Grenze International Journal of Engineering and Technology, June Issue

Artificial Intelligence (AI) to Predict Earth Similarity Index (ESI): An Analysis of Regression Models on ESI Data

Gibin K Jayan[1], Shilpa Aarthi. M[2], Ancy Jenifer. J[3], Golden Nancy. R[4], Joel Mathew[5] and Caleb Stephen[6]
[1-6]Karunya Institute of Technology and Sciences, Coimbatore, India
Email: gibink@karunya.edu.in, shilpaaarthi@karunya.edu, ancyjenifer@karunya.edu, goldennancy@karunya.edu, joelmathew20@karunya.edu.in, calebstephen@karunya.edu.in

***Abstract*— Artificial intelligence has had a significantly smaller impact on space exploration than on other domains. This research analyzes regression models used to predict the Earth Similarity Index (ESI). ESI is an extremely essential statistic for determining exoplanetary habitability. Recognizing the parallels between Earth and exoplanets is critical in determining the habitability of these planets. To capture the complex connections found in ESI data, this study objectively compares the efficiency of various regression models, including Linear Regression, Lasso Regression, Random Forest Regression, Support Vector Regression (SVR), and Huber Regression. These models are utilized to predict ESI, and their performance on this data is assessed to provide a clear picture of AI in predicting ESI. This paper addresses ESI prediction using AI as the initial step in predicting whether exoplanets can potentially support life using AI.**

***Index Terms*— Earth Similarity Index, ESI, exoplanetary habitability, Regression models, Linear Regression, Lasso Regression, Random Forest Regression, Huber Regression, Support Vector Regression, AI**

I. INTRODUCTION

Earth Similarity Index (ESI) has become an important measure in the search for exoplanets that could potentially be habitable and for celestial bodies that have conditions similar to Earth. The ESI was first introduced by Schulze-Makuch et al. (2011) [1] and presents a quantitative way to evaluate how similar exoplanets are to Earth. With the development of astronomical observations producing an ever-growing database of exoplanetary properties, the incorporation of Artificial Intelligence (AI) techniques is becoming essential for effective analysis and prediction. The research delves into the application of artificial intelligence, particularly regression models, to forecast the Earth Similarity Index. The fusion of astrophysics, machine learning, and data science in space exploration has led to substantial advancements in ESI estimation. Utilizing regression models shows promise in predicting ESI by allowing the integration of diverse planetary attributes in the assessment process. The methodology used here aligns with the conclusions made by Schwieterman et al. (2018) [2], who highlighted the value of considering a variety of planet characteristics when evaluating planetary habitability. By combining AI with astronomical data, our study hopes to contribute to the current research on planetary habitability and offer important new information about the possibility of extrasolar life.
The paper's contribution can be summarized as follows:

Grenze ID: 01.GIJET.10.2.258

- A system that can estimate Earth Similarity Index (ESI) if given certain parameters have been developed.
- Various Regression algorithms were used to achieve better results.
- Have done a comparison of all the Regression models used on ESI data.

The remaining content of the paper is organized as follows: A review of relevant research on the use of machine learning for exoplanet discovery using ESI is given in Section 2. Section 3 outlines the process flow and methodologies employed. Section 4 presents the outcomes derived from the application of different regression models. Finally, Section 5 offers concluding remarks.

II. Related Works

This section reviews few papers that have implemented AI techniques in identifying and classifying exoplanets using ESI. In [3], a machine learning (ML) automated approach for identifying and evaluating the likely habitability of transiting exoplanets is presented. The proposed framework integrates various ML algorithms to analyse transit data, demonstrating its efficacy in accurately identifying exoplanets and evaluating their potential habitability, highlighting the crucial role of feature engineering in enhancing model performance. In [4], explore habitability classification of exoplanets using machine learning (ML) techniques. The author presents a methodology employing deep learning to identify exoplanets and assess their potential habitability in [5]. By leveraging advanced neural network architectures, presented approach enables the extraction of intricate patterns from extensive exoplanetary datasets, facilitating accurate detection amidst observational noise. Beyond mere identification, their model incorporates predictive capabilities, offering insights into the likelihood of exoplanetary habitability.

The authors of [6] conduct a preliminary study to investigate the application of machine learning techniques for exoplanet identification. This study looks at how machine learning approaches might improve on conventional detection strategies. Additionally, [7] introduces a deep learning-based anomaly detection method aimed at predicting habitable exoplanets across diverse star systems. This research emphasizes the utilization of deep learning approaches to pinpoint potential habitable exoplanets through the detection of anomalies in observational data.

The writers of [8] employed machine learning methods to examine the habitable zones of planets that are circumbinary. In order to gather insight into the habitable zones surrounding binary star systems and the possible habitability of planets circling these systems, the study applies machine learning methods to the problem. A novel anomaly detection method to infer exoplanetary habitability is proposed in [9]. The study introduces an innovative approach for assessing exoplanet habitability based on anomaly detection techniques, offering insights into identifying potential habitable candidates.

In [10], the authors conduct a study employing analytical and boosted tree methods to theoretically validate potential habitability of recently discovered exoplanets. The research aims to assess the habitability prospects of these exoplanets through rigorous analysis utilizing analytical models and boosted tree algorithms. In [11], machine learning techniques are employed to classify exoplanets as potentially habitable. In order to support the continuous search for habitable planets outside of our solar system, this research focuses on using machine learning algorithms to classify exoplanets based on their potential habitability. The authors of [12] explore the use of machine learning techniques for exoplanet identifying. This study explores the application of machine learning algorithms to improve the effectiveness and accuracy of exoplanet discovery methods.

III. Methodology

Methodologies used mainly include machine learning techniques. The block diagram showing the workflow of the study is shown in Fig. 1

A. Dataset Description

The Habitable Exoplanets Catalog (HWC) is a dataset curated and managed at the University of Puerto Rico at Arecibo (UPR Arecibo)'s Planetary Habitability Laboratory (PHL) [13]. The goal is to provide comprehensive information about exoplanets that are potentially habitable or have characteristics conducive to supporting life as we know it. The HWC dataset offers in-depth data on exoplanets meeting certain criteria linked to habitability, encompassing details like radius, mass, surface temperature, orbital period, and distance from their host star. Moreover, the dataset includes information about the host star, such as spectral type, luminosity, and distance from Earth.

B. Machine Learning Models used

Linear Regression:

Linear regression models the relationship between a dependent variable (ESI) and independent variables (planetary characteristics) assuming a linear association. It calculates coefficients using the least squares method to minimize the differences between observed and predicted values, demonstrating the impact of each independent variable on the dependent variable. While simple and widely used, linear regression struggles with complex nonlinear relationships in data [14] [15]. The linear regression formula is presented in Eq. 1

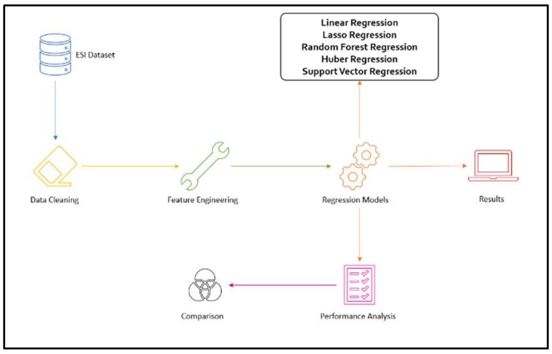

Fig 1. Block Diagram for ESI Prediction and Comparison of Regression Models

$$Z=\gamma_0+\gamma_1 A_1+\gamma_2 A_2+\cdots+\gamma_n A_n+\epsilon \qquad (1)$$

where,

The predicted value is represented by Z

$A_1, A_2, \ldots, A_n$ denotes the predictive variables

$\gamma_0, \gamma_1, \ldots, \gamma_n$ represents the parameters

ϵ represents the residual term.

Lasso Regression:

LASSO (Least Absolute Shrinkage and Selection Operator), a form of linear regression, uses regularization to address overfitting by adding a penalty term to the least squares objective function. By shrinking coefficients towards zero, it encourages sparsity in the model, aiding in feature selection. LASSO automatically identifies key features and eliminates less important ones, proving beneficial for high-dimensional datasets. This technique enhances model interpretability by highlighting essential predictors and preventing overfitting. The formula for lasso regression is given in Eq. 2

$$Z=\beta_0+\beta_1 A_1+\beta_2 A_2+\cdots+\beta_n A_n+\lambda\sum_{i=1}^{n}|\beta_i|+\epsilon \qquad (2)$$

where,

The predicted value is represented by Z

$A_1, A_2, \ldots, A_n$ represents the independent variables

$\beta_0, \beta_1, \ldots, \beta_n$ represents the coefficients

ϵ refers to the error term

λ represents the regularization parameter that regulates the degree of regularization. [16]

Random Forest Regression (RFR):

Decision tree-based regression models like Random Forest Regression is an ensemble learning technique. It creates a finite number of decision trees during the training phase and integrates the predictions made by each to generate the final result. A random subset of the dataset and characteristics are used to train each decision tree in the Random Forest, encouraging variation among the trees. This diversity aids in capturing intricate nonlinear relationships within the data while mitigating the risk of overfitting. Known for its robustness, Random Forest regression excels with large datasets containing noisy or incomplete data. Additionally, it demonstrates resilience to outliers and demands minimal preprocessing of the data [17] [18].

Huber Regression:

Huber regression is a robust regression method designed to address the influence of outliers in a dataset by combining features of least squares regression and absolute deviation regression. Unlike conventional techniques that focus only on squared errors or absolute errors, Huber regression incorporates a mix of both to minimize the total sum of squared errors for data points close to the prediction and the sum of absolute errors for those farther away. This unique approach allows Huber regression to achieve a balance between handling outliers effectively and maintaining strong overall model performance. It is particularly useful when the dataset contains outliers that can significantly affect the estimation of regression coefficients [19] [20]. Huber regression formula is shown in Eq. 3

$$L_\delta (Z,f(A))=\sum_{i=1}^{n} \left(\frac{1}{2} (Z_i-f(A_i))\right)^2 \cdot I(|Z_i-f(A_i)|\le\delta)+\delta\cdot|Z_i-f(A_i)|\cdot I(|Z_i-f(A_i)|>\delta) \quad (3)$$

where,

The observed value is represented by Z

$f(A)$ represents the predicted value

δ represents the threshold parameter

$I(\cdot)$ indicates the indicator function

Support Vector Regression (SVR)

SVR is a regression technique based on SVM, using a kernel function to handle nonlinear relationships by transforming the input space. It finds a hyperplane in the transformed space that maximizes the margin with training data. SVR balances model complexity and accuracy using a regularization parameter, making it useful for high-dimensional datasets and complex nonlinear patterns. SVR is particularly effective at capturing subtle patterns that linear models may miss, making it a valuable tool for handling intricate nonlinear relationships in high-dimensional datasets [21] [22].

C. Evaluation metrics

Mean Squared Error (MSE)

Mean Squared Error (MSE) quantifies the average squared difference between predicted and actual values, providing a comprehensive assessment of a regression model's predictive accuracy. By squaring the errors, MSE emphasizes larger discrepancies, making it responsive to outliers or significant deviations from the true values. Decreasing MSE values suggest smaller discrepancies between predicted and observed values, signifying higher model accuracy. Enhancing prediction accuracy involves refining models to reduce the mean square error (MSE) as a measure of performance. The equation for MSE is provided in Eq. 4

$$MSE=\frac{1}{m} \sum_{i=1}^{n} (z_i-\hat{z}_i)^2 \quad (4)$$

Such that,

m represents the total number of data points

z_i denotes the primary or witnessed value of the response variable

$\hat{z}_i$ signifies the estimated or forecasted value of the response variable

Root Mean Squared Error (RMSE)

RMSE, or Root Mean Squared Error, is the square root of MSE (Mean Squared Error), offering a measure of the average error magnitude in units identical to the target variable. It holds significant importance in regression analysis as a crucial performance metric. Lower RMSE values signify enhanced model effectiveness, indicating reduced disparities between predicted and actual values. When comparing models, RMSE is used as a standard. Reducing RMSE as much as possible while refining a model would increase predicted accuracy. Despite Regression model performance can be assessed and compared using RMSE because of its vulnerability to outliers and its direct interpretation in the original units of the target variable. The formula for RMSE computation is provided in Eq. 5

$$RMSE=\sqrt{MSE} \quad (5)$$

R-Squared (R^2)

R^2, a crucial metric in regression analysis, quantifies the extent to which the independent variable(s) in the model account for the variance observed in the dependent variable. Ranging from 0 to 1, where 1 denotes a perfect fit capturing all variance, higher R^2 values signify enhanced model efficacy. This metric aids in assessing the model's capability to account for variability in the data, providing a clear gauge of its goodness of fit. While invaluable, it's imperative to interpret R^2 alongside other metrics, as its value may artificially inflate with the addition of irrelevant predictors. Hence, validation through cross-validation techniques is essential to ensure robust model evaluation. The formula for R^2 is illustrated in Eq. 6

$$R^2=1-\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \qquad (6)$$

where,

The number of observations is represented by n
y_i denotes the primary or witnessed value of the response variable
$\bar{y}$ denotes the mean of the primary or witnessed value of the response variable
$\hat{y}_i$ denotes the estimated or forecasted value of the response variable.

Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is a reliable metric in regression analysis that evaluates the average discrepancies between predicted and observed values. It provides a simple assessment of model accuracy by treating all errors equally. MAE is less influenced by outliers compared to Mean Squared Error (MSE), making it suitable for datasets with extreme values. Lower MAE values indicate better model performance with smaller differences between predicted and actual values. MAE, along with other metrics like MSE and R-squared, enhances the comprehensive evaluation of model accuracy, aiding in informed decision-making during regression analysis. The equation for calculating MAE is depicted in Eq. 7

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|z_i-\hat{z}_i| \qquad (7)$$

where,

n denotes the sum total of the data points present
z_i denotes the primary or witnessed value of the response variable.
$\hat{z}_i$ refers to the estimated or forecasted value of the response variable

[23]

IV. Results

A. Regression Models on ESI Data

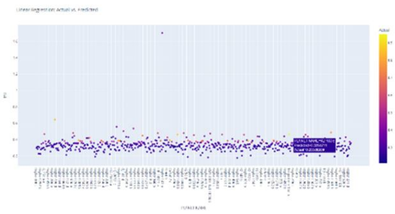

Fig. 2 Linear Regression: Actual ESI vs Predicted ESI values

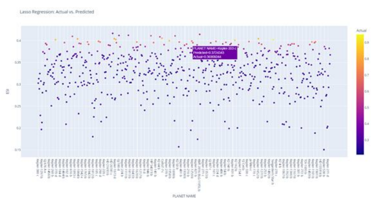

Fig. 3 Lasso Regression: Actual ESI vs Predicted ESI values

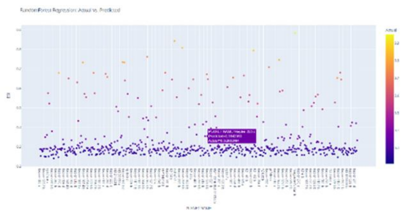

Fig. 4 Random Forest Regression: Actual ESI vs Predicted ESI values

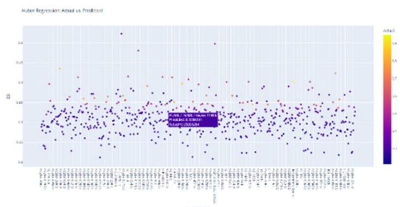

Fig. 5 Huber Regression: Actual ESI vs Predicted ESI values

Fig. 6 Support Vector Regression: Actual ESI vs Predicted ESI values

Fig. 2, Fig. 3, Fig. 4, Fig. 5, and Fig. 6 showcase scatter plots illustrating the correspondence between actual Earth Similarity Index (ESI) values and predicted ESI values derived from the Linear Regression model, Lasso Regression model, Random Forest Regression model, Huber Regression model, and Support Vector Regression model, respectively. Each data point on the plot represents an observation, enabling visual scrutiny of these models' efficacy in accurately predicting ESI values across the dataset.

B. Regression Models' Performance

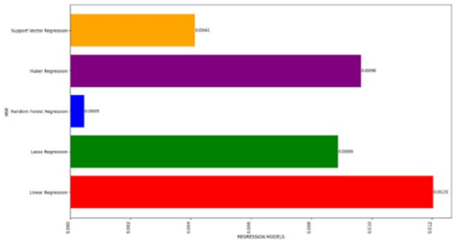

Fig. 7 Comparison of Mean Squared Error (MSE) of all Regression models

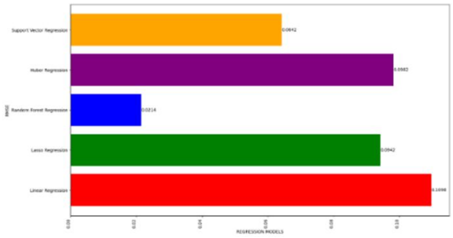

Fig. 8 Comparison of Root Mean Squared Error (RMSE) of all Regression models

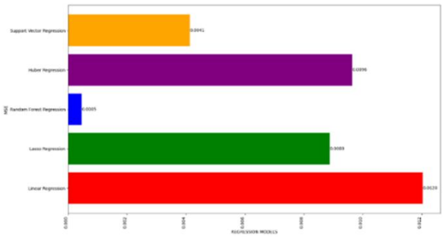

Fig. 9 Comparison of Mean Absolute Error (MAE) of all Regression models

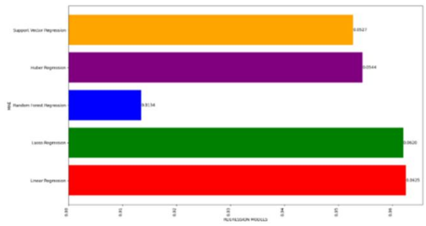

Fig. 10 Comparison of R-squared (R^2) score of all Regression models

Each model's efficacy was measured using key metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) score. The findings revealed significant differences in the predictive accuracy of the models. Random Forest Regression emerged as the top-performing model, boasting the lowest MSE of 0.0005 (refer to Fig. 7), indicating minimal overall variance between predicted and actual ESI values. It was succeeded by Support Vector Regression, Lasso Regression, Huber Regression, and Linear Regression, each displaying relatively low MSE values of 0.0041, 0.0089, 0.0096, and 0.0120, respectively. Furthermore, analysis of Root Mean Squared Error (RMSE) shed light on the mean error dimension in predictions. Random Forest Regression exhibited the lowest RMSE of 0.0214 (see Fig. 8), indicating the most accurate predictions with smaller deviations from actual ESI values. Following suit were Support Vector Regression, Lasso Regression, Huber Regression, and Linear Regression, with RMSE values of 0.0642, 0.0912, 0.0982, and 0.1090, respectively.

The evaluation of Mean Absolute Error (MAE) identified Random Forest Regression as the top performer, achieving the smallest MAE of 0.0134, indicating the lowest average absolute error. This was followed by Support Vector Regression, Huber Regression, Lasso Regression, and Linear Regression. Additionally, R-squared (R^2) scores revealed the amount of variance in ESI explained by the regression models. Random Forest Regression stood out with an impressive R^2 score of 0.9645, explaining 96.45% of the variance in ESI. Support Vector Regression also performed well with an R^2 score of 0.6801, followed by Lasso Regression (0.3115), Huber Regression (0.2519), and Linear Regression (0.0661). These results highlight the effectiveness of artificial intelligence methods, especially Random Forest Regression and Support Vector Regression, in predicting ESI based on planetary attributes.

V. Conclusion

Our research on different regression models' capabilities in estimating the Earth Similarity Index (ESI) based on planetary characteristics has provided valuable insights into the potential application of artificial intelligence (AI) in astrobiology studies. The analysis involved key metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) score, offering a comprehensive evaluation of model performance. Random Forest Regression emerged as the best performer, showing superior results across all metrics with the lowest MSE, RMSE, and MAE values, along with the highest R^2 score. These results emphasize its precision in predicting ESI values and capturing a significant portion of the dataset's variability.

Our study marks a significant milestone in integrating AI technologies into space exploration by using regression models to delve into the cosmos and identify habitable planets. While our focus is on regression modeling, we recognize the potential of classification models and image analysis to enhance accuracy. By analyzing features such as water bodies and vegetation, we can extract data to pinpoint habitable exoplanets. The incorporation of machine learning enables automated habitat detection on distant celestial bodies. As we push the boundaries of astrobiology research, leveraging cutting-edge AI can accelerate the quest for habitable planets and influence the future of space exploration.

REFERENCES

[1] Schulze-Makuch, D., Méndez, A., Fairén, A.G., Von Paris, P., Turse, C., Boyer, G., Davila, A.F., António, M.R.D.S., Catling, D. and Irwin, L.N., 2011. A two-tiered approach to assessing the habitability of exoplanets. Astrobiology, 11(10), pp.1041-1052.

[2] Schwieterman, E.W., Kiang, N.Y., Parenteau, M.N., Harman, C.E., DasSarma, S., Fisher, T.M., Arney, G.N., Hartnett, H.E., Reinhard, C.T., Olson, S.L. and Meadows, V.S., 2018. Exoplanet biosignatures: a review of remotely detectable signs of life. Astrobiology, 18(6), pp.663-708.

[3] Pratyush, P. and Gangrade, A., 2021. Automation of Transiting Exoplanet Detection, Identification and Habitability Assessment Using Machine Learning Approaches. arXiv preprint arXiv:2112.03298.

[4] Basak, S., Mathur, A., Theophilus, A.J., Deshpande, G. and Murthy, J., 2021. Habitability classification of exoplanets: a machine learning insight. The European Physical Journal Special Topics, 230, pp.2221-2251

[5] Mathur, S., Sizon, S. and Goel, N., 2021. Identifying exoplanets using deep learning and predicting their likelihood of habitability. In Advances in Machine Learning and Computational Intelligence: Proceedings of ICMLCI 2019 (pp. 369-379). Springer Singapore.

[6] Jin, Y., Yang, L. and Chiang, C.E., 2022. Identifying exoplanets with machine learning methods: a preliminary study. arXiv preprint arXiv:2204.00721.

[7] Patel, Y., Tiwari, S., Sonbhadra, S.K. and Agarwal, S., 2023, June. Predicting Habitable Exoplanets in Different Star-Systems Using Deep Learning Based Anomaly Detection Approach. In 2023 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.

[8] Kong, Z., Jiang, J.H., Burn, R., Fahy, K.A. and Zhu, Z.H., 2022. Analyzing the Habitable Zones of Circumbinary Planets Using Machine Learning. The Astrophysical Journal, 929(2), p.187.

[9] Sarkar, J., Bhatia, K., Saha, S., Safonova, M. and Sarkar, S., 2022. Postulating exoplanetary habitability via a novel anomaly detection method. Monthly Notices of the Royal Astronomical Society, 510(4), pp.6022-6032.

[10] Saha, S., Basak, S., Safonova, M., Bora, K., Agrawal, S., Sarkar, P. and Murthy, J., 2018. Theoretical validation of potential habitability via analytical and boosted tree methods: An optimistic study on recently discovered exoplanets. Astronomy and computing, 23, pp.141-150.

[11] Hora, K., 2018. Classifying exoplanets as potentially habitable using machine learning. In ICT Based Innovations: Proceedings of CSI 2015 (pp. 203-212). Springer Singapore.

[12] Malik, A., Moster, B.P. and Obermeier, C., 2022. Exoplanet detection using machine learning. Monthly Notices of the Royal Astronomical Society, 513(4), pp.5505-5516.

[13] https://phl.upr.edu/projects/habitable-exoplanets-catalog (Arecibo, 2023)

[14] James, G., Witten, D., Hastie, T., Tibshirani, R. and Taylor, J., 2023. Linear regression. In An introduction to statistical learning: With applications in python (pp. 69-134). Cham: Springer International Publishing.

[15] Montgomery, D.C., Peck, E.A. and Vining, G.G., 2021. Introduction to linear regression analysis. John Wiley & Sons.

[16] Ranstam, J. and Cook, J.A., 2018. LASSO regression. Journal of British Surgery, 105(10), pp.1348-1348.

[17] Segal, M.R., 2004. Machine learning benchmarks and random forest regression.

[18] Jain, N. and Jana, P.K., 2023. LRF: A logically randomized forest algorithm for classification and regression problems. Expert Systems with Applications, 213, p.119225.

[19] Sun, Q., Zhou, W.X. and Fan, J., 2020. Adaptive huber regression. Journal of the American Statistical Association, 115(529), pp.254-265.

[20] Feng, Y. and Wu, Q., 2022. A statistical learning assessment of Huber regression. Journal of Approximation Theory, 273, p.105660.

[21] Zhang, F. and O'Donnell, L.J., 2020. Support vector regression. In Machine learning (pp. 123-140). Academic Press.

[22] Awad, M., Khanna, R., Awad, M. and Khanna, R., 2015. Support vector regression. Efficient learning machines: Theories, concepts, and applications for engineers and system designers, pp.67-80.

[23] Tatachar, A.V., 2021. Comparative Assessment of Regression Models Based On Model Evaluation Metrics. International Journal of Innovative Technology and Exploring Engineering, 8(9), pp.853-860.