Kernel Tricks for Non-Linear Data in Climate Modeling

Submitted by:

Zeel Doshi,Amit Adhikari,Sumit Banerjee

Roll No: M22MA209,M22MA201,M22MA210

M.Tech in DCS, First Year

Supervised by:

Dr. Angshuman Paul

Indian Institute of Technology Jodhpur

Date of Submission: November 2024

Contents

1 Introduction

Climate modeling is crucial in understanding the behavior of weather patterns and making future predictions. Often, climate data exhibits non-linear relationships, making it difficult for traditional linear models to accurately predict outcomes. To address this issue, kernel-based techniques, such as Support Vector Machines (SVMs) and Kernel Principal Component Analysis (K-PCA), offer effective tools to handle non-linear data by mapping it into higher-dimensional spaces where linear separability becomes possible.

This report explores how kernel methods, particularly the application of Support Vector Regression (SVR) and PCA, can enhance climate modeling by transforming the data into a form that is better suited for machine learning models.

2 Problem Statement

This project aims to enhance climate modeling accuracy by using kernel tricks, such as support vector machines (SVMs) and kernel principal component analysis (K-PCA). We will test different kernel functions—linear, polynomial, radial basis function (RBF), and sigmoid—and evaluate their impact on model performance. The goal is to create a climate classification system that outperforms traditional models in predicting and classifying climate patterns.

3 Dataset Description

The dataset used in this study is the daily climate data from Delhi, India, which includes the following variables:
The link of the data set :`https://drive.google.com/file/d/1wUw1J6Xar7439pe1hRMmkpOQCJ16fvha/view?usp=sharing`

- **date**: The date of each observation.
- **meantemp**: Mean temperature recorded for the day (°C).
- **humidity**: Mean relative humidity for the day (%).
- **wind_speed**: Average wind speed (km/h).
- **meanpressure**: Mean atmospheric pressure (hPa).

3.1 Statistical Summary of the Dataset

The dataset was examined to understand the central tendency and dispersion of the variables. The following table summarizes key statistical metrics for the climate features:

	humidity	wind_speed	meanpressure	meantemp
count	1576.000000	1576.000000	1576.000000	1576.000000
mean	60.445229	6.899262	1010.593178	25.221918
std	16.979994	4.510725	175.242704	7.345014
min	13.428571	0.000000	-3.041667	6.000000
25%	49.750000	3.700000	1001.875000	18.500000
50%	62.440476	6.363571	1009.055556	27.166667
75%	72.125000	9.262500	1015.200000	31.142857
max	100.000000	42.220000	7679.333333	38.714286

Figure 1: Descriptive statistics of the climate dataset.

4 Linear vs. Non-Linear Data Analysis

4.1 Defining Linear and Non-Linear Data

Linear data refers to datasets where relationships between variables can be captured with a straight-line equation. Non-linear data, however, requires more complex models to capture patterns such as polynomial or exponential relationships.

5 Exploratory Data Analysis

5.1 Visualization: Scatter Matrix

A scatter matrix was used to visualize the relationships between climate features. It provides a detailed view of the pairwise relationships, highlighting patterns that suggest non-linearity.

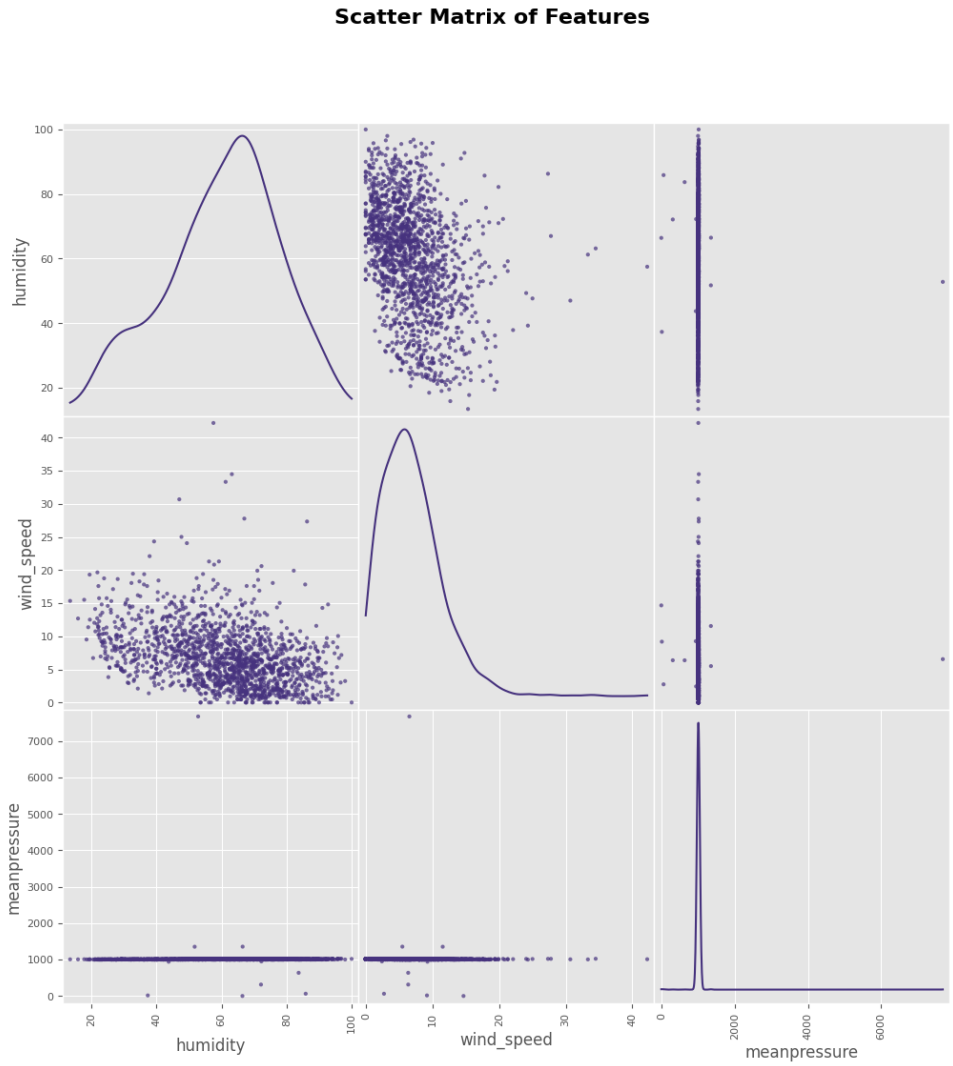

Figure 2: Scatter Matrix of Climate Features

5.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was employed to reduce the dimensionality of the data and examine the linearity of relationships between features. PCA projects the data onto its most significant components, helping to visualize the data in lower dimensions.

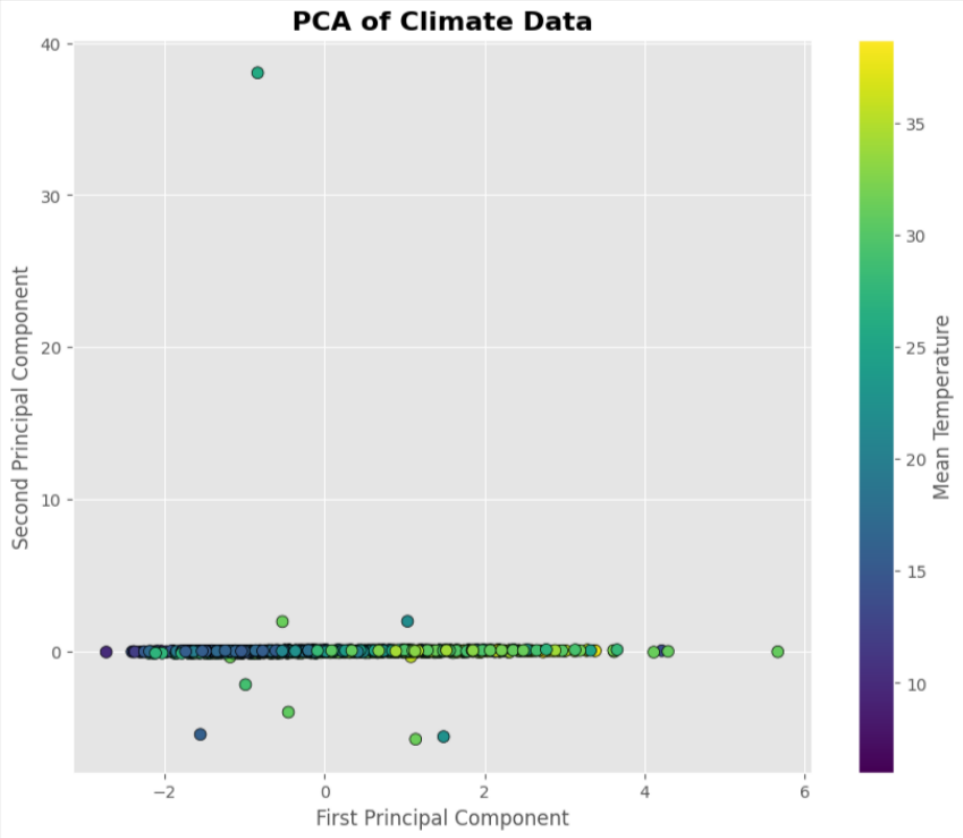

Figure 3: PCA Visualization of Climate Data

Interpretation of PCA Results: The plot shows that the climate data exhibits some structure in the first two principal components, indicating that there are potentially non-linear relationships that could be captured through kernel-based methods.

5.3 Correlation Matrix

The correlation matrix was computed to explore the strength of linear relationships between the variables. The following heatmap visualizes the correlations:

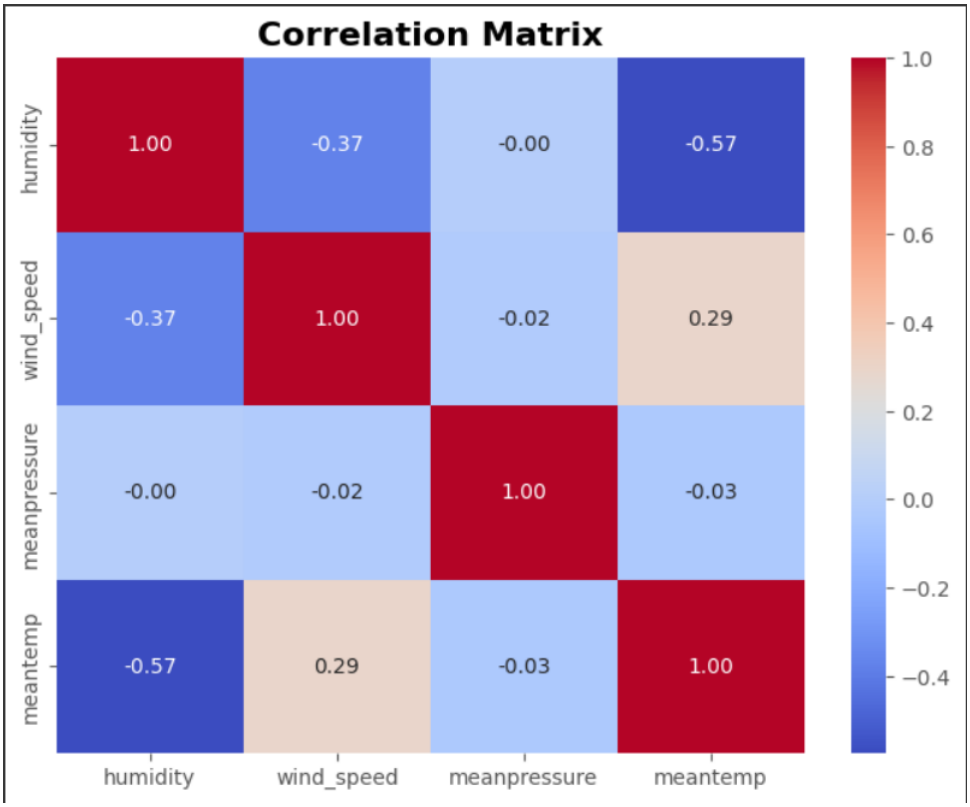

Figure 4: Correlation Matrix of Climate Features

Interpretation: The correlation matrix shows that most of the features exhibit weak to moderate correlations, which suggests that a linear model may not adequately capture the underlying relationships between the variables.

5.4 Residual Analysis

A residual plot was generated to assess the fit of a linear regression model. The residuals show a pattern, indicating that a non-linear model may be more appropriate for this dataset.

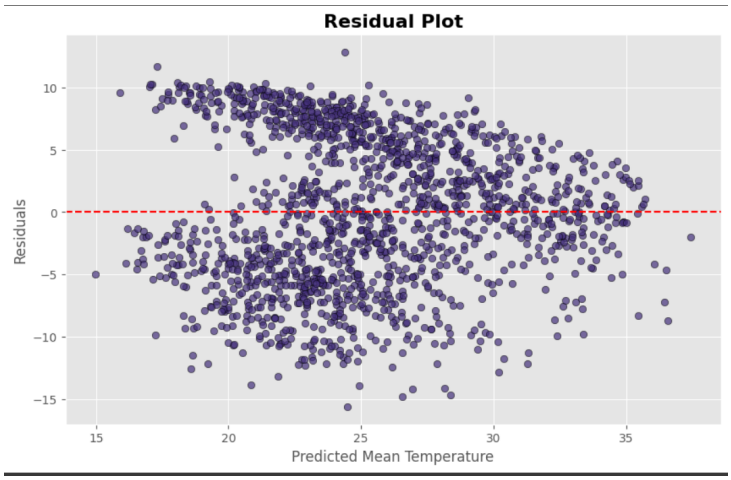

Figure 5: Residual Plot of Linear Regression Model

Interpretation: The distinct pattern in the residuals supports the hypothesis that the data is non-linear and would benefit from kernel-based methods.

5.5 Conclusion from Visualizations

From the exploratory data analysis, including the scatter matrix, PCA plot, and residual analysis, it is evident that the dataset exhibits non-linear relationships between the climate variables. The patterns observed suggest that linear models will not be sufficient to capture the underlying dynamics, and kernel-based methods would provide a more effective approach for modeling the data.

6 Kernel Methods for Non-Linear Modeling

Kernel methods map data into higher-dimensional spaces to make complex, non-linear relationships linear. By using the kernel trick, these methods avoid explicit computation of the higher-dimensional transformations, making them computationally efficient.

6.1 Overview of Kernel Functions

In this study, we test four kernel functions. Each kernel function transforms the input data into a higher-dimensional space, where linear techniques can be applied to solve otherwise non-linear problems. The following are the kernel functions used in the study:

- **Linear Kernel**:
$$K(x, x') = x \cdot x'$$
The linear kernel is the simplest kernel and is used when the data is already linearly separable. It computes the dot product between two data points. This kernel is computationally efficient and works well when the decision boundary is a straight line.

- **Polynomial Kernel**:
$$K(x, x') = (x \cdot x' + c)^d$$
The polynomial kernel computes the dot product between two data points raised to a certain degree d, with an optional constant c. This kernel is suitable for capturing polynomial relationships between the features, which can model more complex patterns in the data.

- **Radial Basis Function (RBF) Kernel**:
$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

The RBF kernel is one of the most widely used kernel functions, effective in capturing localized patterns in the data. The kernel measures the distance between two points and applies an exponential function to it, with a parameter σ controlling the spread of the influence of each data point. It is particularly useful when the data exhibits complex, non-linear relationships.

- **Sigmoid Kernel**:

$$K(x, x') = \tanh(\alpha x \cdot x' + c)$$

The sigmoid kernel is inspired by the activation function used in neural networks. It captures non-linear relationships by mapping data into a hyperbolic tangent function. This kernel is less commonly used, but can still be effective in certain scenarios, especially when data exhibits neural-like patterns.

7 Climate Modeling using Support Vector Machines (SVM)

7.1 Methodology

Support Vector Regression (SVR) was applied to model the climate data. The model was trained using different kernels: linear, polynomial, RBF, and sigmoid. The methodology is based on the following steps:

1. **Data Preprocessing**: The dataset was standardized using the StandardScaler to ensure that all features have the same scale.
2. **Model Training**: SVR models were trained for each kernel.
3. **Model Evaluation**: The models were evaluated based on performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 Score.

7.2 Results and Discussion

The Python code involved scaling the data, performing exploratory data analysis (EDA), and evaluating multiple kernel functions through SVR. The performance of each kernel was assessed based on the following evaluation metrics:

Kernel	**MAE**	**MSE**	**R^2**
Linear	5.23	41.47	0.25
Polynomial	6.45	59.51	-0.07
RBF	**5.10**	**36.34**	**0.34**
Sigmoid	6.44	59.28	-0.06

Table 1: Performance of SVR Models with Different Kernels

The results indicate that the RBF kernel outperforms the other kernels in terms of both error metrics and R^2 score, making it the best choice for this dataset.

8 Conclusion

The study demonstrates that kernel methods are particularly effective in modeling non-linear relationships in climate data. The RBF kernel, in particular, provided the best performance in predicting mean temperature, indicating its suitability for complex, non-linear climate patterns.

9 Future Work

- Exploring deep learning architectures to capture complex patterns in large-scale climate data.
- Implementing time-series analysis with lag features for more accurate climate predictions.
- Optimizing hyperparameters for kernel functions.

10 References

- S. Haykin, *Neural Networks and Learning Machines*, Pearson, 2008.
- C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.