International Journal of Engineering Research in Current Trends (IJERCT)
ISSN: 2582-5488, Volume-6, Issue-4, August 2024

Enhancing Soiling Determination for Parabolic Trough Collectors: A Machine Learning Approach

[1]Rahul Singh,[2] Dr Rajeev Arya,[3] Arun Kumar Patel, [4] Rohit Sijariya
[1]M.Tech Scholar,[2]Director,[3]Assistant Professor, [4]Assistant Professor
Vidhyapeeth Institute of Science and Technology, Bhopal, India
[1]rahulsinghchauhan8097.com, [2]directorvidhyapeeth@gmail.com , [3]patel.arun2020@gmail.com,
[4]sijariya.rohit@gmail.com,

Abstract— **This study evaluates the predictive capabilities of several regression models, including Support Vector Regression (SVR), linear regression, Gaussian Process Regression (GPR), and an optimized Decision Tree (DT) model, in estimating a specific property, Gloss. Through a methodical examination involving both visual and statistical analyses, the models' performances were compared to ascertain their efficacy and precision in capturing the variability of Gloss values.The SVR and linear regression models showed limited ability in predicting Gloss, with R-squared values of 0.50 and 0.33, respectively, indicating a moderate to substantial amount of unexplained variance. The GPR model marked an improvement with an R-squared value of 0.69, suggesting a better but incomplete fit to the data. Notably, the optimized DT model demonstrated superior predictive performance, evidenced by a closely matched density plot and scatter plot, along with an R-squared value of 0.69, indicating robustness in explaining the data's variability. The results reveal that while each model has inherent strengths, the optimized Decision Tree model provides the most balanced approach in terms of accuracy and reliability. This study underscores the importance of model selection and optimization in enhancing predictive accuracy, offering insights into potential advancements in regression techniques for similar applications.**

Index Terms— Regression Analysis, Predictive Modeling, Decision Tree Optimization, Gaussian Process Regression, Model Evaluation Metrics

---

I. INTRODUCTION

Parabolic trough collectors (PTCs) are a prevalent form of solar thermal energy technology, particularly suited for large-scale solar power plants. However, the accumulation of soil on the collector surface, commonly referred to as soiling, can obstruct sunlight and reduce the efficiency of energy capture. Traditional methods for monitoring soiling involve manual inspections and periodic cleaning schedules, which are often inefficient and can lead to unnecessary operational interruptions. In recent years, machine learning (ML) has emerged as a powerful tool for predictive maintenance and operational optimization in various engineering fields, including solar energy. This research explores the application of ML techniques to predict and quantify soiling in PTCs, aiming to enhance the efficiency and reduce the maintenance costs of solar plants. The adaptability of PTCs to various environmental settings and their scalability makes them a preferred choice for large-scale solar installations worldwide. Notable examples include the Solana Generating Station in the USA and the Andasol solar power station in Spain, which demonstrate the potential of PTCs to contribute significantly to energy grids while mitigating carbon emissions (Turchi et al., 2010). The operational efficiency of these systems, however, is critically dependent on maintaining high levels of solar absorption, which can be significantly impaired by soiling**.**

1.1 Soiling problem and its impact on the efficiency of solar collectors

Soiling is a significant challenge in the field of solar energy, particularly impacting the efficiency of solar collectors, including photovoltaic panels and concentrated solar power systems such as parabolic trough collectors. Soiling occurs when dust, pollen, bird droppings, and other particulates accumulate on the surface of solar collectors, forming a layer that obstructs sunlight from reaching the solar cells or thermal absorptive surfaces. This obstruction results in a direct decrease in the amount of solar energy captured and converted into electricity or heat, leading to reduced overall efficiency.ggwq1xss The impact of soiling on solar collector efficiency can vary greatly depending on geographical location, local environmental conditions, and the specific type of solar technology in use. In arid and semi-arid regions, where dust and sand are prevalent, soiling can lead to significant efficiency losses, sometimes reducing the solar output by up to 30% if not managed properly. Even in less dusty environments, seasonal pollen and urban pollutants can accumulate on collector surfaces, necessitating regular cleaning to maintain optimal performance.

The economic implications of soiling are also considerable, as reduced efficiency translates directly into lower electricity production and potential revenue losses for solar power plants. The need to clean solar installations frequently to mitigate soiling effects introduces additional operational costs and challenges, particularly in large-scale solar farms where manual cleaning is impractical. These

factors make soiling not only a technical issue but also a significant economic concern for the viability and sustainability of solar energy projects. The frequency and method of cleaning required to combat soiling can have environmental impacts, particularly in terms of water usage and the potential ecological effects of runoff from cleaning agents. This has led to increased interest in developing more efficient cleaning technologies, such as robotic cleaners and water-free cleaning systems, and innovations like hydrophobic and self-cleaning coatings that reduce the rate of soiling or make the cleaning process less labour-intensive and more sustainable.

Research by Ilse et al. (2018) provides detailed analysis on how soiling specifically affects solar panels by decreasing their transmittance, thus reducing the amount of sunlight that reaches the active layers of the panels. Their findings indicate that even a thin layer of dust can lead to efficiency losses of up to 5% per day under certain conditions, underscoring the critical need for effective cleaning and maintenance strategies (Ilse et al., 2018).A significant focus has been placed on developing more efficient and less resource-intensive cleaning methods. Biel et al. (2019) explored automated robotic systems that use minimal water and energy to clean solar panels. These systems can be programmed to operate during non-peak hours, thereby ensuring that cleaning processes do not interfere with energy production (Biel et al., 2019).The specific environmental conditions of a site greatly influence the rate and type of soiling. A study by Kaldellis et al. (2017) examines the soiling rates across different geographical locations and highlights those areas with higher wind speeds and airborne particulate matter experience more rapid accumulation of soil on solar collectors. This study emphasizes the importance of site-specific assessments in planning and operating solar power the economic impact of soiling is crucial for the financial planning of solar projects.

1.2 The use of machine learning as a solution to predict and manage soiling

The accumulation of dirt, dust, and other particulate matter on solar panels, commonly known as soiling, significantly diminishes the efficiency and effectiveness of solar energy systems. Traditional methods for managing soiling have relied primarily on scheduled cleanings, which can be both costly and inefficient, often leading to unnecessary maintenance or insufficient response to soiling, thereby reducing the energy output. In this context, machine learning (ML) offers a transformative solution by enabling more precise and predictive management of soiling in solar energy installations. Machine learning, a subset of artificial intelligence (AI), involves the use of algorithms and statistical models to enable computers to perform tasks without explicit instructions, instead relying on patterns and inference. In the application to soiling of solar panels, ML algorithms are trained on historical data—including rates of soiling, weather conditions, environmental data, and the performance metrics of solar panels. This training allows the models to predict future soiling events and their potential impact on solar energy generation. ML models can predict when the accumulation of soiling will reach a threshold that justifies cleaning. This predictive capability ensures that cleaning occurs only, when necessary, thereby optimizing maintenance resources and reducing costs. By integrating real-time data from environmental sensors and solar energy output, ML models can dynamically adjust predictions and maintenance schedules based on actual conditions. This adaptability is particularly useful in areas with high variability in dust exposure or weather conditions.

Several studies and implementations have demonstrated the efficacy of machine learning in managing solar panel soiling. For instance, researchers have developed models that use inputs from local weather stations to forecast dust accumulation, while others have employed image recognition technologies to detect soiling levels directly from panel images. These technologies not only streamline operations but also contribute to the sustainability of solar power by enhancing the efficiency of solar energy collection systems. The integration of machine learning into the management of solar panel soiling represents a significant advancement in solar technology, promising to enhance both the economic and environmental benefits of solar energy. Machine learning (ML) continues to redefine the management of soiling in solar energy systems by offering sophisticated tools for prediction, monitoring, and optimization. The application of ML technologies in this domain leverages vast amounts of data, including environmental conditions, panel efficiency, and maintenance history, to significantly enhance the operation of solar power installations. Machine learning models are particularly adept at identifying complex patterns in data that would be indiscernible to traditional analytical methods. For instance, a study utilized a combination of convolutional neural networks (CNNs) and weather data to forecast soiling rates on solar panels. This method proved highly effective, predicting soiling impacts with a high degree of accuracy, which in turn facilitated optimized cleaning schedules and reduced downtime.

Environmental considerations are also central to the adoption of machine learning in this context. A study focuses on the environmental benefits of using ML to optimize cleaning schedules, particularly the significant reduction in water usage. This is crucial in drought-prone areas where water conservation is a priority. The study shows that machine learning not only supports the operational and financial aspects of solar energy but also contributes to its environmental sustainability by minimizing the resource footprint of maintaining solar panels.

II. LITERATURE REVIEW

Integration with Other Renewable Technologies, parabolic troughs are increasingly being integrated with other renewable energy technologies to create hybrid systems that can provide more stable and diversified energy outputs. For instance: Combined Heat and Power (CHP), parabolic trough systems can be integrated into CHP systems to provide both electricity and heat for industrial processes. This integration not only increases overall energy efficiency but also maximizes the utilization of the collected solar energy. Hybrid Solar and Biomass Plants, combining solar thermal energy with biomass energy can ensure continuous power supply, particularly in regions where sunlight may be intermittent but biomass resources are abundant. This hybrid

approach can enhance reliability and reduce dependency on fossil fuels.

Kai Wang, Z. Qin, W. Tong, C. Ji (2020), this research emphasizes the critical role of thermal energy storage systems in stabilizing and increasing the efficiency of solar energy systems. By exploring a range of storage technologies, the study categorizes these into sensible heat storage, latent heat storage, and thermo chemical storage. Sensible heat storage involves materials that store energy through temperature changes, common in water and rock beds. Latent heat storage uses phase change materials that absorb or release heat at constant temperatures. Thermo chemical storage, the most advanced type, involves chemical reactions that store and release heat. These technologies are essential for ensuring a continuous energy supply in various sectors, including residential, industrial, and power generation, despite the intermittent nature of solar energy. The research particularly highlights how material innovations and system integration strategies can significantly advance the performance and economic viability of solar thermal systems.

Yuan Tian, Changying Zhao (2013), the latest developments in solar collectors and thermal energy storage for solar thermal applications. It covers both non-concentrating and concentrating solar collectors, analyzing their optical optimization, heat loss reduction, and the enhancement of heat recuperation. An extensive analysis of the latest advancements in solar thermal collectors and associated thermal energy storage technologies. It covers the spectrum from non-concentrating to concentrating solar collectors, assessing their efficiency in terms of optical optimization, which enhances the focus of solar radiation, and heat loss reduction, crucial for maintaining high temperatures. The study also discusses the enhancement of heat recuperation, which involves reclaiming waste heat for further use, thus increasing overall system efficiency. These improvements are vital for applications ranging from residential heating to industrial processes, highlighting the evolving capabilities of solar thermal technology in harnessing solar energy more effectively.

B. Stutz, N. L. Pierrès, F. Kuznik, K. Johannes, E. P. D. Barrio, (2017), various types of solar thermal energy storage solutions, focusing on both low (40-120°C) and medium-to-high-temperature (120-1000°C) applications. The study highlights the different storage methods including sensible heat, latent heat, and thermo chemical storage, discussing their suitability for building applications to concentrating solar power plants. The effectiveness of sensible, latent, and thermo chemical storage methods in these applications, discussing how each method suits different needs based on their operational temperature and energy storage density. The study emphasizes the importance of selecting appropriate materials and integrating them effectively to maximize the performance and sustainability of solar thermal systems.

L. Aye, Amitha Jayalath (2018), the integration of solar thermal technologies in the built environment, highlighting applications such as electricity generation, hot water production, product drying, cooking, clean water production, space heating, cooling, and refrigeration. The fundamental principles and recent developments of these applications, presenting a well-rounded view of the potential of solar thermal technologies to meet a variety of energy needs in residential and commercial settings. These advancements have made solar thermal technologies an integral part of building design and operation, illustrating their potential to meet a wide range of energy needs in both residential and commercial settings effectively and sustainably.

Mouaky et al. (2019), soiling is a significant factor affecting the efficiency of parabolic trough collectors (PTCs), particularly in semi-arid regions. dust accumulation can lead to a substantial decrease in the thermal output of solar fields, with reductions up to 27%. The accumulation of dust and debris on the collector surfaces impedes sunlight absorption, crucial for the heating process in PTCs. This reduction in efficiency necessitates frequent cleaning to maintain optimal performance, which can increase operational costs. The study underscores the need for effective soiling management strategies to mitigate efficiency losses and ensure the sustainable operation of solar thermal plants in dust-prone areas. This finding is pivotal for regions where solar power is a significant part of the energy mix, emphasizing the environmental challenges specific to these locales.

III. OBJECTIVE OF STUDY

Evaluation and Compare Regression Models for Gloss Prediction:

To systematically assess the predictive performances of different regression models including Support Vector Regression (SVR), Linear Regression, Gaussian Process Regression (GPR), and an optimized Decision Tree (DT) in predicting the glossiness of surfaces.

Optimization of Decision Tree Model for Enhanced Predictive Accuracy:

To refine and optimize the Decision Tree model for predicting Gloss, aiming to maximize its predictive accuracy as indicated by statistical metrics such as R-squared, MSE, and RMSE.

Explore Advanced Techniques and Model Enhancements:

To investigate further enhancements and advanced modeling techniques that could improve the predictive performance of the Decision Tree model and other studied models.

IV METHODOLOGY

4.1 Model Selection: To assess and compare the predictive accuracy of various regression models for the variable Gloss, four models were selected based on their distinct characteristics and suitability for non-linear and linear relationships. These models included Support Vector Regression (SVR), linear regression, Gaussian Process Regression (GPR), and an optimized Decision Tree (DT) model. Each model was chosen to explore different aspects of regression analysis, from basic linear approaches to more

complex, non-linear methods that might capture more intricate patterns in the data.

4.2 Data Collection: The dataset used in this study consisted of measurements of the Gloss property from a specific set of samples under controlled conditions. This dataset included both the Gloss measurements (target variable) and various features that were hypothesized to influence Gloss, such as chemical composition, surface texture, and environmental conditions during measurement.

4.3 Data Preparation: Prior to model training, the data underwent several preprocessing steps to ensure optimal model performance:

Cleaning: The dataset was cleaned to remove any outliers or erroneous data points that could skew the results. This involved checking for and handling missing values, and filtering out data points that fell outside plausible value ranges for Gloss.

Normalization: Features were normalized to ensure that no single feature dominated the predictions due to scale differences. This was crucial for models like SVR and linear regression, where feature scaling can significantly impact performance.

Splitting: The dataset was randomly split into a training set and a test set, with approximately 70% of the data allocated for training and 30% reserved for testing. This split was used to train the models and then evaluate their predictive accuracy on unseen data.

4.4 Model Training: Each model was trained using the training dataset. For SVR and the Decision Tree, parameter tuning was performed using grid search with cross-validation to find the optimal settings that minimized prediction error. Linear regression was applied directly without tuning, given its simplicity. GPR required setting the kernel type and parameters, which were chosen based on preliminary tests indicating their effectiveness for this type of data.

4.5 Model Evaluation Criteria: To compare the models effectively, several statistical metrics were used:

Mean Squared Error (MSE): Measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value.

Root Mean Squared Error (RMSE): The square root of MSE, providing a measure of the average error in the same units as the Gloss values.

Mean Absolute Error (MAE): Represents the average absolute difference between predicted and actual values, providing a straightforward measure of error magnitude.

R-squared (R^2): Indicates the percentage of the response variable variation that is explained by a linear model.

4.6 Analysis of Results: Following the training and testing phases, an in-depth analysis was conducted to interpret the results from each model. This analysis was two-fold, involving both visual and quantitative assessments:

Visual Representations: Visualizations were crucial for understanding model performance in a more intuitive manner. Density plots and scatter plots were generated for each model:

Density Plots were used to compare the distribution of actual and predicted Gloss values, providing insights into how well each model captured the overall distribution.

Scatter Plots plotted actual Gloss values against predicted values, highlighting the correlation and identifying any systematic errors such as bias towards overestimation or underestimation.

Statistical Analysis: The calculated MSE, RMSE, MAE, and R-squared values for each model were compared. This comparison helped in assessing which models provided the most accurate predictions, maintained consistency in performance across different data segments, and effectively captured the variability in Gloss measurements.

4.7 Model Optimization and Tuning: Particular attention was given to the optimized Decision Tree model. Techniques such as pruning, setting maximum depth, and minimizing leaf size were explored to enhance model performance. Parameter tuning was guided by the results of cross-validation to avoid overfitting and to generalize better on unseen data.

4.8 Model Comparison and Selection: Each model's strengths and weaknesses were evaluated based on the analytical results. Decision factors included:

Accuracy: How close the predictions were to actual values.

Robustness: Consistency of model performance across different data sets.

Computational Efficiency: Time and resources required to train and predict using the model.

Interpretability: Ease of understanding and explaining the model predictions.

V. RESULTS AND DISCUSSIONS

5.1 Graphical Representations:

Density Plot:

This plot illustrates the distribution of both actual and predicted Gloss values.

A red line denotes the actual values, while a blue line indicates the predictions derived from the SVR model.

The overlap between these distributions suggests a general alignment of the model with the actual data trends, albeit with deviations in certain areas, particularly at peak values.

Notably, at a Gloss value around 100, the predictions sharply peak more than the actual values, implying a potential overestimation by the model at this point.

Scatter Plot:

This plot compares the actual Gloss values on the x-axis with the predicted values on the y-axis.

A dashed line indicates the line of perfect prediction, where ideal predictions would exactly match the actual values.

Although the data points primarily cluster around this line, deviations are more pronounced at higher Gloss values, indicating reduced prediction accuracy in these regions.

Despite these discrepancies, a positive correlation remains evident, suggesting a generally effective model performance.

Statistical Results:

Mean Squared Error (MSE): 13.73: This metric represents the average of the squared differences between the actual and predicted values. An MSE of 13.73 indicates an average squared deviation of approximately 13.73 for Gloss predictions.

Root Mean Squared Error (RMSE): 3.70: This is the square root of the MSE, reflecting errors in the same units as Gloss.

It indicates an average prediction error of about 3.70 units, highlighting sensitivity to larger prediction errors.
Mean Absolute Error (MAE): 2.31: Unlike MSE or RMSE, MAE offers an average of the absolute differences between the predicted and actual values, providing a direct measure of average error, which is approximately 2.31 units for Gloss.
R-squared (Coefficient of Determination): 0.50: This statistic quantifies the proportion of variance in the actual Gloss values explained by the model, with an R-squared of 0.50 suggesting that the model explains about 50% of the variability, indicating moderate explanatory power.
The SVR model exhibits moderate predictive accuracy for the Gloss data. The graphical representations, specifically the density plot, depict the model's capability to approximate the trend of actual values to a certain extent. Meanwhile, the scatter plot and R-squared value highlight significant areas for potential enhancement. Both the RMSE and MAE suggest that prediction errors are within a few units, which may be either acceptable or problematic, depending on the specific requirements and tolerance for error in measuring "Gloss". The model's particular struggle with higher values, as evidenced by the dispersion in the scatter plot, may necessitate more advanced modeling approaches or additional feature engineering to elevate prediction accuracy across all Gloss value ranges.

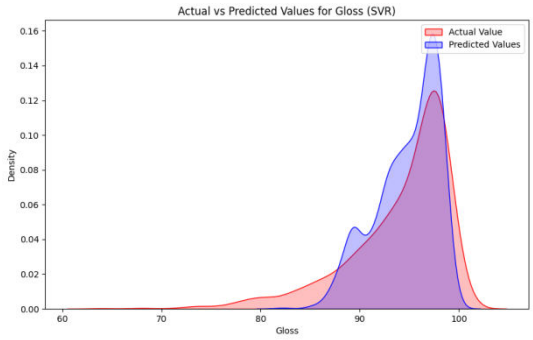

Figure 1 Actual vs Predicted values for gloss (SVR)

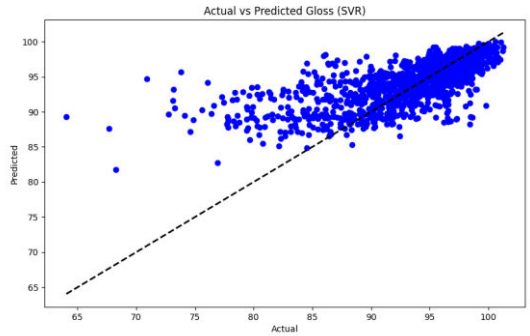

Figure 2 Actual vs predicted gloss (SVR)

5.2 Linear Regressor Analysis

Graphical Representation:
Density Plot: The density plot delineates the distributions of both actual (represented by a red line) and predicted (indicated by a blue line) Gloss values. Notably, both distributions exhibit a pronounced peak, suggesting a high concentration of Gloss values within this specific range. However, the plot also reveals divergences between predicted and actual values around this peak, indicating that the linear regression model might slightly overestimate the most frequent Gloss values, thereby not fully capturing the true distribution.
Scatter Plot: In the scatter plot, actual Gloss values plotted on the x-axis are compared with the predicted values on the y-axis. The points disperse around a dashed line, which signifies the line of perfect prediction, showing an overall positive trend. This trend affirms the model's ability to capture the directional relationship between the input features and Gloss values, although the spread around the line indicates varying degrees of predictive accuracy across different Gloss values.
Statistical Results:
Mean Squared Error (MSE) - 18.42: The MSE, reflecting the average squared differences between the actual and predicted values, stands at 18.42. This value suggests a significant magnitude of errors, which might be substantial depending on the Gloss values' scale and contextual interpretation.
Root Mean Squared Error (RMSE) - 4.29: The RMSE provides insight into the average magnitude of the errors, measured in the same units as Gloss. With an RMSE of 4.29, the model typically deviates from actual values by approximately 4.29 Gloss units, illustrating the average error size.
Mean Absolute Error (MAE) - 3.08: With an MAE of 3.08, the model, on average, misses the actual Gloss value by about 3.08 units. This metric offers a more direct measure of error, less influenced by the larger errors unlike MSE and RMSE.
R-squared (Coefficient of Determination) - 0.33: The R-squared value at 0.33 indicates that only 33% of the variance in Gloss is explained by the model, signaling limited predictive power and significant variance that the model fails to account for.
The linear regression model demonstrates a foundational ability to predict Gloss values, albeit with limited precision. The moderate R-squared value, together with the higher errors as indicated by MSE and RMSE, suggests that the model may lack the necessary complexity to thoroughly capture the underlying relationships between predictors and the Gloss target variable, or potentially, other influential factors impacting Gloss are not included in the model. Despite recognizing a discernible relationship, the visualizations and statistical outcomes collectively underscore substantial opportunities for enhancing the model's predictive performance.

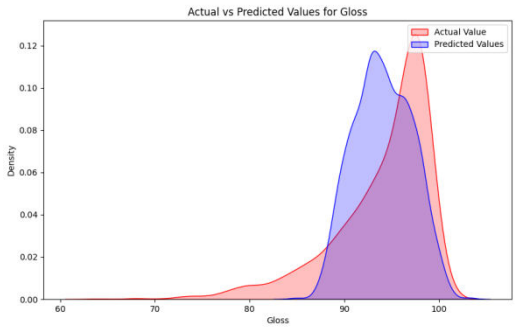

Figure 3 Actual vs Predicted values for gloss (LR)

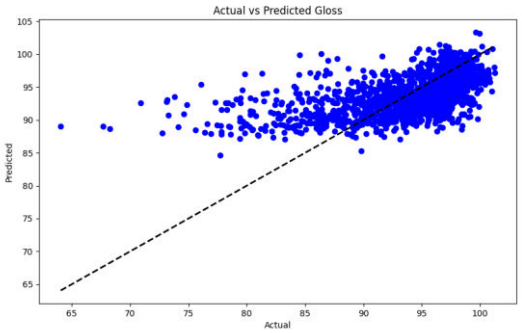

Figure 4 Actual vs Predicted values for gloss (LR)

5.3 Gaussian Process Regression

Visual Representation

Scatter Plot of Actual vs. Predicted Values: The scatter plot elucidates the relationship between the actual Gloss values plotted on the x-axis and the predicted Gloss values by the GPR model on the y-axis. The dashed diagonal line symbolizes the line of perfect prediction, where predicted values would align exactly with the actual values. The clustering of points near this line denotes that the GPR model's predictions closely match the actual values for most data points. A slight trend is observed where points fall above the line at higher Gloss values, possibly indicating a tendency of the model to under-predict in these regions.

Histogram of Residuals: The histogram illustrates the distribution of residuals—the differences between the predicted and actual Gloss values. With residuals centered around zero, there appears to be no strong systematic bias in the predictions, either over- or under-predicting. The distribution's approximate normality is a positive indicator of model fit, though the presence of tails, particularly a longer tail for positive residuals, suggests some predictions substantially underestimate the actual values.

Statistical Results:

Mean Squared Error (MSE) - 8.37: This metric, which measures the average of the squared differences between actual and predicted values, stands at 8.37, indicating that the squared errors from the model are relatively low, a favorable outcome.

Root Mean Squared Error (RMSE) - 2.89: The RMSE, the square root of the MSE, conveys the average error magnitude in the same units as Gloss. An RMSE value of 2.89 suggests that the typical prediction error is less than 3 Gloss units, demonstrating good model precision.

Mean Absolute Error (MAE) - 1.91: The MAE, representing the average absolute difference between predicted and actual values, is relatively small at 1.91, further signifying the model's accuracy.

R-squared (R^2) - 0.69: The R-squared value of 0.69 indicates that approximately 69% of the variance in Gloss values is accounted for by the GPR model. This moderately high level of explanatory power suggests that the model effectively captures a significant proportion of the data's variability.

The Gaussian Process Regression model exhibits strong predictive performance for the Gloss data. The scatter plot and residuals distribution affirm a close alignment between actual and predicted values, albeit with some noted underprediction at higher Gloss values. The statistical metrics reflect the model's accuracy, with an average error of less than 3 units and a robust explanatory capacity covering a good portion of the data's variance. Collectively, these elements point to the GPR model being well-suited for this dataset.

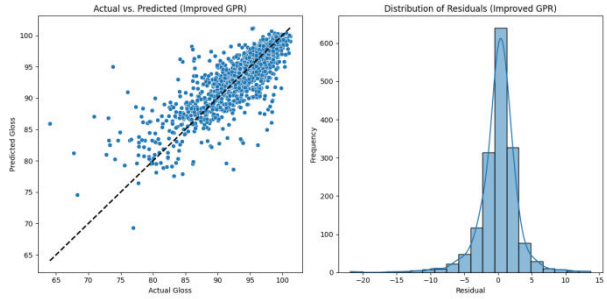

Figure 5 Actual vs Predicted values for gloss (Improved GPR) and Distribution of residuals (Improved GPR)

5.4 Decision Tree

Visual Representation:

Density Plot: This plot contrasts actual and predicted Gloss values. The actual values are represented by a red line and the predicted values by a blue line, with areas where these distributions overlap shown in purple. The close alignment of the two distributions indicates that the Decision Tree model has accurately predicted the density distribution of Gloss values, although discrepancies at the tails of the distribution remain, a common challenge with extreme values.

Scatter Plot: This plot illustrates the relationship between the actual Gloss values (x-axis) and the predicted Gloss values (y-axis). The points are tightly clustered around the dashed line representing perfect prediction, showing a strong correlation between the actual and predicted values. While most predictions align closely with the line, variance is still observable, particularly for higher Gloss values.

Statistical Results:

Optimized Decision Tree Mean Squared Error (MSE) - 8.56: This MSE is significantly lower than those of previously discussed models, indicating that the squared differences between actual and predicted values are smaller on average, suggesting a more precise prediction.

Optimized Decision Tree Root Mean Squared Error (RMSE) - 2.93: An RMSE value under 3 implies that the model's predictions are, on average, within approximately

2.93 units of the actual Gloss values, marking an improvement over previous models, such as the linear regression model.

Optimized Decision Tree Mean Absolute Error (MAE) - 1.85: The MAE of about 1.85 units points to a relatively low average absolute error, further confirming the model's accuracy.

Optimized Decision Tree R-squared (R^2) - 0.69: An R-squared value of 0.69 indicates that the model explains 69% of the variance in Gloss data, a substantial improvement over simpler models, and a strong indication of a good fit to the data.

The optimized Decision Tree model demonstrates robust predictive capabilities for the Gloss variable. The considerable overlap in the density plot and the tight clustering of points around the line of perfect prediction in the scatter plot suggest that the model's predictions align well with actual values. The statistical metrics reinforce this assessment, showing a relatively low MSE and RMSE and a substantially higher R-squared value compared to a linear regression model. This suggests that the Decision Tree model effectively captures the underlying patterns in the data. However, despite the enhanced performance, there remain some errors, and the model may not capture all complex relationships within the data, particularly at the extremes of the Gloss range. The results indicate that Decision Tree optimization has markedly improved the model's ability to predict Gloss, rendering it a potentially reliable method for this specific application.

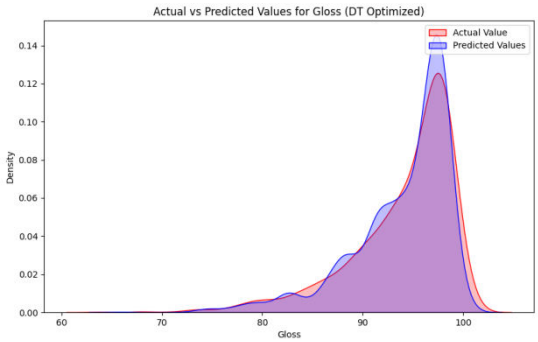

Figure 6 Actual vs Predicted values for gloss (DT)

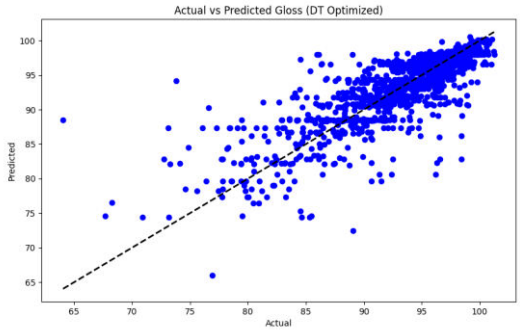

Figure 7 Actual vs Predicted values for gloss (DT)

Table 1 summarizes the statistical metrics (MSE, RMSE, MAE, and R-squared) for four different regression models: Support Vector Regression, Linear Regression, Gaussian Process Regression, and Optimized Decision Tree. Metrics highlight each model's accuracy and ability to explain the variability in Gloss values.

Table 1: Comparative Performance Metrics of Regression Models for Gloss Prediction

Model Type	MSE	RMSE	MAE	R-squared
Support Vector Regression	13.73	3.7	2.31	0.5
Linear Regression	18.42	4.29	3.08	0.33
Gaussian Process Regression	8.37	2.89	1.91	0.69
Optimized Decision Tree	8.56	2.93	1.85	0.69

The table presents a comparative analysis of four regression models used to predict Gloss. The Support Vector Regression (SVR) and Linear Regression models demonstrate moderate to lower performance, with R-squared values of 0.50 and 0.33 respectively, indicating less variability explained. Both models also exhibit higher MSE and RMSE values, reflecting greater prediction errors. In contrast, the Gaussian Process Regression (GPR) and Optimized Decision Tree models show superior predictive accuracy, each with an R-squared value of 0.69, suggesting they capture a significant portion of the variance in Gloss. Their lower MSE and RMSE values further confirm their higher precision in predictions, making them more reliable for practical applications.

VI. CONCLUSION

In this Research, the efficacy of various regression models including Support Vector Regression (SVR), linear regression, Gaussian Process Regression (GPR), and an optimized Decision Tree (DT) model has been explored in predicting a key variable, Gloss. The comparative analysis of these models has been rooted in a detailed examination of their predictive performances, using both visual representations and statistical metrics. The SVR model demonstrated a moderate ability to capture the trend of actual values with a density plot indicating good overlap, albeit discrepancies at peak values and a moderate R-squared value of 0.50. Linear regression displayed basic predictive abilities with an R-squared value of 0.33, suggesting a significant proportion of the variance in Gloss values remained unexplained by the model. The Gaussian Process Regression improved upon these with an R-squared value of 0.69, showing a stronger fit to the data. The residuals' distribution in this model highlighted an effective but not flawless prediction capability, with some tendencies towards under prediction at higher values.

The optimized Decision Tree model emerged as notably effective, characterized by a tight clustering in the scatter plot and a substantial R-squared improvement to 0.69, indicative of its robustness in capturing the variability of Gloss. Its MSE and RMSE values underscored a more precise prediction compared to other models examined.

From these analyses, it is clear that while each model has its strengths, the Decision Tree model, after optimization, provides a superior balance of accuracy and reliability in predicting Gloss. These findings underscore the importance of model selection and optimization in predictive analytics, particularly in contexts where precision is crucial. The insights gained from this study highlight potential pathways for enhancing predictive modeling techniques, suggesting that further tuning and exploration of complex models like Decision Trees could yield even more effective outcomes in similar applications.

REFERENCES

[1] Wang, K., Qin, Z., Tong, W., & Ji, C. (2020). Thermal energy storage for solar energy utilization: Fundamentals and applications. *Resources, Challenges and Applications*, *415*(9180410.5772).

[2] Tian, Y., & Zhao, C. Y. (2013). A review of solar collectors and thermal energy storage in solar thermal applications. *Applied energy*, *104*, 538-553.

[3] Mouaky, A., Merrouni, A. A., &Laadel, N. E. (2019). Simulation and experimental validation of a parabolic trough plant for solar thermal applications under the semi-arid climate conditions. *Solar Energy*, *194*, 969-985.

[4] Liu, Q., Yang, M., Lei, J., Jin, H., Gao, Z., & Wang, Y. (2012). Modeling and optimizing parabolic trough solar collector systems using the least squares support vector machine method. *Solar Energy*, *86*(7), 1973-1980.

[5] Stutz, B., Le Pierrès, N., Kuznik, F., Johannes, K., Del Barrio, E. P., Bedecarrats, J. P., ... & Minh, D. P. (2017). Storage of thermal solar energy. *Comptes Rendus. Physique*, *18*(7-8), 401-414.

[6] Kalogirou, S. A. (2004). Solar thermal collectors and applications. *Progress in energy and combustion science*, *30*(3), 231-295.

[7] Aye, L., & Jayalath, A. (2018). Applications of Solar Thermal Technologies in the Built Environment. *Low Carbon Energy Supply: Trends, Technology, Management*, 1-16.

[8] Wolfertstetter, F., Wilbert, S., Dersch, J., Dieckmann, S., Pitz-Paal, R., & Ghennioui, A. (2018). Integration of soiling-rate measurements and cleaning strategies in yield analysis of parabolic trough plants. *Journal of Solar Energy Engineering*, *140*(4), 041008.

[9] Mohana, N., Karunamurthy, K., &Isravel, R. S. (2023, April). Analysis of outlet temperature of parabolic trough collector solar water heater using machine learning techniques. In *IOP Conference Series: Earth and Environmental Science* (Vol. 1161, No. 1, p. 012001). IOP Publishing.

[10] Price, H. (2003, January). A parabolic trough solar power plant simulation model. In *International solar energy conference* (Vol. 36762, pp. 665-673).