ISSN: 2278-4632
Vol-14, Issue-9, September: 2024

An enhanced K-NN regression model based on Principal Component analysis in big data processing for weather station data

S. Tamilarasan[#1], Dr. S. K. Mahendran[*2], Dr. S. Pandikumar[#3]

[#1] Research Scholar, Department of Computer Science, Centre for Research and Evaluation, Bharathiar University, Coimbatore, TamilNadu, India

[*2]Assistant Professor, P.G. & Research Department of Computer Science Government Arts College, Coimbatore, TamilNadu, India

[#3]Assistant Professor, Department of Computer Applications, Yadava College, Madurai, TamilNadu, India

[#1] tamilarasanonline@gmail.com
*2 sk.mahendran@yahoo.co.in
[#3]pandikmr19986@gmail.com

ABSTRACT:

This study analyses the application of a PCA-based enhanced k-Nearest Neighbors (k-NN) Regression model for big data processing in weather station data. Weather data, characterized by high dimensionality and large volumes due to continuous collection across numerous locations, poses challenges in computational efficiency and noise management. Principal Component Analysis (PCA) is employed to reduce the dimensionality of the data by transforming it into principal components that capture the most significant variance, thus simplifying the dataset while retaining essential information. The enhanced k-NN Regression model is then applied to the reduced dataset to predict weather variables effectively. This approach leverages PCA's ability to enhance computational efficiency and reduce noise, making the E-k-NN algorithm feasible for large-scale data analysis. The integration of PCA reduces the complexity and runtime of E-k-NN, which traditionally suffers from high computational costs in high-dimensional spaces. Experimental results demonstrate that the PCA-based E-k-NN model improves prediction accuracy of 89.67 % and processing speed, making it suitable for near real-time weather forecasting and analysis.

Keywords: Feature Selection, Machine Learning, Big Data, PCA, E-kNN

1. INTRODUCTION

Weather prediction is a complex task involving massive datasets from various sources, including satellite data, sensor networks, and historical records. Machine learning (ML) has emerged as a powerful tool to model and predict weather patterns. However, the presence of redundant and irrelevant features can

ISSN: 2278-4632
Vol-14, Issue-9, September: 2024

significantly hamper model performance, making feature selection crucial for effective ML applications. This paper focuses on feature selection techniques for processing big data in weather monitoring, aiming to optimize model accuracy and reduce computational costs. Renukadevi G et. al (2019) In big data processing, there are many algorithms available to process the huge data. The weather station data is used to find the weather conditions as well as prediction. To process these types of data, the regression model is suitable, and the prediction based on the available data can be easily implemented using this regression model. The data stored in the SQL Server can be retrieved using the authorized access, and process those data using the algorithm developed for the data prediction. The data from weather station is very huge, because it is updated on every 30 seconds [9].

2. RELATED WORK

Suvendra Kumar Jayasingh, Jibendu Kumar Mantri, and Sipali Pradhan (2022) the accuracy evaluation of the models shows that the machine learning models perform better than the traditional models. These models made use of the dataset collected from predefined recourses in which the maximum accuracy is observed up to 81.67% [12]. Hui Yie Teh1, Andreas W. Kempa-Liehr and Kevin I-Kai Wang (2020) There are 16 different types of methods presented for error detection, which are obtained from 32 papers out of the 57 selected papers that introduced techniques for the respective problem. The two most common approaches are principal component analysis (PCA) and artificial neural networks (ANN). They are both used to model the normal sensor behaviour and the newly observed readings will be compared to the model to determine if it is anomalous. Other techniques for fault detection include Ensemble Classifiers, Support Vector Machines, Clustering, and hybrid methods [15].

3. METHODOLOGY

3.1 Data Collection

The DHT22 sensor is commonly used in weather monitoring due to its ability to provide accurate and reliable measurements of temperature and humidity. The DHT22 measures ambient temperature with reasonable accuracy, which is essential for understanding weather conditions, predicting weather patterns, and conducting climate studies. The sensor provides relative humidity readings, which are critical for assessing atmospheric moisture levels. This information is valuable for weather forecasting, agricultural planning, and environmental monitoring.

In interfacing a DHT22 sensor with an ESP8266 and SQL Server

ISSN: 2278-4632
Vol-14, Issue-9, September: 2024

1. DHT22 Sensor: This sensor measures temperature and humidity. It sends data to the ESP8266 microcontroller.

2. ESP8266: The ESP8266 reads data from the DHT22 sensor. This is a Wi-Fi module that reads data from the DHT22 sensor and can process it. It then communicates with a server to send the data. The ESP8266 then sends this data to a web server or a cloud service. The web server or cloud service processes the data and can store it in a database.

3. SQL Server: SQL Server manages the database where ESP8266 data is stored, allowing for querying, analysis, and retrieval. This is a database management system used to store and manage data. In this setup, SQL Server would be used to store the data collected by the ESP8266.By integrating DHT22 sensors into weather stations, continuous data on temperature and humidity can be collected. This data is crucial for

Climate Analysis: Understanding long-term trends and variations in temperature and humidity. Weather Forecasting: Improving the accuracy of short-term and long-term weather forecasts.

The dataset used in this study comprises historical weather data from DHT22 sensor with ESP8266 such as temperature, humidity, wind speed, pressure, and precipitation levels, spanning a period of 10 years.

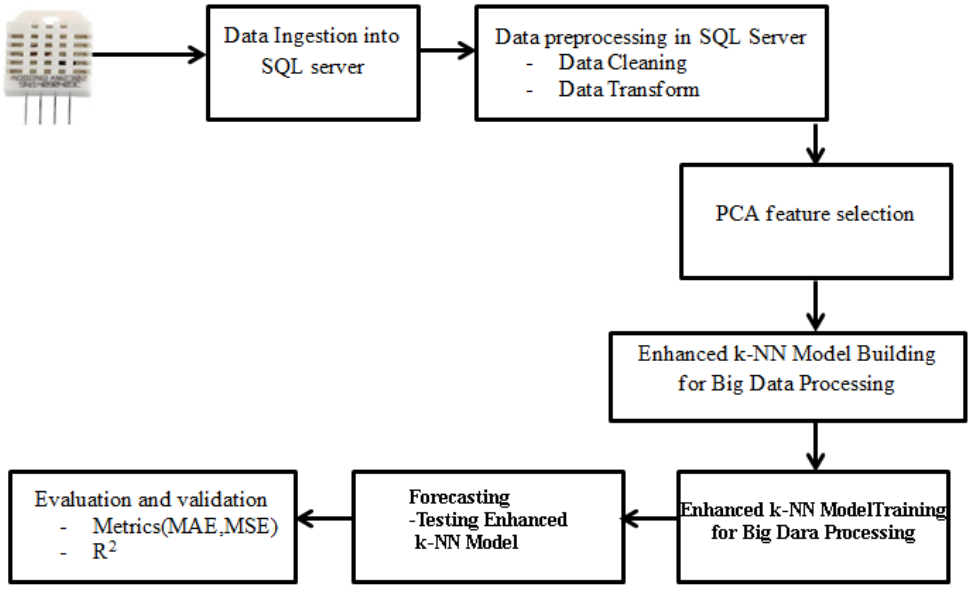

Fig.1. Block diagram of PCA feature selection to forecast weather

ISSN: 2278-4632
Vol-14, Issue-9, September: 2024

3.2 Data Preprocessing:

Data preprocessing involves cleaning, normalizing, and transforming the data to ensure consistency. Missing values are handled using interpolation, and outliers are removed to avoid skewing the results.

3.3 Feature Selection Techniques:

Principal Component Analysis (PCA): PCA feature selection technique is used to reduce data dimensionality by transforming the original variables into a new set of uncorrelated variables (principal components) that capture the most variance. The fundamental benefit of PCA is that the principal components of every dataset are orthogonal to each other, so there is no redundant data after pre-processing

Algorithm of PCA:
Input: Data Matrix
Output: Reduced set of arrangements
Step 1: Create N x d information framework (X) with one row has a data point xn
Step 2: Subtract mean x from each line vector xn.
Step 3: **Σ** *of the yield in step 2 is the Covariance matrix of*
Step-4: Find eigenvectors and Eigen values of Σ.
Step 5: Identify the largest value of Eigenvector in Principle Component
Step-6: Output PCs

3.4 Machine Learning Technique Used:

Enhanced K-nearest neighbors (E-kNN) is a supervised learning method that may be used for both regression and classification, however it is usually utilized for classification. E-kNN aims to predict the right class of testing data given a set with various classes by calculating the distance between both the testing data and all the training points. It then chooses the k points that are the most similar to the test.

```
generate_pred_knn("EKNN",KNN,x_train,x_test,y_train,y_test)
```

```
------Evaluation metrics for training data set--------
modelname-EKNN
rmse is 8.889448004262103
Rsqr is  91.24
-------Evaluation metrics for test dataset--------
modelname-EKNN
rmse is 12.324286094952836
Rsqr is  83.08
```

Fig. 2. Evaluation metrics of Enhanced k-NN

ISSN: 2278-4632
Vol-14, Issue-9, September: 2024

4. RESULTS & DISCUSSIONS

4.1 Performance Measure:

To evaluate the performance of a PCA based k-NN regression model, the following evaluation performance metrics will be calculated.

Mean Absolute Error (MAE): this Measures the average magnitude of the errors in the predictions, without considering their direction. It is calculated as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Squared Error (MSE): this metric measures the average of the squared of the errors. It gives a higher weight to larger score.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE): this is the root of MSE, which gives a measure of the error in the same units as the target variable.

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$$

R-Squared (Coefficient of Determination): this statistic indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. It is calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

The tested environment has following features processor and platform: Intel i3, sixth era processor, OS: Ubuntu 20.04 and RAM 8 GB, python 3.7.6, and Jupyter notebook 6.03.

Name of the Model	Description	Purpose
Enhanced k-NN	Machine learning algorithms with Deterministic Approach	Predicting the weather
MAE, MSE, RMSE & R^2	Mathematical formulation	forecasting weather performance indicator

ISSN: 2278-4632
Vol-14, Issue-9, September: 2024

Table 1: Performance Metrics and its description

Name	Datatype
Humidity	Number
TempC	Number
TmepF	Number
HeatIndexC	Number
HeatIndexF	Number

Table 2: Variables used in Weather forecasting dataset with datatypes

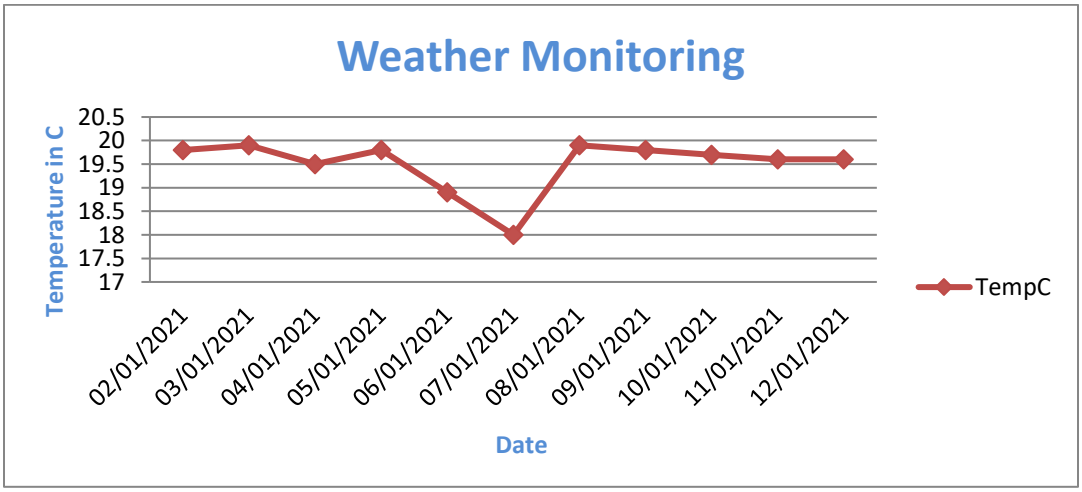

Fig. 3. Sample Temperature Readings during Feb, 2021

Date	TempC	Humidity	Status
02/02/2021	19.8	26.5	Partly Sunny
03/02/2021	19.9	27.5	Partly Sunny
04/02/2021	19.5	27.5	Rainy
05/02/2021	19.8	26.5	Partly Sunny
06/02/2021	18.9	22.8	Cloudy
07/02/2021	18	24	Cloudy
08/02/2021	19.9	27.5	Partly Sunny
09/02/2021	19.8	26	Partly Sunny
10/02/2021	19.7	26.5	Cloudy
11/02/2021	19.6	26.7	Rainy
12/02/2021	19.6	26.8	Rainy
Average	**19.5**	**26.20**	**Partly Sunny**

Table 3.Weather Readings of Feb, 2021 with Status

ISSN: 2278-4632
Vol-14, Issue-9, September: 2024

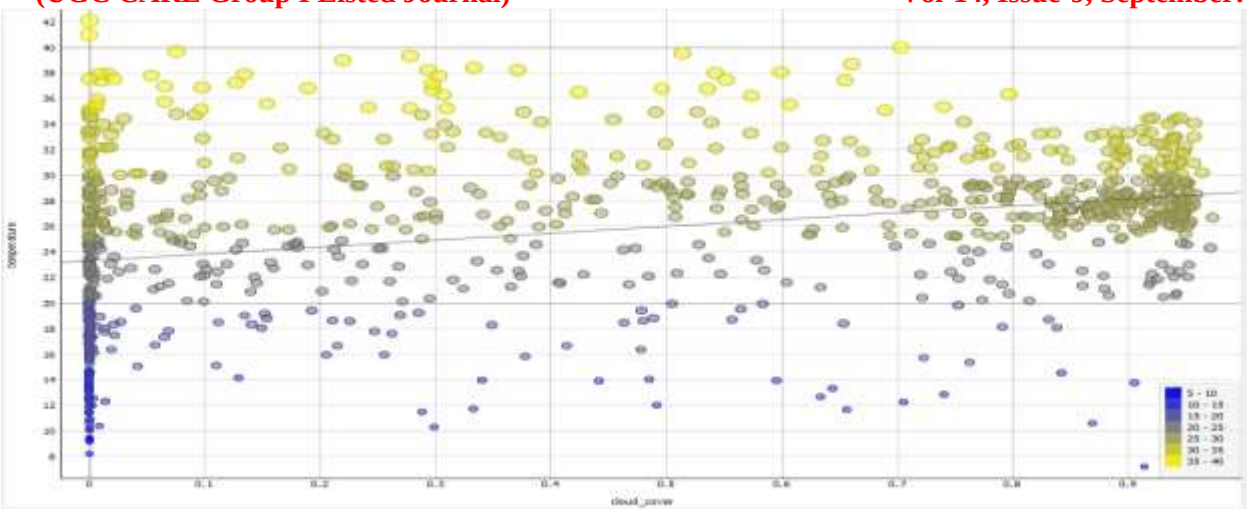

Fig. 4. Scatter plot temperature Vs Humidity

5. CONCLUSION

PCA effectively reduces the complexity of the weather station data by transforming it into a lower-dimensional space. This not only speeds up the enhanced k-NN algorithm but also helps in mitigating the curse of dimensionality, leading to more accurate and interpretable results. By leveraging PCA, the k-NN regression model becomes more computationally efficient. The reduced number of dimensions allows for faster distance computations and model training, which is crucial when dealing with large-scale weather data. The use of PCA helps in filtering out noise and irrelevant features, which can improve the performance of the enhanced k-NN regression model. The model's predictions become more reliable, as PCA focuses on the principal components that capture the most variance in the data. The proposed system gives high accuracy of 89.67% on weather prediction.

ACKNOWLEDGMENT

I would like to thank with overwhelmed gratitude, the enormous support and guidance rendered to us by Dr. S. Venkatakrishnan, Assistant Professor, Department of Computer Science, Annamalai University, TN, India.

REFERENCES

[1] Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., & Ayyash, M. (2015). Internet of Things: A survey on enabling technologies, protocols, and applications. *IEEE Communications Surveys & Tutorials, 17*(4), 2347-2376.

ISSN: 2278-4632
Vol-14, Issue-9, September: 2024

[2] Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2015). Internet of Things (IoT): A vision, architectural elements, and future directions. *Future Generation Computer Systems, 29*(7), 1645-1660.

[3] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems, 47*, 98-115.

[4] Wang, S., Wan, J., Li, D., & Zhang, C. (2016). Implementing smart factory of Industrie 4.0: An outlook. *International Journal of Distributed Sensor Networks, 2016*, 3159805.

[5] Zanella, A., Bui, N., Castellani, A., Vangelista, L., & Zorzi, M. (2016). Internet of Things for smart cities. *IEEE Internet of Things Journal, 1*(1), 22-32.

[6] Hu, H., Wen, Y., Chua, T. S., & Li, X. (2016). Toward scalable systems for big data analytics: A technology tutorial. *IEEE Access, 2*, 652-687.

[7] Farooq, M. U., Waseem, M., Khairi, A., & Mazhar, S. (2017). A critical analysis on the security concerns of Internet of Things (IoT). *International Journal of Computer Applications, 111*(7), 1-6.

[8] Abdel-Basset, M., Manogaran, G., & Gamal, A. (2018). A novel method for evaluating activities of daily living based on Dempster-Shafer theory of evidence for internet of medical things. *IEEE Access, 6*, 63025-63035.

[9] Renukadev Gi et. al (2019), An Implementation of Regression Model in Big data processing for Weather station data, IJRAR March 2019, Volume 6, Issue 1,886-891

[10] Khajeh-Hosseini, A., Greenwood, D., Smith, J. W., & Sommerville, I. (2019). The cloud adoption toolkit: Supporting cloud adoption decisions in the enterprise. *Software: Practice and Experience, 42*(4), 447-465.

[11] Borgia, E. (2019). The Internet of Things vision: Key features, applications, and open issues. *Computer Communications, 54*(1), 1-31.

[12] Suvendra Kumar Jayasingh, Jibendu Kumar Mantri, and Sipali Pradhan (2022), SmartWeather Prediction Using Machine Learning, Springer Nature Singapore Pte Ltd., 571-583

[13] Fadlullah, Z. M., Tang, F., Mao, B., Kato, N., Akashi, O., Inoue, T., & Mizutani, K. (2020). State-of-the-art deep learning: Evolving machine intelligence toward tomorrow's intelligent network traffic control systems. *IEEE Communications Surveys & Tutorials, 19*(4), 2432-2455.

ISSN: 2278-4632
Vol-14, Issue-9, September: 2024

[14] Sun, Y., Song, H., Jara, A. J., & Bie, R. (2021). Internet of Things and Big Data analytics for smart and connected communities. *IEEE Access, 4*, 766-773.

[15] Hui Yie Teh1, Andreas W. Kempa-Liehr and Kevin I-Kai Wang (2020), Sensor data quality: a systematic review,Journal of Big Data, 1-49