Spatiotemporal extreme event prediction over the Indogangatic plane using Machine Learning

Name:	**Prachi Chachondhia**
Registration No./Roll No.:	2311003
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	August 17, 2023
Date of Submission:	November 19, 2023

1 Introduction

Widespread low visibility poses challenges across sectors, notably in places like Delhi, impacting safety and activities, especially during winter due to factors like fog and smog. Machine learning, including regression models with historical weather data, aids in predicting visibility. Classification models, using urban built-up data, inform urban planning for mitigating low visibility impacts. These models contribute to safety protocols, transportation management, and urban development policies. Continuous innovation and interdisciplinary collaboration are crucial for refining models and addressing the evolving challenges in low visibility risks [1].

The dataset spans from 1942 to the present day, capturing extensive weather data from various Indian locations. With 934,807 training and 103,868 testing recordings, featuring 122 features, it presents challenges of noise and imbalance. To enhance usability, we performed pre-processing, retaining 22 features after eliminating columns with predominant null values. Key features include geographic details (STATION, LATITUDE, LONGITUDE, and ELEVATION) and weather-related metrics (Altimeter Setting, Dew Point Temperature). Time details (YEAR, MONTH, DAY, HOUR, MINUTES) enable temporal pattern analysis. The TARGETS variable underwent outlier removal, focusing on values within the 0 to 20 km range. The dataset serves as a valuable resource for spatiotemporal weather exploration in India.[2]

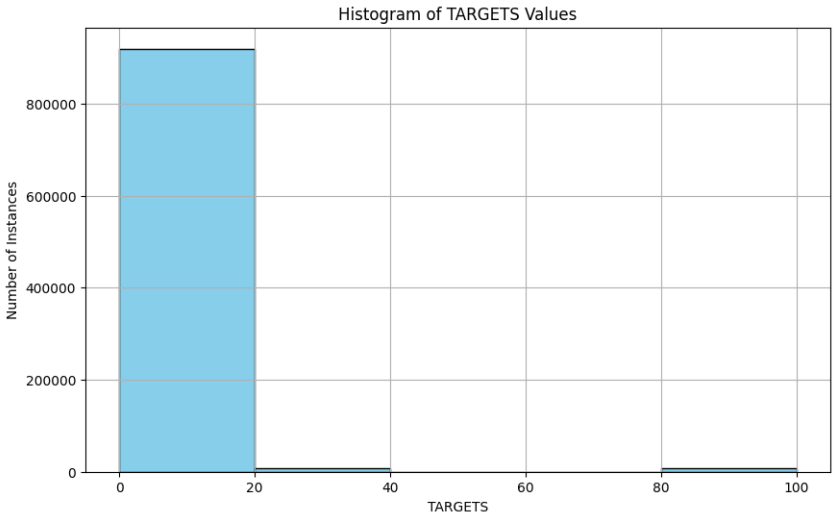
2 Methods

2.1 Data Preprocessing

In the initial phase of data preprocessing, we optimized the dataset for subsequent analyses by eliminating columns with predominantly null values, resulting in a streamlined dataset comprising 22 columns. However, as only one feature vector remained without null values, we applied data-filling techniques to enhance usability.

2.2 Encoding

To address missing values in categorical variables like "HourlyPresentWeatherType" and "HourlySkyConditions," a pivotal step involved converting categorical values into numeric ones. This conversion was essential to enable the application of kNNImputer, a technique that exclusively operates with numeric variables. The conversion process entailed creating a mapping that assigned each category a corresponding numeric value.



2.3 Imputation Methods

Handling missing values in spatiotemporal data, where time and location information are intertwined, requires specialized methods. We explored two approaches:

Yearly Imputation: Instead of using a single median for the entire dataset, we calculated medians for each year separately, considering different locations. This nuanced approach allowed us to fill in missing values based on the median corresponding to the respective year and location, providing a more tailored imputation.`https://github.com/PCprachi/spationtemporal_data_ML`

KNN Imputation: Specifically designed for spatiotemporal data, KNN imputation considers both spatial and temporal dimensions when filling in missing values. It identifies the k-nearest neighbors in terms of both space and time and utilizes their known values to estimate and fill in the missing data point. This tailored approach acknowledges the unique characteristics of spatiotemporal data, where both geographical location and timing play crucial roles.

2.4 Classification

For the classification task, we excluded data before the year 2000 and utilized Landsat 8 satellite imagery to identify built-up areas across 30 stations using Google Earth Engine. For datasets with numeric visibility targets, a discretization process categorized visibility levels into bins (e.g., very low, low, medium, high). Three distinct approaches were explored:

PCA (Principal Component Analysis): Streamlining features through PCA facilitated the identification of classification algorithms with higher accuracy. Evaluation involved the use of a confusion matrix to assess performance.

Withoversampler Method: To address dataset imbalance, oversampling on the validation set involved replicating samples from the minority class to achieve balance.

SMOTE (Synthetic Minority Over-sampling Technique): This technique focused on the feature space to generate new instances through interpolation between positive instances close to each other, mitigating overfitting.

Subsequently, four classification models (KNN, Decision Tree, Random Forest, and Gradient Boosting) were applied, and their performance was systematically compared. Hyperparameter tuning, specifically on the oversampled data, further optimized the models.

2.5 Regression

For regression tasks, we employed GridSearchCV in conjunction with Random Forest (RF) and Linear Regression (LR) models. GridSearchCV systematically explored hyperparameter combinations to identify optimal configurations for each model, enhancing their predictive accuracy.

These comprehensive preprocessing steps lay the foundation for robust analyses and model development in addressing the challenges posed by low visibility. [3]

3 Experimental Setup

3.1 Classification

The performance metrics of various classification models were evaluated under different scenarios: With PCA (Principal Component Analysis), With SMOTE (Synthetic Minority Over-sampling Technique), and With Oversampler.

	With PCA				**With SMOTE**				**With Oversampler**			
Classifiers	RF	DT	GB	KNN	RF	DT	GB	KNN	RF	DT	GB	KNN
Weighted F1 score	0.99	0.98	0.98	0.98	0.66	0.59	0.51	0.51	0.62	0.66	0.32	0.32
Accuracy	0.99	0.99	0.99	0.99	0.60	0.60	0.42	0.42	0.64	0.66	0.42	0.42

Table 1: Performance of different models in classification

Under the PCA scenario, each classifier (RF, DT, GB, KNN) demonstrates robust performance with Weighted F1 scores close to 1.0 and high accuracy nearing 0.99, indicating overall model robustness.

However, with the use of SMOTE and Oversampler techniques, there is a noticeable reduction in both Weighted F1 score and accuracy compared to the PCA scenario, despite maintaining generally high performance. Random Forest (RF) consistently outperforms other classifiers in this context.

	With SMOTE				**With Oversampler**			
Classifiers	RF	DT	GB	KNN	RF	DT	GB	KNN
Weighted F1 score	0.56	0.59	0.15	0.33	0.62	0.65	0.15	0.32
Accuracy	0.60	0.60	0.28	0.42	0.64	0.66	0.28	0.42

Table 2: Performance of different models with hyperparameter tuning

During the application of GridSearchCV with StratifiedKFold (k=5) on a RandomForest classification model, the hyperparameters were set as follows: min_samples_leaf=1, min_samples_split=2, and n_estimators=200. To address class imbalance, specified class weights were provided for each class during training. Additionally, SMOTE was employed for data augmentation, and the cross-validation accuracy obtained was 0.96.

3.2 Regression

	Global Imputation			**KNN Imputation**		
Regressor	LR	RF	DT	LR	RF	DT
R^2	-0.284	0.527	0.66	1.0	0.99	0.99
MAE	0.5088	0.42	0.30	1.41	9.54	6.99

Table 3: Performance of different models on different Imputation methods

In our analysis for KNN imputation, a Decision Tree (DT) model was employed with the following hyperparameters: max_depth of 10, min_samples_split of 2, and min_samples_leaf of 1. For the Random Forest (RF) model, the hyperparameter configuration included 200 estimators, max_depth of 10, min_samples_split of 2, and min_samples_leaf of 1.

Global Imputation:

The Decision Tree (DT) model exhibits the highest R^2 (0.66) and lowest MAE (0.30), making it a strong candidate for the best model. The Random Forest (RF) model also performs well with an R^2 of 0.527 and an MAE of 0.42. Linear Regression (LR) appears to be less effective with a negative R^2 and higher MAE.

KNN Imputation:

Linear Regression (LR) shows a perfect R^2 (1.0), but the MAE is high (1.41), suggesting potential overfitting. Decision Tree (DT) and Random Forest (RF) models are consistent with high R^2 values (0.99) but have higher MAE values compared to the Global Imputation results.

In summary, the Decision Tree (DT) model with Global Imputation appears to be the most favorable among the models based on the provided metrics. However, it's crucial to consider other factors and potentially explore additional evaluation metrics or cross-validation for a comprehensive assessment.

4 Results & Discussions

Throughout this project, we discovered that traditional imputation methods fall short when dealing with spatiotemporal datasets. Spatial techniques such as KNN prove more effective in filling missing data. While these techniques may not completely fill all gaps, they significantly enhance dataset usability without causing a substantial decline in feature-target correlations.

5 Conclusion

In the future, exploring various interpolation techniques could enhance performance on this spatiotemporal weather dataset. Additionally, there is potential for experimentation with multiple time series analysis methods to extract valuable insights. Another avenue for improvement lies in exploring advanced deep learning techniques, which could contribute to achieving better model performance.

6 Readme

For classifcation dataset used :merged_df.csv For regression dataset used: a) for knnimputation imputated_data and for simpleimputation given dataset.

References

[1] Sandy Wong Johnathan Rush Itai Kloog Allan C. Just Iván Gutiérrez-Avila, Kodi B. Arfer. A spatiotemporal reconstruction of daily ambient temperature using satellite data in the megalopolis of central mexico from 2003 to 2019. 2021.

[2] Arkapal Panda, Tanmay Basu, and Vaibhav Kumar. An ensemble learning framework for visibility prediction in indo-gangetic region. 2023.

[3] Y Zhu L Yang L Ge C Luot Y Zhang, Y Wang. Visibilty prediction based on machine learning algorithms. 2022.