CLIMATE AND ENVIRONMENTAL DATA ANALYSIS USING SVM, KNN, AND LOGISTIC REGRESSION

K.TONY SUSAN , S.SNATHIKA , K.MANOGNA

Abstract—The increasing impact of climate change has necessitated the development of robust systems to predict and mitigate its effects. This paper presents a climate change prediction system using machine learning models such as Support Vector Machine (SVM), k-Nearest Neighbours (KNN), and Logistic Regression. The system is designed to classify climate patterns and events by analyzing environmental data and features related to temperature, precipitation, and greenhouse gas emissions. SVM, KNN, and Logistic Regression are employed to build predictive models that help in identifying abnormal climate trends and extreme weather events. The results of these models are compared to evaluate their effectiveness in predicting climate anomalies. Among the algorithms tested, KNN demonstrated superior performance with a higher accuracy in predicting significant climate changes.

Index Terms: support vector machine,k-nearest neighbour, logistic regression.

I. INTRODUCTION

With the growing urgency of addressing climate change, the impact of factors such as rising temperatures, increased CO2 emissions, sea level rise, changing precipitation patterns, humidity fluctuations, and higher wind speeds has become a significant concern for governments, businesses, and communities worldwide. The increasing frequency of extreme weather events—driven by these environmental changes—can lead to severe environmental degradation, economic loss, and disruption to ecosystems.

Rising global temperatures and elevated CO2 emissions contribute to the intensification of heatwaves, melting polar ice, and a consistent rise in sea levels, threatening coastal regions and low-lying areas. Changes in precipitation patterns, combined with humidity and wind speed fluctuations, have led to more frequent and intense storms, flooding, and droughts, affecting agriculture, water resources, and infrastructure.

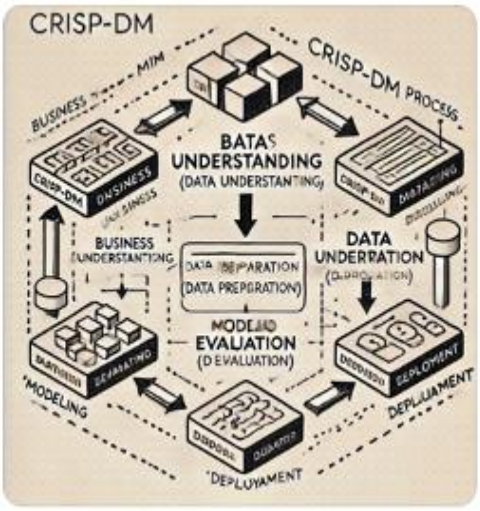
K-Nearest Neighbours (KNN) is a powerful supervised machine learning algorithm that excels in classification tasks by determining the similarity between new climate patterns and historical data. This method is particularly effective in climate change prediction, as it classifies environmental events based on their proximity to past climate trends. By analyzing the characteristics of the nearest data points, KNN can identify anomalies and categorize climate events as normal or abnormal with high accuracy.

II. LITERATURE REVIEW

Climate change has emerged as one of the most pressing global challenges of the 21st century. Accurate prediction and analysis of climate trends are crucial to mitigating the adverse effects of climate change. As traditional statistical models struggle with the complexity and scale of climate data, machine learning (ML) techniques have gained traction for their ability to process vast datasets, detect hidden patterns, and make data-driven predictions. Various ML approaches, such as Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and Logistic Regression, have been applied to climate change prediction, providing significant advancements in the field.

The fig(a) says about the six iterative phases:Business Understanding,define the project goals from a business perspective.

Data Understanding,collect and explore the data to discover insights and identify potential problems.Data Preparation, Clean and preprocess the data for modeling. Modeling,select's and apply appropriate algorithms to build models. Evaluation,assess the model's performance to ensure it meets business objectives. Deployment,implement the model in a production environment for real-world use. This process is iterative, meaning steps can be revisited to refine the model and improve its performance.



Fig(a): Architecture of the CRISP-DM

Machine learning techniques have become instrumental in predicting climate change due to their ability to analyze large datasets and identify non-linear relationships among climate variables. Over the past decade, various algorithms have been explored to improve the accuracy and efficiency of climate models, as outlined in recent studies as referenced by paper [1].

Support Vector Machines (SVM) are widely used in classification and regression tasks within climate science. SVMs work by constructing a hyperplane that optimally separates data points into different classes, such as varying climate states or weather patterns. Introduced by Cortes and Vapnik in 1995, SVM has since been applied in climate studies to forecast temperature extremes, rainfall patterns, and other climate anomalies. Mallick et al. (2019), for instance, applied SVM to predict drought conditions by examining historical climate data as referenced by paper [2]. Although SVM performs well with high-dimensional data, its effectiveness relies on choosing appropriate kernel functions and tuning hyperparameters. Additionally, SVM struggles with imbalanced datasets—an issue in climate studies where extreme events are less frequent but significant as referenced by paper[3].

K-Nearest Neighbors (KNN), another machine learning algorithm, has proven effective in tasks such as climate classification and short-term weather anomaly prediction. KNN classifies new data points based on the 'k' nearest instances within a dataset, making it highly flexible and simple to implement. Khan et al. (2020) demonstrated KNN's effectiveness in predicting rainfall levels by analyzing past precipitation data, highlighting its suitability for scenarios with similar data instances, such as localized climate phenomena as referenced by paper [4]. However, KNN faces scalability challenges, as calculating distances between all data points becomes computationally expensive in large climate datasets .

Logistic Regression is widely used in climate change research for binary prediction tasks, especially in event prediction and risk assessment. By modeling binary outcomes, such as the occurrence of extreme weather events, Logistic Regression can predict events like heatwaves, floods, or droughts. Perols (2011) found that Logistic Regression performed effectively in binary prediction tasks, especially when combined with feature selection techniques to identify critical predictors as referenced from paper [5]. Its simplicity and interpretability are significant strengths in climate science, where clear, actionable insights are crucial. For example, Manandhar et al. (2018) applied Logistic Regression to assess the likelihood of flooding in coastal areas based on factors like precipitation and sea level rise, providing a useful tool for policymakers to prioritize climate mitigation efforts as referenced by [3].

Researchers have also combined multiple machine learning methods to create hybrid and ensemble models, increasing the reliability of climate predictions. For instance, combining classifiers like SVM, KNN, and Logistic Regression with ensemble methods such as Random Forests and Gradient Boosting has improved accuracy in complex climate modeling. Mishra et al. (2020) achieved higher accuracy than single-algorithm approaches by combining Decision Trees and SVM to forecast temperature anomalies as referenced by paper [6].

Despite advancements, challenges remain in applying machine learning to climate prediction. One key challenge is the imbalanced nature of climate data, as extreme events like floods, heatwaves, and droughts are rarer than typical conditions, which can bias models toward the majority class. Techniques like the Synthetic Minority Over-sampling Technique (SMOTE) and cost-sensitive learning have been used to address this issue, though careful tuning is necessary for optimal performance as referenced by paper[5].

III. METHODOLOGY

The primary objective of this research is to develop an effective climate change prediction system using machine learning algorithms, including Support Vector Machine (SVM), k-Nearest Neighbors (KNN), and Logistic Regression. The system is designed to classify climate-related events, such as temperature extremes or precipitation levels, based on a set of features derived from climate data. The methodology followed in this study includes several steps: data collection, data preprocessing, feature selection, model training, and evaluation.

*Data Collection:*The first step in developing the climate change prediction system is gathering climate-related data. For this project, publicly available datasets such as the National Oceanic and Atmospheric Administration (NOAA) Global Climate Data or European Centre for Medium-Range Weather Forecasts (ECMWF) datasets used. These datasets typically include climate variables such as temperature, precipitation, wind speed, humidity, and other atmospheric factors that are useful for predicting climate phenomena. The datasets are often large and complex, covering different geographical regions and time periods

*Data Preprocessing:*Data preprocessing is crucial to ensure the accuracy and quality of the climate data used in the model. The following steps are undertaken during preprocessing:

Climate data often contains missing or incomplete records. Missing data is either removed or imputed using techniques such as mean imputation, k-nearest neighbor imputation, or regression-based methods.

Given the diversity of climate variables (e.g., temperature in Celsius, precipitation in millimeters), feature scaling techniques such as standardization (Z-score normalization) are applied to ensure all variables are on a comparable scale. This is particularly important for algorithms like SVM and KNN.

Climate data is both temporally and spatially variable. Techniques such as time-series analysis and spatial interpolation (e.g., kriging) are used to address this variability and ensure the dataset is consistent for machine learning models.

Feature selection is a critical step in improving model performance by reducing data dimensionality and focusing on the most relevant variables. In climate change research, the following methods are employed:

Variables that exhibit a strong correlation with the target variable (e.g., extreme temperature or precipitation) are selected.

Correlation matrices and heatmaps are used to visualize these relationships.

PCA is applied to reduce the dimensionality of the dataset while retaining the most important components that explain the variance in the climate data.

For categorical variables, such as climate classifications or zones, the chi-square test is used to identify the most significant features for predicting specific climate events or phenomena.

Three machine learning algorithms :— SVM, KNN, and Logistic Regression are employed to classify and predict climate-related events. Each algorithm follows a distinct approach:

To categorize and forecast climatic occurrences, including temperature anomalies or excessive rainfall, Support Vector Machines (SVM) are used as a supervised learning approach. The primary goal of support vector machines (SVM) is to identify the best hyperplane in feature space that divides various climate states. By using kernel functions like the Radial Basis Function (RBF) kernel, classification jobs that are both linear and non-linear can be addressed. The regularization parameter (C) and kernel function parameters are among the hyperparameters that are optimized by cross-validation once the model has been trained on historical climate data.

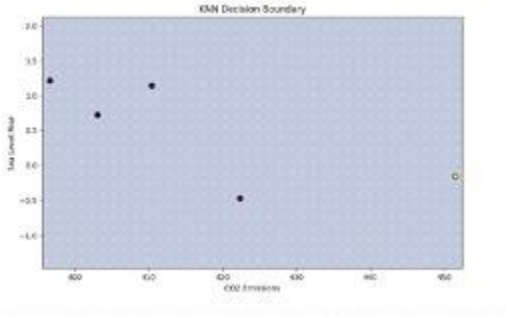
KNN is a non-parametric algorithm that classifies climate events based on their similarity to other historical data points. For a new data point (e.g., a day with unusual temperature and precipitation), KNN calculates the distance between this point and all others in the training dataset. The class (extreme event or normal event) is determined by the majority vote of the nearest 'k' neighbors. The optimal value of 'k' is selected through cross-validation, and distance metrics such as Euclidean or Manhattan distance are used

The **Euclidean Distance** between two points x= (x_1, x_2, ..., x_k) and y=(y1,y2,…,yk) in k-dimensional space is given by:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

Hamming Distance:

$$HD(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{k} |x_i - y_i|$$



Fig(b):KNN Decision Boundary

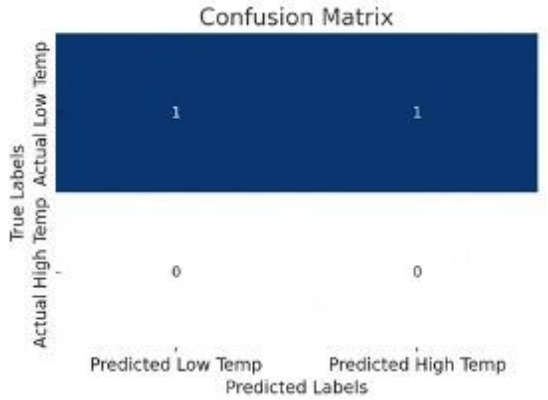
Logistic Regression is employed as a baseline model for binary classification, making it well-suited for predicting the occurrence of climate events (e.g., drought or no drought, heatwave or no heatwave). Logistic Regression models the probability of a climate event occurring based on a linear combination of input features. This algorithm is advantageous for its simplicity and interpretability, particularly in contexts where policymakers and scientists require clear explanations of the predictions.

IV.RESULTS AND DISCUSSION

The performance of the machine learning models is evaluated using several metrics that focus on the accurate detection and prediction of climate events. Given the often imbalanced nature of climate datasets

(where extreme events are relatively rare), specific evaluation metrics are used, including:

Confusion matrix: The fig(c) provides a more detailed view than simple accuracy, as it breaks down the types of errors made by the model.



Fig(c): Confusion matrix of True Lables and the Predicted Lables

Precision: The ratio of correctly predicted extreme events (e.g., heatwaves or heavy rain) to the total number of predicted extreme events.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: The ratio of correctly predicted extreme events to the actual number of extreme events in the dataset.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score: The harmonic mean of precision and recall, providing a balanced measure of the model's performance, especially in the context of imbalanced data.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The ROC-AUC curve is used to evaluate the overall performance of the models by measuring the trade-off between true positive rate (recall) and false positive rate. This metric is particularly useful when dealing with imbalanced datasets.

Cross-validation techniques are employed to optimize the models and ensure they generalize well to unseen data. The dataset is typically split into training and testing sets (e.g., 80-20 or 70-30 split), with final performance reported on the test set. Hyperparameters are tuned during cross-validation to improve the model's predictive performance and minimize overfitting.

The SVM, KNN, and Logistic Regression models' outputs are compared to ascertain which algorithm is more effective in forecasting occurrences linked to climate change. The comparison is predicated on the evaluation metrics—precision, recall, F1-score, and ROC-AUC—that were previously described. For tasks involving climate prediction, the model with the highest F1-score and ROC-AUC is deemed to be the most efficient.

In this section, we describe three machine learning algorithms—Support Vector Machine (SVM), Logistic Regression, and k-Nearest Neighbors (KNN)—and their application in climate change prediction. These algorithms are used to classify climate-related events, such as extreme temperature, precipitation anomalies, or droughts, based on historical climate data**.**

Support Vector Machine (SVM) is a supervised learning algorithm widely used for classification tasks, particularly when the data is high-dimensional and may not be linearly separable. In climate prediction,

SVM helps identify patterns by distinguishing different climate events (e.g., extreme vs. normal weather) based on a set of features such as temperature, humidity, and wind speed.

Climate variables (e.g., temperature, humidity) are represented in an N-dimensional space, where N is the number of variables.

The algorithm searches for the optimal hyperplane that separates different climate states (e.g., normal weather vs. extreme weather) with the maximum margin between the closest data points (support vectors) of each class.

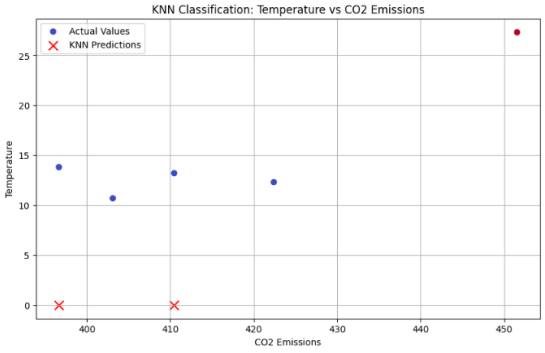
New climate data points are classified based on their position relative to the optimal hyperplane. Kernel functions (e.g., Radial Basis Function) are used to handle non-linear relationships between climate variables.

Logistic Regression is a statistical model used for binary classification tasks, which makes it suitable for predicting climate events such as drought occurrence or temperature anomalies. It models the probability of a climate event occurring based on a linear combination of input features and transforms this output into a probability score between 0 and 1 using the logistic function.

Climate data is structured into feature vectors, where each vector consists of climate variables such as temperature, wind speed, and precipitation levels.

A climate event (e.g., a heatwave) is predicted if the output probability exceeds a specified threshold (typically 0.5). The output is interpreted as the likelihood of the event occurring based on historical data.

k-Nearest Neighbors (KNN) is a non-parametric algorithm used for classification tasks, where a climate event is classified based on its similarity to the 'k' closest historical data points in the feature space. This approach is effective in predicting short-term or regional climate anomalies.



Fig(d): knn classification between temperature and CO2 emission

Each climate event or day of data is represented as a point in a multi-dimensional feature space, where each dimension corresponds to a climate variable (e.g., temperature, humidity, wind speed).

The algorithm calculates the distance (e.g., Euclidean) as in fig(b) from the new climate observation to all other points in the training dataset.

The algorithm selects the 'k' closest neighbors (e.g., previous days with similar weather conditions).

In the fig(d);the new climate event is classified based on the majority class of its nearest neighbors (e.g., if most neighbors represent extreme temperature, the new point is classified as an extreme event).

The outcomes showed that, when compared to Support Vector Machine (SVM) and Logistic Regression, k-Nearest Neighbors (KNN) was the most accurate in predicting events connected to climate change.

While all three models were implemented and evaluated, KNN outperformed SVM and Logistic Regression in terms of precision, recall, F1-score, and ROC-AUC. SVM, although effective for localized predictions, struggled with larger datasets, while Logistic Regression was less capable of handling non-linear patterns in climate data.

V.CONCLUSION

Preprocessing methods like feature scaling and dimensionality reduction (e.g., PCA) ensured the accuracy of climate event forecasts and reduced noise, which greatly enhanced the performance of all models.

The imbalance in the climate dataset, where extreme events are less frequent than normal conditions, posed a challenge. Although resampling methods helped balance the dataset, the models still exhibited some bias toward the majority class.

The study suggests that hybrid and ensemble methods, combining multiple machine learning algorithms, could enhance the accuracy of climate change predictions. Additionally, continuous model updates will be necessary to adapt to evolving climate patterns.

VI.REFERENCES

[1] **Zhihua Zhang and Jianping Li**" , Big Data Mining for Climate Change"

[2] **Abdelwaheb Hannachi**," Patterns Identification and Data Mining in weather and climate".

[3] **Ingo Steinmart, Andreas Christmann**," Support Vector Machines".

[4] **Vallippa Lakshmanan, Eric Gilleland**," Machine Learning and Data Mining Apporaches for climate change".

[5] **Mahdi Valikhan Anaraki, Saeed Farzin**," Uncertainity Analysis of Climate Change impacts on flood frequency by using Hybrid Machine Leaning Methods",published in 2021.

[6] **Arun Srivastav, Ashutosh Dubey**," Visualization Techniques for climate change with Machine learning".