BUILDING REGRESSION MODELS TO PREDICT VISIBILITY DISTANCE UNDER VARIOUS CLIMATIC CONDITIONS

Pratim Mukherjee
Department of Mathematics,
School of Advance Science
Vellore Institute of Technology
Vellore, INDIA, 632014
pratim.mukherjee2023@vitstudent.ac.in

Soumyadip Tikader
Department of Mathematics,
School of Advance Science
Vellore Institute of Technology
Vellore, INDIA, 632014
soumyadip.tikader2023@vitstudent.ac.in

Mohammad Sarim Fayajoddin Kazi
Department of Mathematics,
School of Advance Science
Vellore Institute of Technology
Vellore, INDIA, 632014
mohammad.sarim2023@vitstudent.ac.in

Amrit Das
Department of Mathematics,
School of Advance Science
Vellore Institute of Technology
Vellore, INDIA, 632014
amrit.das@vit.ac.in

***Abstract*— The accurate prediction of visibility distance is vital for diverse human activities, from transportation and aviation to outdoor recreation and air quality monitoring. This atmospheric parameter is intricately linked to climatic factors for example temperature, humidity, wind speed, precipitation, and atmospheric pressure. Employing regression models presents a promising avenue for forecasting visibility distance, with techniques like linear regression, polynomial regression, decision trees, random forests, and support vector machines being viable options. Recent interest has surged in leveraging machine learning for visibility prediction due to its capacity to discern complex, non-linear relationships within data. However, challenges persist, including the scarcity of high-quality visibility data and the intricate, non-linear connections between predictor and target variables. Researchers have addressed data limitations through imputation and synthetic data generation techniques. Mitigating the risk of overfitting in machine learning models involves employing cross-validation and regularization techniques. Despite these challenges, regression models emerge as a valuable tool for developing accurate forecasting tools adaptable to various climatic conditions, offering reliability for multiple applications.**

Keywords—Visibility-distance,Regression-models,Climatic factors, Machine learning

I. INTRODUCTION

Visibility is a critical element influencing safety and operational efficiency in transportation [1], particularly when adverse weather conditions, such as fog, rain, and snow, pose challenges for drivers, pilots, and navigators. Accurate prediction of visibility distance becomes essential for implementing preemptive safety measures[2]. Traditional visibility prediction methods often fall short in capturing the intricate relationships between climatic variables and visibility which causes accidents on roads[3], Reduced visibility caused by air pollutants and relative humidity (RH) has become a serious environmental problem [4] necessitating a more sophisticated approach also, visibility is a simple indicator of air quality, and sharp decreases in visibility are often associated with increased aerosol concentrations [5] and during periods of high pollution, a low visibility of ≤5 km is closely related to severe haze [6], [7]. This study addresses this gap by employing advanced regression modeling techniques to create precise visibility prediction models. These models consider the dynamic nature of atmospheric interactions, accounting for fluctuations in temperature, humidity, precipitation, and atmospheric pressure.[[8]–[10]]

The research utilizes a comprehensive dataset from diverse geographical locations and temporal periods, ensuring the robustness of the models across various climates and seasons. Beyond transportation safety, the study's findings extend to meteorology, climate science, and environmental monitoring, enhancing our understanding of atmospheric processes. The resultant regression models are poised to provide practical insights for decision-makers, transportation authorities, and meteorological services. They can serve as invaluable tools for real-time visibility assessment, enabling adaptive strategies to enhance safety and mitigate the impact of adverse climatic conditions on transportation systems. As climate change transforms global weather patterns, the ability to predict visibility becomes increasingly crucial for the resilience and sustainability of transportation networks, aligning with endeavors to formulate proactive measures addressing evolving climatic patterns.

II. DATA COLLECTION AND FILTERING

In this section, the fundamental phases of data acquisition and data preprocessing are elucidated, offering a comprehensive overview of the foundational processes involved in the research study. The data collection section encompasses information regarding the dataset's origin and the specific procedures involved in its acquisition. Subsequently, the data preprocessing phase will expound upon the essential steps, including data cleaning, transformation, integration, and quality assurance measures.

A. Data Collection

In this study, 'Meteostat'[11] Python package has been used to access up-to-date information. The data retrieved was stored as a CSV file and contains daily summaries for New York City. This dataset covers the period from January 1, 1960, to October 16, 2023, providing a comprehensive historical view of weather conditions. The dataset comprises 9 attributes, including average temperature, minimum temperature, maximum temperature, total daily precipitation, average wind direction, average wind speed, wind peak gust, average sea-level pressure, total sunshine duration. With 23,301 records, it serves as a valuable resource for our model. Utilizing the 'meteostat' package ensures a reliable and standardized approach to obtaining weather-related information, supporting the integrity and accuracy of our analyses in this project.

B. Data Filtering

In the data filtering phase, essential decisions were made to optimize the dataset for analysis. Redundant attributes, namely ‘tmin’, ‘tmax’ , ‘snow’, ‘wdir’, ‘wpgt’ and ‘tsun’ were removed. Then rows with at least 4 non-null values are kept. All leftover null values were filled using statistical methods like linear interpolation for ‘tavg’ and ‘wspd’ and mean for ‘pres’.In conclusion, the calculation of visibility was performed explicitly, as it was not provided as a parameter in the raw data; this aspect is integral to the study.

Nomenclature

tmin: Minimum Temperature
tmax: Maximum Temperature
wdir: Wind Direction
wpgt: Wind peak gust
tsun: Total Sunshine
tavg: Average Temperature
wspd: Windspeed
pres: Average sea-level pressure

III. PARAMETER SELECTION AND PREDICTION

For the model, it is essential to know about the correlations between all the variables. Hence correlation heatmap has been used.

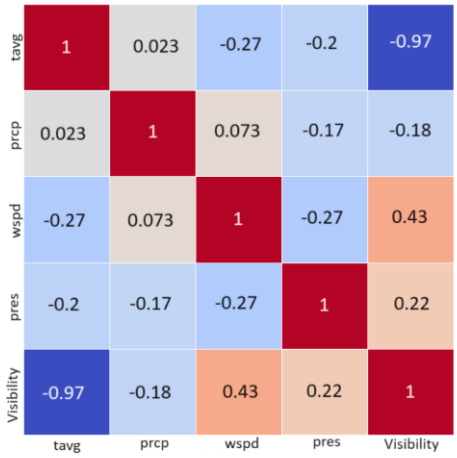


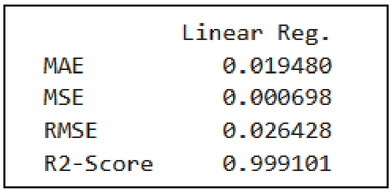
Fig1: Correlation Heatmap showing relationship between attributes.

With the help of heatmap it is evident that there is high correlation between average temperature and visibility.

In this study, visibility is predicted, with the dependent variable being visibility, and the other attributes serving as independent variables.

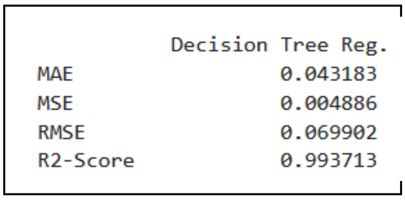
Evaluation of Model

This article uses four cost functions of sklearn to evaluate the four regression models, the first being the Mean Absolute Error(MAE), the second being Mean Squared Error(MSE) method, the third being Root Mean Squared Error(RMSE) method and the fourth being R-squared(R^2) method.



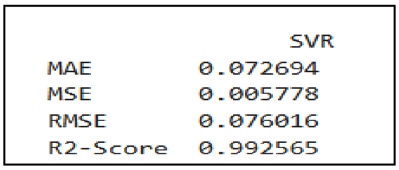
```
              Linear Reg.
MAE             0.019480
MSE             0.000698
RMSE            0.026428
R2-Score        0.999101
```

Fig2: Cost Function scores for Linear Regression



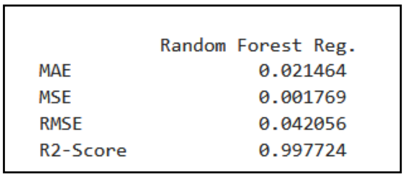
```
               Decision Tree Reg.
MAE                      0.043183
MSE                      0.004886
RMSE                     0.069902
R2-Score                 0.993713
```

Fig3: Cost Function score for Decision tree regression



```
                   SVR
MAE          0.072694
MSE          0.005778
RMSE         0.076016
R2-Score     0.992565
```

Fig4: Cost Function scores for Support vector regression



```
             Random Forest Reg.
MAE                    0.021464
MSE                    0.001769
RMSE                   0.042056
R2-Score               0.997724
```

Fig5: Cost Function scores for Random Forest regression

IV. RESULT

Performance metrics in machine learning models gauge the accuracy of the model. The assessment of model accuracy in this study involves the utilization of Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R² scores.

MAE: Mean Absolute Error refers to the magnitude of difference between the prediction of an observation and the true value of that observation. It takes the average of absolute errors for a group of predictions and observations as a measurement of the magnitude of errors for the entire group. Its value increases as the model error increases.[12]

MSE: Mean Squared Error is defined as the mean or average of the squared differences between the actual and estimated values. Its value increases as the model error increases.[13]

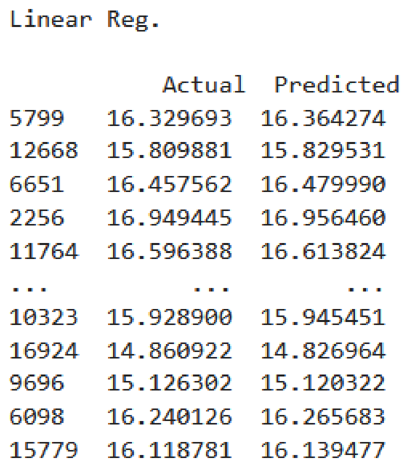
RMSE: It measures the average difference between values predicted by a model and the actual values. It provides an estimation of how well the model can predict the target value (accuracy). The lower the value of the Root Mean Squared Error, the better the model is.[14]

R^2: R-Squared determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model. The more value of R2, the more accurate the model is.[15]

V. Conclusion

As the attribute “average temperature” and “windspeed” follows a linear relationship with “Visibility”, it is evident that the Linear Regression model is well suited for the dataset.

This study makes notable contributions to the field of weather visibility prediction, showcasing remarkable performance metrics in the cost functions for the Linear Regression model: a 1% Mean Absolute Error (MAE), approximately 0% Mean Squared Error (MSE), a 2% Root Mean Squared Error (RMSE), and a 99% coefficient of determination (R^2), surpassing the comparative models. Subsequently, based on meticulous data analysis and extensive cost function evaluation, the conclusion is drawn that the Linear Regression model is the most fitting and effective choice for the dataset.



```
Linear Reg.

            Actual  Predicted
5799     16.329693  16.364274
12668    15.809881  15.829531
6651     16.457562  16.479990
2256     16.949445  16.956460
11764    16.596388  16.613824
...            ...        ...
10323    15.928900  15.945451
16924    14.860922  14.826964
9696     15.126302  15.120322
6098     16.240126  16.265683
15779    16.118781  16.139477
```

Fig6: Comparison between Actual and Predicted Value

VI. Future scope

the near future, using this machine learning model further advancement can be made by integrating in different platform like an web-application, building APIs,etc

References

[1] L. C. Ortega, L. D. Otero, M. Solomon, C. E. Otero, and A. Fabregas, “Deep learning models for visibility forecasting using climatological data,” *Int. J. Forecast.*, vol. 39, no. 2, pp. 992–1004, Apr. 2023, doi: 10.1016/j.ijforecast.2022.03.009.

[2] Y. Hu and F. Xiao, “A novel method for forecasting time series based on directed visibility graph and improved random walk,” *Phys. Stat. Mech. Its Appl.*, vol. 594, p. 127029, May 2022, doi: 10.1016/j.physa.2022.127029.

[3] C. Peláez-Rodríguez, J. Pérez-Aracil, C. Casanova-Mateo, and S. Salcedo-Sanz, “Efficient prediction of fog-related low-visibility events with Machine Learning and evolutionary algorithms,” *Atmospheric Res.*, vol. 295, p. 106991, Nov. 2023, doi: 10.1016/j.atmosres.2023.106991.

[4] Y.-C. Ting, L.-H. Young, T.-H. Lin, S.-C. Tsay, K.-E. Chang, and T.-C. Hsiao, “Quantifying the impacts of PM2.5 constituents and relative humidity on visibility impairment in a suburban area of eastern Asia using long-term in-situ measurements,” *Sci. Total Environ.*, vol. 818, p. 151759, Apr. 2022, doi: 10.1016/j.scitotenv.2021.151759.

[5] A. Molnár, K. Imre, Z. Ferenczi, G. Kiss, and A. Gelencsér, “Aerosol hygroscopicity: Hygroscopic growth proxy based on visibility for low-cost PM monitoring,” *Atmospheric Res.*, vol. 236, p. 104815, May 2020, doi: 10.1016/j.atmosres.2019.104815.

[6] X. Shen *et al.*, “A novel method of retrieving low visibility during heavily polluted episodes in the North China plain,” *Atmospheric Environ. X*, vol. 9, p. 100101, Jan. 2021, doi: 10.1016/j.aeaoa.2021.100101.

[7] “National Center for Biotechnology Information.” Accessed: Nov. 11, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/

[8] C.-Y. Chen and H.-M. Feng, “Hybrid intelligent vision-based car-like vehicle backing systems design,” *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7500–7509, May 2009, doi: 10.1016/j.eswa.2008.09.057.

[9] M. K. Singh, S. Chaube, S. Pant, S. K. Singh, and A. Kumar, “An integrated image visibility graph and topological data analysis for extracting time series features,” *Decis. Anal. J.*, vol. 8, p. 100253, Sep. 2023, doi: 10.1016/j.dajour.2023.100253.

[10] D. Pavlyuk and I. Jackson, “Potential of vision-enhanced floating car data for urban traffic estimation,” *Transp. Res. Procedia*, vol. 62, pp. 366–373, Jan. 2022, doi: 10.1016/j.trpro.2022.02.046.

[11] “The Weather’s Record Keeper,” Meteostat. Accessed: Nov. 10, 2023. [Online]. Available: https://meteostat.net/en/

[12] “Islamic Financial Services Board (IFSB).” Accessed: Nov. 11, 2023. [Online]. Available: https://ifsb.org/

[13] “EPJ Data Science,” SpringerOpen. Accessed: Nov. 11, 2023. [Online]. Available: https://epjdatascience.springeropen.com/

[14] “SAP Help Portal.” Accessed: Nov. 11, 2023. [Online]. Available: https://help.sap.com/docs/

[15] “Research and Reviews - International Journals.” Accessed: Nov. 11, 2023. [Online]. Available: https://www.rroij.com/