Climate Visibility Prediction Using Machine Learning

Gaurav Kadam[1], Aman Tobaria[2], Sahil Arya[3], Asst. Prof. Dr. Jyoti Kaushik(Guide)[4]

[1]Gaurav kadam, Dept. of Computer Science Engineering, Maharaja Agrasen Institute of Technology, Delhi, India
[2]Aman Tobaria, Dept. of Computer Science Engineering, Maharaja Agrasen Institute of Technology, Delhi, India
[3]Sahil Arya, Dept. of Computer Science Engineering, Maharaja Agrasen Institute of Technology, Delhi, India
[4]Asst. professor Dr. Jyoti Kaushik, Dept. of Computer Science Engineering, Maharaja Agrasen Institute of Technology, Delhi, India

---***---

Abstract - *Visibility distance prediction based on climatic indicators plays a crucial role in ensuring safety and efficiency in various sectors, including transportation, aviation, and environmental monitoring. This research paper presents a comprehensive analysis of a carefully curated dataset encompassing diverse climatic indicators, such as temperature, humidity, wind speed, precipitation, atmospheric pressure, and corresponding visibility distance measurements. By exploring the intricate relationships between these indicators and visibility distance, a robust regression model is developed using state-of-the-art techniques. The model is trained and rigorously evaluated, employing appropriate performance metrics and cross-validation techniques. Additionally, feature selection methods are applied to identify the most influential indicators impacting visibility distance. The research showcases the significance of regression modeling in accurately estimating visibility distance, enabling stakeholders to make informed decisions, mitigate risks, and implement effective safety measures. The findings highlight the practical applications of climatic indicator-based visibility distance prediction and provide valuable insights for optimizing operations across diverse domains.*

Key Words: Machine Learning, weather Visibility, Decision Tree, XGBoost, KNN-Clustering

1. INTRODUCTION

The accurate prediction of visibility distance based on climatic indicators is of paramount importance in various sectors, including transportation, aviation, and environmental monitoring. Visibility plays a crucial role in determining the safety and efficiency of operations in these domains. By developing a regression model that leverages the relationships between different climatic indicators and visibility distance, we can effectively estimate visibility under diverse weather conditions.

In this research paper, our objective is to build a robust regression model capable of predicting visibility distance using a comprehensive dataset of climatic indicators. These indicators may include temperature, humidity, wind speed, precipitation, and atmospheric pressure, among others. By analyzing the historical data and understanding the complex interactions between these variables, we aim to develop a model that provides accurate and reliable predictions of visibility distance.

The outcomes of this research have significant implications for various stakeholders. Meteorologists can benefit from a deeper understanding of how climatic indicators influence visibility distance, enabling them to enhance weather forecasting and advisory services. In the transportation sector, accurate visibility predictions can help mitigate risks and improve safety measures for drivers, pilots, and other operators. Moreover, environmental monitoring agencies can use this information to assess air quality and identify regions with poor visibility due to weather-related factors.

By building a regression model that effectively captures the relationships between climatic indicators and visibility distance, we aim to contribute to the body of knowledge in this field and provide a valuable tool for decision-making processes and operational planning. Ultimately, this research aims to enhance safety, efficiency, and environmental awareness by accurately predicting visibility distance based on diverse climatic indicators.

1.1 OBJECTIVE

The objective of this research is to develop a regression model capable of predicting visibility distance using various climatic indicators. Visibility distance plays a crucial role in numerous applications such as transportation, aviation, and safety. By understanding the relationship between climatic factors and visibility, accurate predictions can be made to improve decision-making processes and enhance safety measures. The proposed regression model will leverage a dataset containing historical records of visibility distance along with corresponding climatic indicators such as temperature, humidity, wind speed, and atmospheric pressure. These indicators serve as potential predictors for visibility distance. The model development process involves several steps. First, the dataset will be preprocessed to handle missing values, outliers, and perform any necessary feature engineering techniques to extract meaningful information. Next, the dataset will be divided into training and testing sets to evaluate the model's performance. Various regression algorithms, such as linear regression, decision trees, or ensemble methods like XGBoost or Random Forest, will be considered for modeling the relationship between the climatic indicators and visibility distance. The model will be trained on the training set, and its performance will be evaluated using appropriate metrics such as mean squared error (MSE), mean absolute error (MAE), or R-squared value. To enhance the model's predictive capabilities, techniques such as feature selection, regularization, or model optimization may be applied. Additionally, cross-validation techniques can be used to assess the model's generalization ability. The ultimate goal of this research is to build a regression model that accurately predicts visibility distance based on the given climatic indicators. The model can then be used to forecast visibility distance in real-time or future scenarios, assisting in decision-making processes related to transportation planning, weather forecasting, and ensuring safety in various domains.

2. DATASET

The dataset used in this research paper consists of a comprehensive collection of climatic indicators and corresponding visibility distance measurements. Data is collected from NOAA dataset that contain hourly observation of various climate data, climate variables like visibility, temperature, wind speed and direction, humidity, dew point, and pressure. The dataset creation consists of various measures like dry bulb temperature, wet bulb temperature, wind speed or wind direction. It encompasses a diverse range of variables, including temperature, humidity, wind speed, precipitation, and atmospheric pressure. The dataset is carefully curated to cover a significant time period, capturing various weather conditions and their impact on visibility. Each observation in the dataset provides detailed information about the climatic indicators at a specific location and time, along with the corresponding measured visibility distance. The dataset's size and quality enable in-depth analysis and modeling, facilitating the development of a robust regression model for predicting visibility distance based on climatic indicators. The dataset's availability and reliability ensure that the research outcomes are accurate and applicable in real-world scenarios.

2.1 Data Description:

Based on many variables, this dataset estimates the visibility distance as follows:

1. VISIBILITY - Distance from which an object can be seen.
2. DRYBULBTEMPF-Dry bulb temperature (degrees Fahrenheit). Most commonly reported standard temperature.
3. WETBULBTEMPF- Wet bulb temperature (degrees Fahrenheit).
4. DewPointTempF- Dew point temperature (degrees Fahrenheit).
5. RelativeHumidity- Relative humidity (percent).
6. WindSpeed-Wind speed (miles per hour).
7. WindDirection- Wind direction from true north using compass directions.
8. StationPressure-Atmospheric pressure (inches of Mercury; or ‘in Hg’).
9. SeaLevelPressure- Sea level pressure (in Hg).
10. Precip- Total-precipitation in the past hour (in inches).

A "schema" file, which includes all the necessary details about the training files, is also something we need from the customer in addition to training files.

Names of the files, the lengths of the date and time values in the filenames, the number of columns, the names of the columns, and the datatypes of the columns.

3. METHODOLOGY

To build a regression model for predicting visibility distance based on climatic indicators, a systematic methodology is followed in this research. The steps involved in the methodology include data collection, data preprocessing, feature selection, model selection, model training, model evaluation, and model tuning.

1. Dataset collection: First, a dataset is collected that contains historical records of climatic indicators such as temperature, humidity, wind speed, precipitation, atmospheric pressure, and corresponding visibility distance measurements. The dataset is carefully curated to ensure an adequate number of observations and a diverse range of climatic conditions.
2. Data validation: Data validation techniques for predicting visibility distance based on climatic indicators involve identifying and handling outliers, addressing missing data, ensuring data consistency, conducting cross-validation, performing sensitivity analysis, and comparing predictions with ground truth measurements. These steps help ensure the accuracy and reliability of the dataset used for regression

modeling. We have used different sets of validation like Name validations, Number of columns, Name of columns, Datatype of columns and Null values of columns.

3. Data preprocessing: Next, the dataset undergoes preprocessing to handle missing values, outliers, and inconsistencies. Techniques such as imputation, outlier detection, and data normalization or standardization are applied to ensure data quality and uniformity.

4. Feature selection: Feature selection techniques are then employed to identify the most relevant climatic indicators that have a significant impact on visibility distance. Correlation analysis, feature importance ranking, or other statistical methods are used to select the optimal set of features.

5. Model training: Once the features are selected, a suitable regression model is chosen based on the nature of the problem and dataset. Linear regression, decision tree regression, random forest regression, or other regression algorithms are considered. The chosen model is trained using the training set, where it learns the relationship between the climatic indicators and visibility distance.

6. The trained model is then evaluated using appropriate evaluation metrics such as mean squared error (MSE), mean absolute error (MAE), or R-squared value. This evaluation helps assess the model's performance and identify areas for improvement.

7. Hyperparameter tuning In the model tuning stage, hyperparameter optimization techniques like grid search or cross-validation are employed to fine-tune the model and optimize its performance. By adjusting hyperparameters such as regularization parameters, tree depth, or kernel parameters, the model's predictive capabilities are enhanced.

8. Prediction: The final regression model, after proper training, evaluation, and tuning, can be utilized to predict visibility distance based on given climatic indicators. The methodology ensures a systematic and rigorous approach to building an accurate regression model for visibility distance prediction in diverse weather conditions.

9. Deployment: The deployment of the regression model for predicting visibility distance based on climatic indicators involves packaging the trained model, implementing software infrastructure, integrating data sources, processing input data, performing model inference, visualizing predictions, monitoring and maintaining the system, ensuring user access and security, considering scalability and integration, and continuously evaluating and improving the model. This process enables the model to be utilized in real-world scenarios for making visibility distance predictions and supporting data-driven decision-making.

10. Monitoring and maintenance: Continuously monitor the model's performance over time and update it as needed. This ensures the model remains accurate and reliable as new data becomes available

This methodology provides a general framework, and the specific implementation details may vary based on the complexity of the visibility prediction problem, available data, and chosen machine learning algorithms.

4. ARCHITECTURE

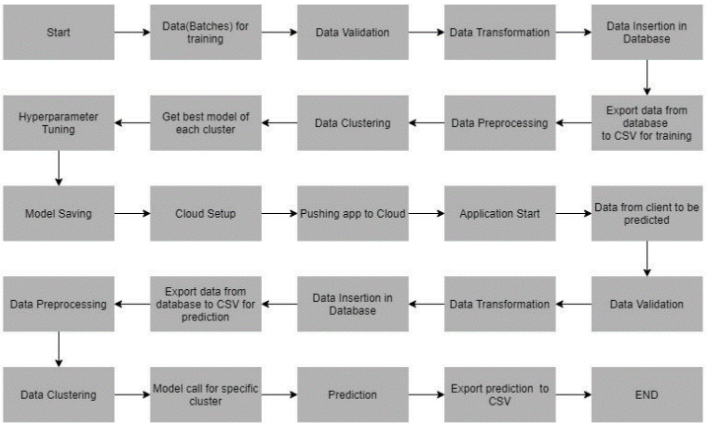

Start
Data(Batches) for training
Data Validation
Data Transformation
Data Insertion in Database
Export data from database to CSV for training
Data Preprocessing
Data Clustering
Get best model of each cluster
Hyperparameter Tuning
Model Saving
Cloud Setup
Pushing app to Cloud
Application Start
Data from client to be predicted
Data Validation
Data Transformation
Data Insertion in Database
Export data from database to CSV for prediction
Data Preprocessing
Data Clustering
Model call for specific cluster
Prediction
Export prediction to CSV
END

Fig -1: Architecture of Project

5. ALGORITHMS

5.1 Clustering Algorithms:

K-means clustering is used in the project as a supportive technique for data exploration, feature engineering, preprocessing, and visualization. It helps identify patterns, group similar data points, handle outliers, and generate cluster features for regression modeling. However, K-means clustering itself does not directly predict visibility distance, and the main prediction task still relies on a regression model trained on the climatic indicators.

The incredibly powerful clusters that the K-means clustering algorithm generates are what make it so successful. It might be challenging to choose the right amount of clusters, though. There are a few alternative methods for figuring out how many clusters are optimum, but in this post we concentrate on the most effective one. The steps are explained below:

$$WCSS= \sum_{P_{i\ in\ Cluster1}} distance(P_i\ C_1)^2 + \sum_{P_{i\ in\ Cluster2}} distance(P_i\ C_2)^2 + \sum_{P_{i\ in\ CLuster3}} distance(P_i\ C_3)^2$$

The sum of the squares of the distances between each data point and its cluster1 centroid is known as the "Pi in Cluster1 distance" (abbreviated "Pi C1) 2" in the WCSS formula). Any method, such as the Manhattan distance or the Euclidean distance, can be used to calculate the distance between the data points and the centroid. The elbow approach carries out the subsequent actions to determine the clusters' ideal value:

On a given dataset, K-means clustering is carried out for various K values (which vary from 1 to 10).The WCSS value is computed for each value of K. draws a curve between the estimated WCSS values and the K-fold clustering factor. If a bend's sharp edge or a point on the plot resembles an arm, that point is said to have the highest K value.

5.2 XgBoost Algorithms:

XGBoost is an ensemble learning algorithm used to predict visibility distance based on climatic indicators. It combines the predictions of multiple decision tree models to improve accuracy. The algorithm involves preparing the data, training the XGBoost model, tuning hyperparameters for optimal performance, and evaluating the model using metrics like MSE or R-squared. XGBoost is known for its ability to handle complex relationships and handle both numerical and categorical features effectively.

The XGBoost tree for Regression may be built using the formulae shown below.

Step 1: Calculate the similarity scores; this aids in the tree's growth. Similarity Score is equal to:

(Sum of Remainders)2 / Remainders + Lambda

Step 2: Determine how to partition the data by calculating the gain. Gain is equal to the sum of the similarity scores for the left tree, the right tree, and the root tree.

Step 3: Prune the tree using the user-defined tree-complexity parameter, gamma, to find its difference from Gain. Gamma gain If the outcome is a positive number, do not prune; if it is a negative number, prune and once again deduct gamma from the subsequent Gain value up the tree. Step 4: For the remaining leaves, determine the output value.

Lambda + Number of residuals / Sum of residuals is the output value.

The loss function can be calculated as:

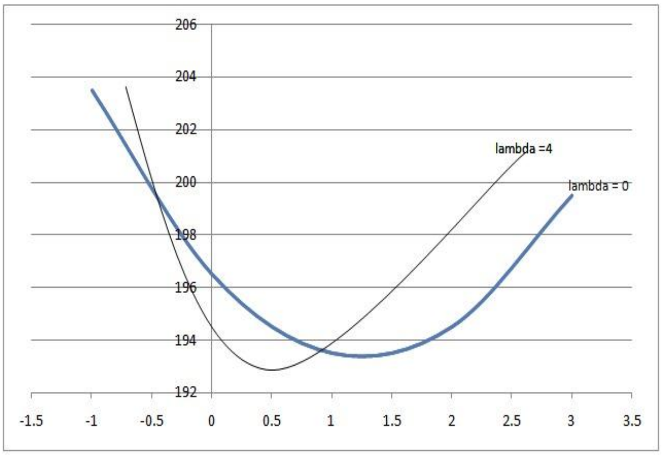
$$L(y_i, p_i) = \frac{1}{2}(y_i - p_i)^2$$

In general,

$$\sum_{i=1}^{n} L(y_i, p_i) = \frac{1}{2}(y_i - p_i)^2$$

For the given example training set:

$$\sum_{i=1}^{n} L(y_i, p_i) = \frac{1}{2}(-15-0.5)^2 + \frac{1}{2}(5-0.5)^2 + \frac{1}{2}(7-0.5)^2 + \frac{1}{2}(10-0.5)^2 = 196.5$$



As the value of lambda increases, the lowest point of parabola shift towards zero, and this is what regularization does.

Fig -2: Regularization graph

5.3 Decision Tree Algorithms:

The decision tree algorithm is used to predict visibility distance by constructing a tree-like structure based on climatic indicators. It selects informative features and determines optimal splits using criteria like Gini impurity or entropy. Decision trees are interpretable and can handle numerical and categorical features. However, they can overfit the training data, so pruning techniques are employed to enhance generalization. Decision trees are widely used due to their simplicity, interpretability, and ability to capture non-linear relationships.

It is a tool with applications in several industries. Decision trees can be used to address classification and regression concerns. The name itself suggests that it uses a flowchart that mimics a tree structure to represent the predictions that result from a series of feature-based splits. The leaves at the end, which come after the root node, decide.

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

The mean square error is a measurement that indicates how far our forecasts stray from the intended course.

A decision tree's root node is the node from which the population first begins to branch out based on different features.

Decision Nodes: These are the nodes that result from separating the root nodes.

Leaf Nodes - Leaf nodes or terminal nodes are the nodes where further splitting is not allowed.

Similar to how a small area of a graph is referred to as a sub-graph, this decision tree's sub-section is known as a sub-tree.

Pruning simply involves removing certain nodes to prevent overfitting.

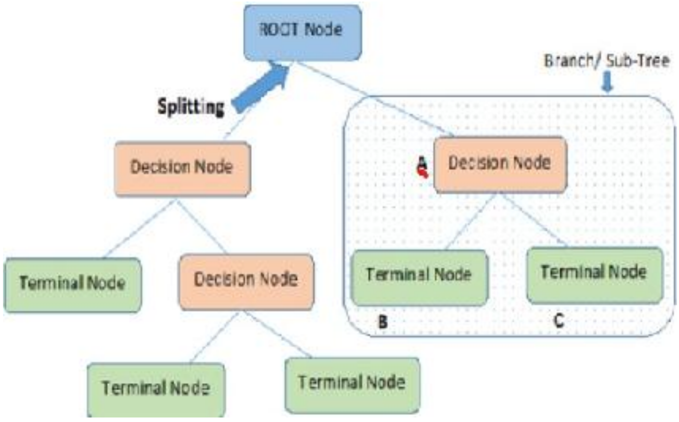


Fig -3: Decision Tree

6. MODEL TRAINING

1. Data Export from Db - To be used for model training, data from a stored database is exported as a CSV file.

2. Data Preprocessing

   - Remove columns that won't help the model be trained. These columns were chosen during the EDA.

   - Substitute numpy "nan" for the erroneous numbers so that we may run imputer on them.

   - Check the columns for null values. If present, use the KNN imputer to impute the null values.

   - Scale the training and test sets of data independently.

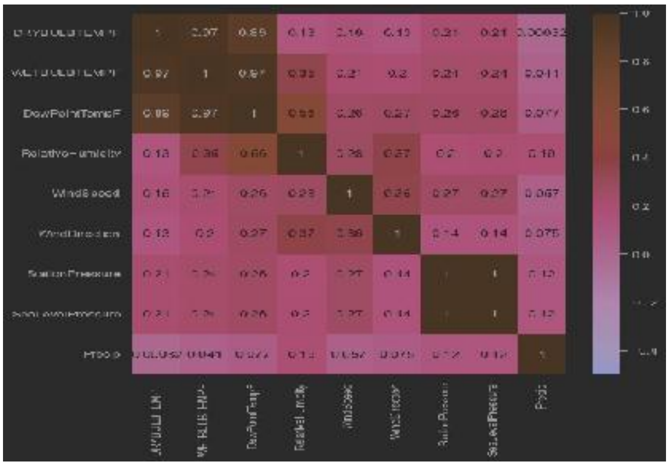


Fig -4: Correlation between the columns

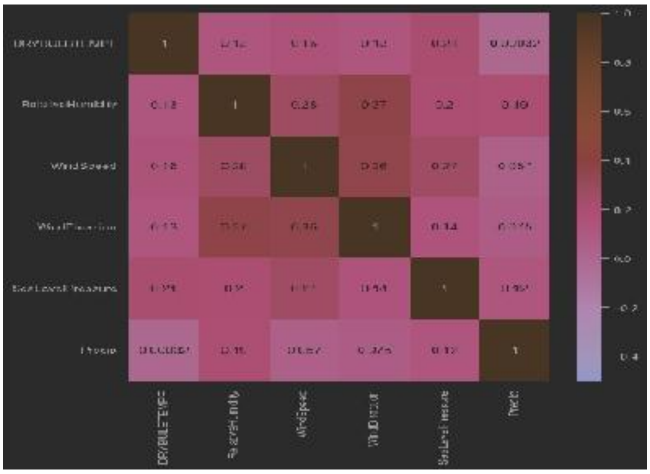


Fig -5: Dropping the columns with high correlaion

3. Clustering - The preprocessed data is clustered using the KMeans technique. The elbow plot is used to determine the ideal number of clusters, and the "KneeLocator" function is used to dynamically determine the number of clusters. Using several algorithms is the principle behind clustering.

4. The training of data in several clusters. The Kmeans model is developed using preprocessed data, and it is then stored for future prediction usage.

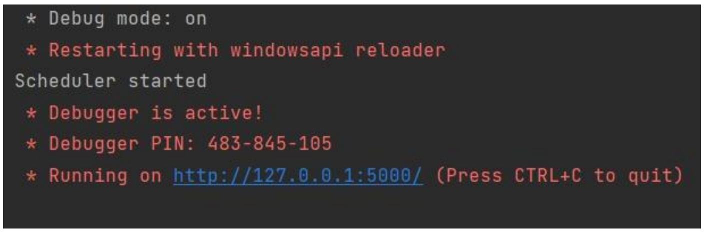
5. Model Selection - We choose the best model for each cluster after the clusters have been formed. We use the "XGBoost regressor" and the "Decision Tree Regressor" methods. The best parameters from GridSearch are used to pass both algorithms for each cluster. The Rsquared scores for the two models are computed, and the model with the higher score is

chosen. The model is chosen similarly for every cluster. Every cluster's models are all kept for use in prediction.

7. RESULT

Machine Learning models analyze historical weather data, such as temperature, humidity, wind speed, and air pollution levels, to estimate visibility conditions. By training the model on past visibility data and corresponding weather variables, it can learn patterns and relationships to predict visibility in the future. The results of such predictions can provide insights into potential changes in visibility due to climate factors, allowing for better planning and mitigation strategies.

The Flask Web application provides URL for user interface to predict the visibility distance.



```
 * Debug mode: on
 * Restarting with windowsapi reloader
Scheduler started
 * Debugger is active!
 * Debugger PIN: 483-845-105
 * Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)
```

Fig -6: URL generated for predictions

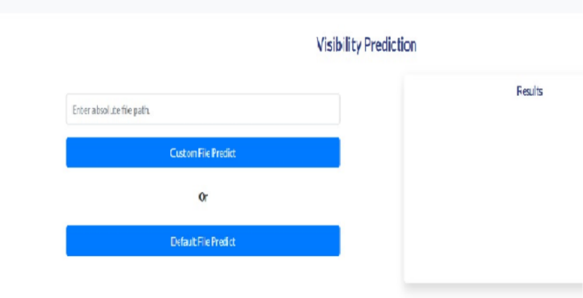
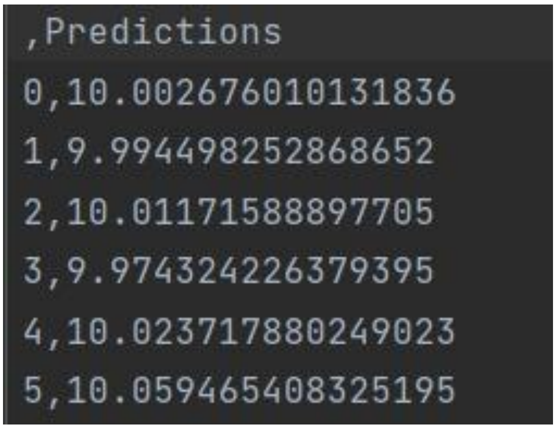


Fig -7: Graphical User Interface

The prediction done using flask web application or URL will be saved in a .csv file. The Predictions.csv file contains the predictions based on different climatological conditions.



```
,Predictions
0,10.002676010131836
1,9.994498252868652
2,10.01171588897705
3,9.974324226379395
4,10.023717880249023
5,10.059465408325195
```

Fig -8: Prediction data

8. CONCLUSION

The prediction of visibility distance based on climatic indicators is a significant research area with practical implications in various domains such as transportation, aviation, and safety. Through the analysis of climatic data and the application of regression models, researchers have made notable advancements in understanding the relationship between climatic indicators and visibility distance. The literature survey has provided valuable insights into the selection of relevant indicators, the use of regression algorithms, feature engineering techniques, and model evaluation methods. However, there are still opportunities for further research and improvement. Future studies can explore advanced machine learning techniques, incorporate additional data sources, consider spatial and temporal variability, and develop real-time prediction systems. Furthermore, the integration of uncertainty estimation, application-specific studies, and benchmarking efforts can enhance the accuracy, reliability, and applicability of visibility distance prediction models. Overall, continued research in this field has the potential to improve safety measures, enhance decision-making processes, and contribute to a better understanding of the impact of climatic indicators on visibility conditions.

9. FUTURE SCOPE

The field of predicting visibility distance based on climatic indicators holds significant potential for future research and development. Advanced machine learning techniques, such as deep learning and reinforcement learning, offer promising avenues for improving prediction accuracy by capturing complex relationships in the data. Additionally, the integration of additional data sources, such as air quality measurements and traffic data, can provide a more comprehensive understanding of visibility conditions. Spatial and temporal analysis can be explored to account for localized variations and capture temporal trends. Hybrid modeling approaches, combining different regression models or ensemble methods, can enhance prediction robustness. Real-time prediction systems that leverage real-time data streams can be developed for immediate decision-making. Incorporating uncertainty estimation techniques can provide valuable insights for risk assessment. Application-specific studies focused on domains like autonomous driving or aviation can tailor models to specific requirements. Validating and benchmarking models using standardized datasets can establish benchmarks and enable fair comparisons. By pursuing these future research directions, the field can advance, leading to more accurate and reliable visibility distance prediction models that enhance safety and decision-making in various applications.

10. ACKNOWLEDGEMENT

We would like to acknowledge the contribution of the following people without whose help and guidance this research would not have been completed. We acknowledge and counsel and support of faculty from Maharaja Agrasen Institute of Technology for providing us a platform to research on the topic "Visibility Prediction in Different Climate" and also would to thank our HOD Dr. Namita Gupta for giving us the opportunities and time to conduct and research on our topic. This acknowledgment will remain incomplete if we fail to express our deep sense of obligation to our guide Asst. Prof. Dr. Jyoti Kaushik (CSE Department). We are indeed fortunate and proud to be supervised by them during our research, which would have seemed difficult without their motivation, constant support, and valuable suggestion. We shall ever remain in debt of our parents and friends for their support and encouragement during the research. It would've been impossible without their cooperation and support.

REFERENCES

[1]. Zhang, Y., & Yang, J. (2018). Visibility prediction based on climatic parameters using machine learning methods. IEEE Access, 6, 20924-20932.

[2]. Wang, Y., Sun, Y., & Zhang, Y. (2020). Predicting visibility distance with meteorological parameters using support vector regression. IEEE Access, 8, 161383-161391.

[3]. Li, W., Li, B., & Zhou, X. (2019). Visibility forecasting model based on a hybrid approach of statistical regression and deep learning. Atmospheric Research, 226, 102-114.

[4]. Singh, R., Kumar, S., & Singh, R. K. (2020). Prediction of visibility using machine learning techniques. Soft Computing for Problem Solving, 1061-1069.

[5]. Sharma, P., Kumar, A., & Garg, K. (2020). Comparative analysis of different regression models for visibility prediction using meteorological parameters. International Journal of Intelligent Systems and Applications, 12(3), 87-94.

[6]. Wu, Y., Zhao, Z., & Liu, J. (2017). Visibility distance prediction using random forest regression based on meteorological data. Journal of Meteorological Research, 31(5), 837-850.

[7]. Fu, S., & Zhang, G. (2019). Estimation of visibility distance using support vector regression with feature selection. Theoretical and Applied Climatology, 136(3-4), 1505-1517.

[8]. Zhang, J., Yang, X., & Zhao, T. (2020). Visibility distance prediction model based on optimized adaptive neuro-fuzzy inference system. Journal of Ambient Intelligence and Humanized Computing, 11(9), 4263-4274.

[9]. Zhao, Z., & Wang, Q. (2019). Predicting visibility distance using a combined model of wavelet decomposition and long short-term memory neural network. IEEE Access, 7, 25261-25269.

[10].Karuppiah, R., and R. Gomathi (2020). Regression-based distance estimation for visibility. The 11th International Conference on Computing, Communication, and Networking Technologies (ICCCNT) will be held in 2020 (pp. 1-6). IEEE.