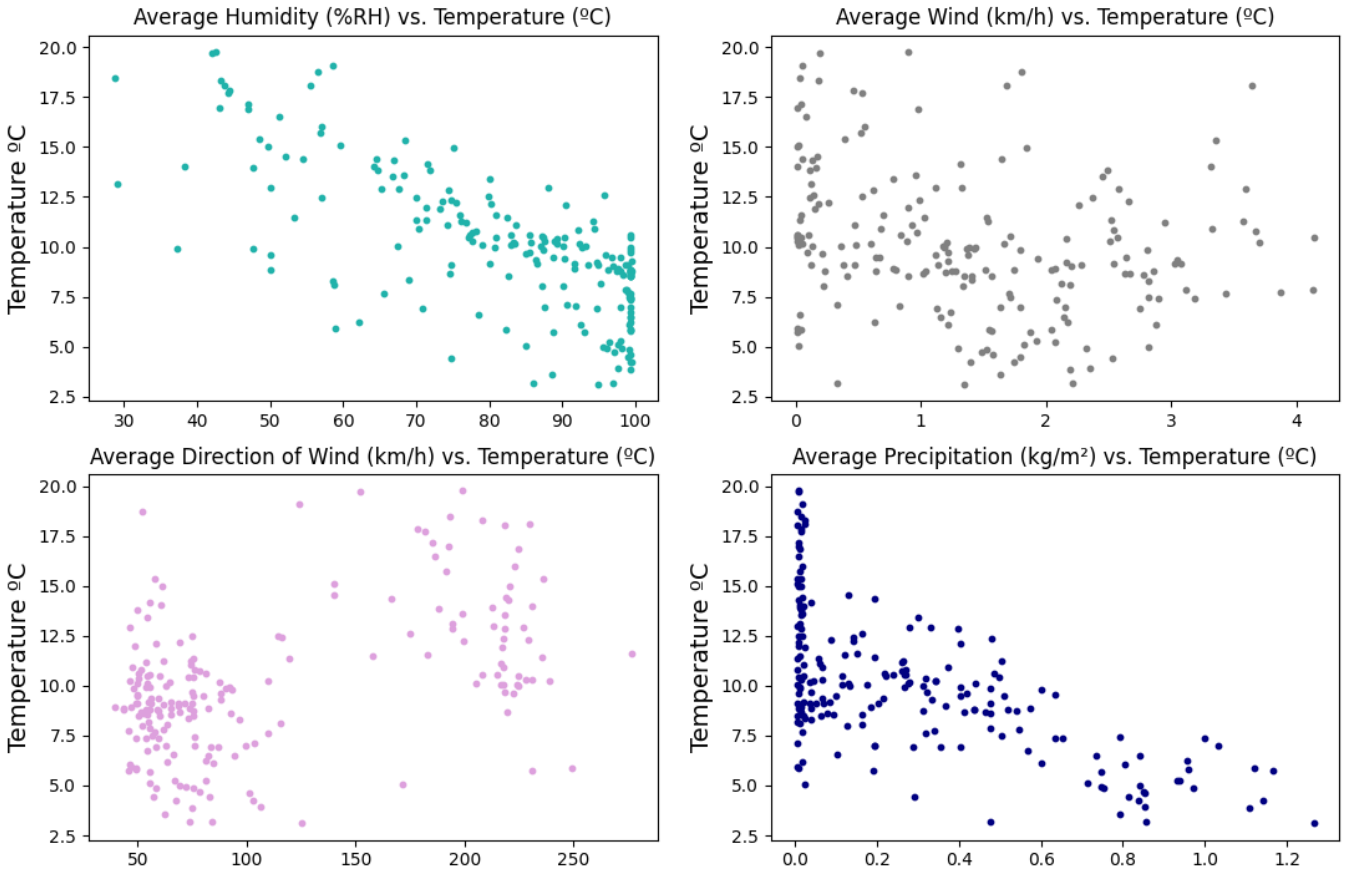

Figure 4. Visualization of relationships between inputs and outputs (exact zero values deleted)

2.2. Regression algorithms

Many regression algorithms can be used for air temperature prediction, including linear (Avdakovic et al., 2013), ridge (Lan & Zhan, 2017), lasso (Al-Obeidat et al., 2020), SV (Paniagua-Tineo et al., 2011), RF (He et al., 2022), and PR (Bahrami & Mahmoudi, 2022). LR is a simple and widely used method that models the linear relationship between the independent variables and the dependent variable. Ridge regression and lasso regression are both variations of LR that introduce regularization terms in the objective function to prevent overfitting and improve model generalization. SV regression is a non-linear method that uses the idea of SV to find the hyperplane that maximally separates the data points in the predictor space and then uses this hyperplane to make predictions. RF is an ensemble method that combines the predictions of multiple decision trees trained on different subsets of the data, resulting in improved predictive accuracy and reduced overfitting. PR is a method that models the relationship between the independent variables and the dependent variable using a polynomial function of a specified degree. These algorithms can be trained on historical temperature data and used to make predictions of future temperature values. The study includes linear, polynomial, lasso, ridge, SV, and RF regression algorithms for air temperature prediction and compares these methods according to their performance.

2.2.1. Linear regression

LR is a method for modeling the linear relationship between a dependent variable and one or more independent variables. It is based on the idea of finding the line (or hyperplane in the case of multiple independent variables) that best fits the data. To fit a LR model, parameters must be found that minimize the sum of the squared errors between the predicted values and the actual values. The objective function for LR is shown in Equation 1, where $\tilde{y}$ is the predicted output value and $\bar{y}$ is the actual output value.

$$\sum_{i=1}^{n} (\bar{y_i} - \tilde{y_i})^2 \quad (1)$$

2.2.2. Polynomial regression

PR is a method for modeling the relationship between the predictor variables and the response variable by fitting a polynomial function of a specified degree. It is based on the idea of extending LR by adding extra features that are powers of the original features. The equation of a PR model with one independent variable is given in Equation 2.

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_d x^d \tag{2}$$

The objective function for PR is similar to the objective function defined for LR. Only the polynomial function given in Equation 2 is used instead of the simple linear function.

2.2.3. Lasso regression

Lasso regression is a variation of LR that introduces a regularization term in the objective function to prevent overfitting and improve model generalization. It is based on the idea of finding the line or hyperplane that best fits the data while also limiting the complexity of the model by reducing the magnitude of the coefficients. To fit a lasso regression model, an optimization algorithm must be used that can process the absolute value of the coefficients in the term regularization. A commonly used method is the coordinate descent method, which iteratively optimizes each coefficient while keeping the others constant. The objective function for lasso regression is shown in Equation 3.

$$\sum_{i=1}^{n} \left(y_i - b_0 - \sum_{i=1}^{p} b_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |b_j| \tag{3}$$

2.2.4. Ridge regression

Ridge regression and lasso are both techniques for regularizing LR models. Regularization is a way to prevent overfitting by adding a penalty term to the objective function. In lasso regression, the penalty term is the sum of the absolute values of the coefficients, while in ridge regression it is the sum of the squares of the coefficients. This has the effect of shrinking the coefficients towards zero and can help reduce overfitting. The objective function for ridge regression is shown in Equation 4.

$$\sum_{i=1}^{n} \left(y_i - b_0 - \sum_{i=1}^{p} b_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} b_j^2 \tag{4}$$

2.2.5. Support vector regression

SV regression is based on the idea of finding the hyperplane in the high-dimensional feature space that maximally separates the data points and then using this hyperplane to make predictions. To fit a SV regression model, the optimization problem given in Equation 5 must be solved.

$$\text{MIN} \frac{1}{2} \| \mathbf{w} \|^2 \text{ with the constraint of } |y_i - b_i x_i| \leq \varepsilon. \tag{5}$$

2.2.6. Random forest regression

RF regression is an ensemble method that combines the predictions of multiple decision tree regressors to make predictions. A decision tree regressor is a tree-like model that makes predictions by recursively partitioning the data based on the feature values. The prediction for a given sample is computed as the mean response value of the samples in the leaf node to which it belongs. The RF architecture for classification and regression analysis is given in Figure 5.

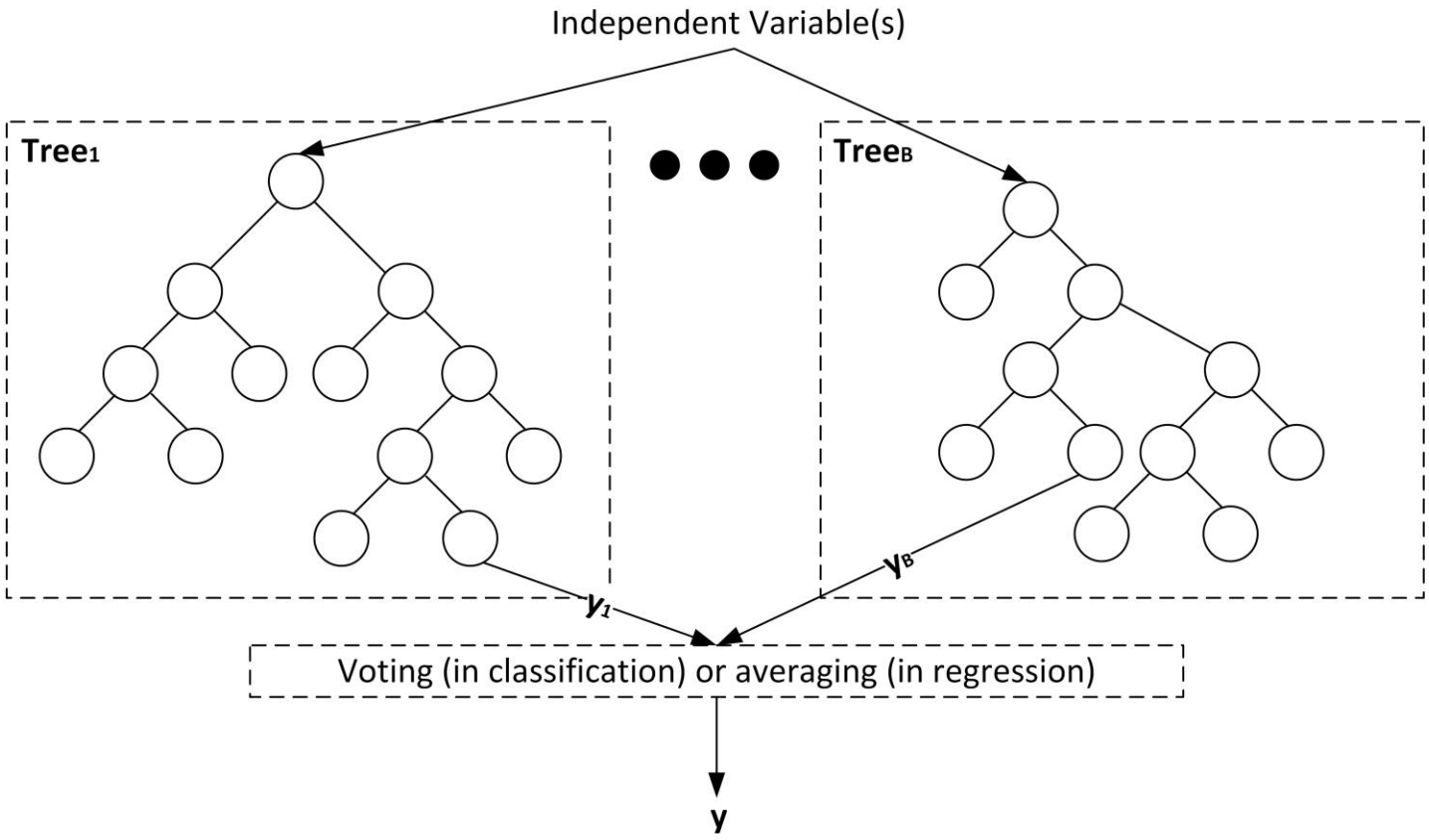

Figure 5. Architecture of the RF model (Verikas et al., 2016)

2.3. Implementation of regression algorithms

In the study, Python programming language was preferred for regression analysis due to the abundance of libraries and resources specially designed for data analysis (Stančin & Jović, 2019) and statistical modeling (Seabold & Perktold, 2010). Python has many libraries such as NumPy, Pandas, and Scikit-learn that make it easy to do regression analysis (Massaron & Boschetti, 2016). Pandas library was used to collect weather data, Matplotlib library for visualization of raw and processed data, Statsmodel library for accurate assignment of model inputs (data preparation), and Scikit-learn library for model development (training, evaluation, tuning) in the study. The development steps for the air temperature prediction regression model in Python and which libraries are used in which steps are shown in Figure 6.

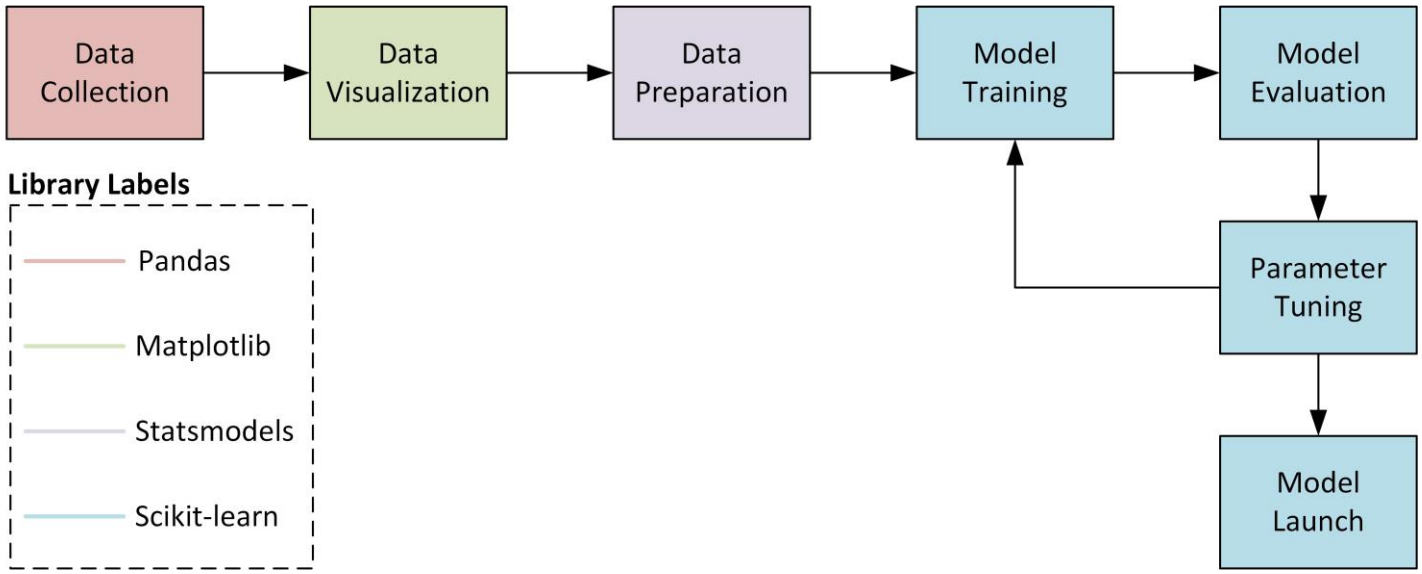

Figure 6. Development steps for the air temperature prediction regression model

3. Results

3.1. Results of raw data

The section presents the performances of the regression algorithms trained with the raw data set in Figure 3. 70% of the data is reserved for training purposes and 30% for testing purposes. The adjustable parameters of the linear, polynomial, lasso, ridge, SV, and RF regression algorithms are optimized to give the best results. The R^2 scores obtained as a result of applying the relevant regression algorithms to the test data are shown in Figure 7.

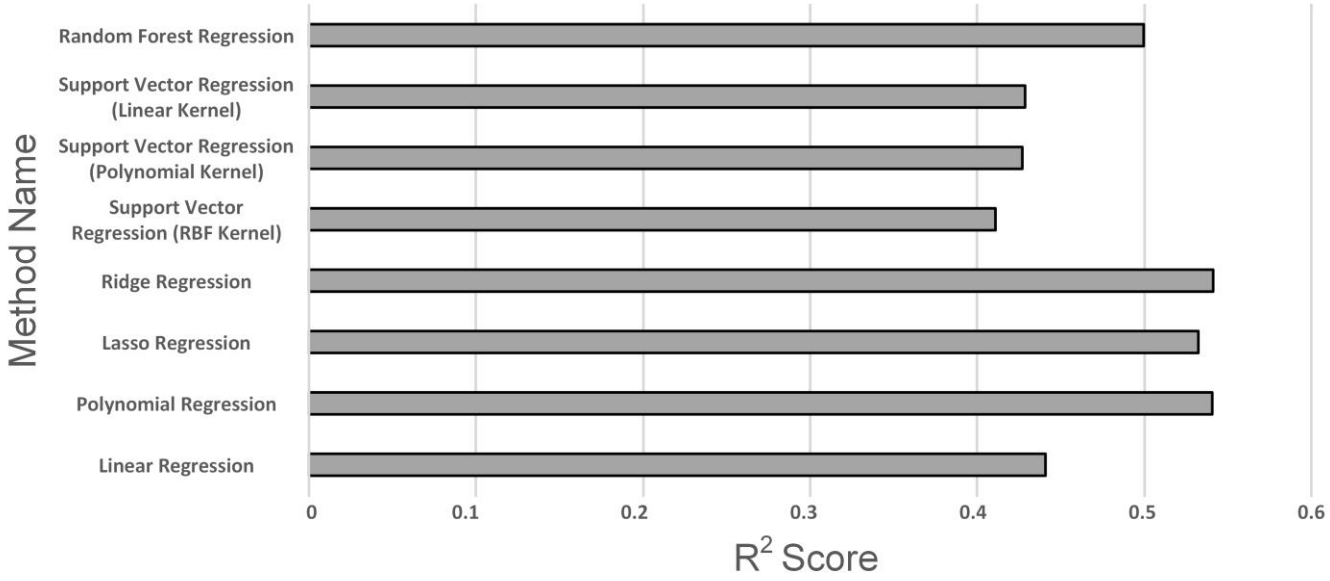

Figure 7. R^2 scores in air temperature prediction of regression models (raw data)

3.2. Results of processed data

The section presents the performances of the regression algorithms trained with the processed data set in Figure 4. 70% of the data is reserved for training purposes and 30% for testing purposes. The adjustable parameters of the linear, polynomial, lasso, ridge, SV, and RF regression algorithms are optimized to give the best results. The R^2 scores obtained as a result of applying the relevant regression algorithms to the test data are shown in Figure 8.

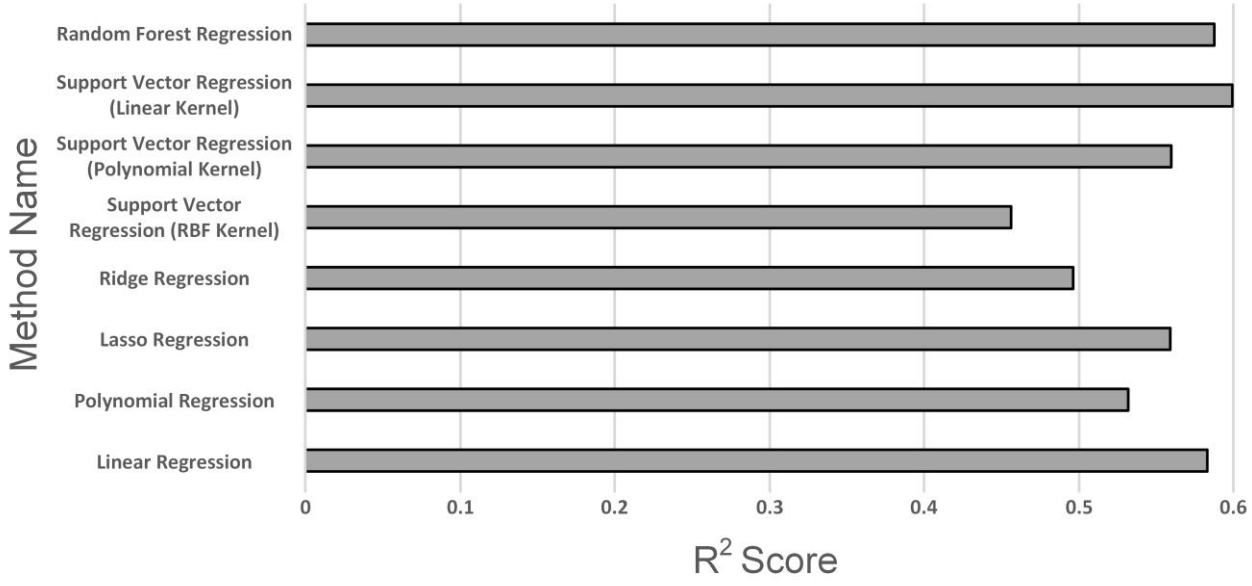

Figure 8. R^2 scores in air temperature prediction of regression models (processed data)

3.3. Combined results

Since there are two different types of data (raw and processed) in the study, it would be reasonable to compare the performance of the regression algorithms by averaging the R^2 scores obtained from these data. An average of the results obtained in Figure 7 and Figure 8 is given in Figure 9. The average scores obtained will be used to compare the performance of the respective regression algorithms in predicting air temperature.

Figure 9. Average R^2 scores in air temperature prediction of regression models (combined data)

4. Conclusions

Temperature prediction is a challenging problem that has received considerable attention in recent years. A wide variety of methods have been proposed for tackling this problem, with varying degrees of success. One approach that has shown promise is the use of regression algorithms. Regression-based models can be used to learn the relationships between temperature and a variety of potential predictive factors, such as historical temperatures, atmospheric conditions, and so on. These models can then be used to make predictions about future temperatures. The development of robust regression algorithms is therefore an important area of research with the potential to improve the ability to predict future temperature changes.

The study compares the performance of multiple regression models in predicting air temperature. For the training and testing of the models, 1-month measurement data of the district of Besiktas, Istanbul, dated April 2021, was used. While the inputs of the regression models were selected as average humidity, average wind, the average direction of the wind, and average precipitation, the output was determined as average temperature. The data set is divided into raw data and processed data. The processed data is created by deleting rows that are exactly equal to zero in the raw data. Among the models trained and tested with raw data, the three most successful models were the ridge (R^2 score of 0.5417), polynomial (R^2 score of 0.5407), and lasso (R^2 score of 0.5329) regression models, respectively. Among the models trained and tested with processed data, the three most successful models were the SV (linear kernel, R^2 score of 0.6000), RF (R^2 score of 0.5875), and LR (R^2 score of 0.5830) models, respectively. Since these two discrete results cannot be used to rank regression models, the R^2 score average of the two results was taken. Accordingly, the three most successful models for temperature prediction were the lasso (R^2 score of 0.5461), RF (R^2 score of 0.5438), and PR (R^2 score of 0.5364) models, respectively.

An analysis of the R^2 score values for the models trained on both processed and raw data revealed that lasso regression is the most effective method for predicting air temperature. However, when the analysis was limited to models trained only on processed data, a different outcome emerged. Linear kernel-structured SV regression performs superiorly in this case. This difference is due to the dependence of regression model performance on the quality and characteristics of the training dataset. Notably, the results presented in this paper could be further enhanced by expanding the dataset or implementing advanced data processing techniques. Nonetheless, the paper serves as a useful guide for researchers and practitioners in selecting suitable regression models for air temperature estimation. The study indicates that, with well-processed data, linear kernel-structured SV regression may be a better choice than lasso regression for air temperature prediction, according to the paper.

References

Abdel-Aal, R. E. (2004). Hourly temperature forecasting using abductive networks. Engineering Applications of Artificial Intelligence, 17(5), 543–556.

Al-Obeidat, F., Spencer, B., & Alfandi, O. (2020). Consistently accurate forecasts of temperature within buildings from sensor data using ridge and lasso regression. Future Generation Computer Systems, 110, 382–392.

Alaruri, S. D., & Amer, M. F. (1993). Empirical regression models for weather data measured in Kuwait during the years 1985, 1986, and 1987. Solar Energy, 50(3), 229–233.

Avdakovic, S., Ademovic, A., & Nuhanovic, A. (2013). Correlation between air temperature and electricitydemand by linear regression and wavelet coherence approach: UK, Slovakia and Bosnia and Herzegovina case study. Archives of Electrical Engineering, 62(4).

Bahrami, M., & Mahmoudi, M. R. (2022). Long-term temporal trend analysis of climatic parameters using polynomial regression analysis over the Fasa Plain, southern Iran. Meteorology and Atmospheric Physics, 134(2), 1–12.

Bastien, P., Vinzi, V. E., & Tenenhaus, M. (2005). PLS generalised linear regression. Computational Statistics & Data Analysis, 48(1), 17–46.

Benyahya, L., Caissie, D., St-Hilaire, A., Ouarda, T. B. M. J., & Bobée, B. (2007). A review of statistical water temperature models. Canadian Water Resources Journal, 32(3), 179–192.

Chevalier, R. F. (2008). Air temperature prediction using support vector regression and GENIE: The Georgia Extreme-weather Neural-network Informed Expert. University of Georgia.

Duan, S., Yang, W., Wang, X., Mao, S., & Zhang, Y. (2019). Grain pile temperature forecasting from weather factors: A support vector regression approach. 2019 IEEE/CIC International Conference on Communications in China (ICCC), 255–260.

He, Y., Chen, C., Li, B., & Zhang, Z. (2022). Prediction of near-surface air temperature in glacier regions using ERA5 data and the random forest regression method. Remote Sensing Applications: Society and Environment, 28, 100824.

Holmstrom, M., Liu, D., & Vo, C. (2016). Machine learning applied to weather forecasting. Meteorol. Appl, 10, 1–5.

Houthuys, L., Karevan, Z., & Suykens, J. A. K. (2017). Multi-view LS-SVM regression for black-box temperature prediction in weather forecasting. 2017 International Joint Conference on Neural Networks (IJCNN), 1102–1108.

Jakaria, A. H. M., Hossain, M. M., & Rahman, M. A. (2020). Smart weather forecasting using machine learning: a case study in tennessee. ArXiv Preprint ArXiv:2008.10789.

Karna, N., Roy, P. C., & Shakya, S. (2018). Temperature Prediction using Regression Model.

Lan, Y., & Zhan, Q. (2017). How do urban buildings impact summer air temperature? The effects of building configurations in space and time. Building and Environment, 125, 88–98.

Massaron, L., & Boschetti, A. (2016). Regression analysis with Python. Packt Publishing Ltd.

Paniagua-Tineo, A., Salcedo-Sanz, S., Casanova-Mateo, C., Ortiz-García, E. G., Cony, M. A., & Hernández-Martín, E. (2011). Prediction of daily maximum temperature using a support vector regression algorithm. Renewable Energy, 36(11), 3054–3060.

Riordan, D., & Hansen, B. K. (2002). A fuzzy case-based system for weather prediction. Engineering Intelligent Systems for Electrical Engineering and Communications, 10(3), 139–146.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference, 57(61), 10–25080.

Shafin, A. A. (2019). Machine learning approach to forecast average weather temperature of Bangladesh. Global Journal of Computer Science and Technology, 19(3), 39–48.

Stančin, I., & Jović, A. (2019). An overview and comparison of free Python libraries for data mining and big data analysis. 2019 42nd

International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 977–982.

Verikas, A., Vaiciukynas, E., Gelzinis, A., Parker, J., & Olsson, M. C. (2016). Electromyographic patterns during golf swing: Activation sequence profiling and prediction of shot effectiveness. Sensors, 16(4), 592.

Vicente-Serrano, S. M., Saz-Sánchez, M. A., & Cuadrat, J. M. (2003). Comparative analysis of interpolation methods in the middle Ebro Valley (Spain): application to annual precipitation and temperature. Climate Research, 24(2), 161–180.

Zhang, Q., Cheng, J., & Wang, N. (2021). Fusion of All-Weather Land Surface Temperature From AMSR-E and MODIS Data Using Random Forest Regression. IEEE Geoscience and Remote Sensing Letters, 19, 1–5.