YUGATO

ISSN: 0387-5695, eISSN: 0387-5695
Vol. 76 No. 1 (2024)

Received: November 16, 2023 / Revised: December 11, 2023 / Accepted: January 09, 2024 / Published: January 30, 2024

AUDIENCE ENGAGEMENT PREDICTION USING RANDOM FOREST REGRESSOR

Dr.R.Mary Metilda
Professor & Head, SREC Business School, Sri Ramakrishna Engineering College

Kanishkaa.K
Student, SREC Business School, Sri Ramakrishna Engineering College

A. Sagayarani
Assistant Professor, SREC Business School, Sri Ramakrishna Engineering College

Abstract

The concept of the study revolves around predicting the audience for a TV channel, which hasvarious applications in the broadcasting industry. Accurate audience prediction helps broadcasters and advertisers make informed decisions regarding scheduling, advertising strategies, and resource allocation. The study utilizes the Random Forest Regressor algorithm,which combines multiple decision trees to enhance accuracy and robustness in forecasting audience engagement. The Random Forest Regressor is an ensemble learning algorithm that handles high-dimensional data modeling and can handle missing values, and continuous, categorical, and binary data. It overcomes the problem of overfitting and does not require tree pruning. The study aims to evaluate the effectiveness of this algorithm in forecasting audienceengagement. Overall, the study aims to provide insights into audience behavior and optimize audience engagement using the Random Forest Regressor algorithm.

Keywords: *Audience Engagement, Random Forest Regressor*

Introduction

The ability to predict the audience for a TV channel holds immense significance and offers a wide range of applications within the broadcasting industry. Accurately forecasting the expected viewership for a specific channel empowers broadcasters and advertisers to make well-informed decisions and optimize various aspects of their operations. In the TV industry, in many areas, such as Hong Kong and the US, around 80% of advertising slots are sold in advance and 20% are retained for sale at higher prices in the scatter market, (Song, L., et.al.,2020). One key implication is the optimization of scheduling, whereby audience prediction enables broadcasters to strategically plan the airing of programs by identifying the most opportune time slots and the types of content that are likely to attract a larger audience. This, in turn, can result in increased viewership and improved audience engagement. Despite the increase in the number and types of media, television still receives the largest proportion ofadvertising expenditures, (Schweidel Moe, 2016), Moreover, audience prediction plays a crucial role in shaping advertising strategies. Advertisers can leverage the predictions to effectively target their advertisements by placing them during programs or time slots projectedto have a higher viewership.

Predicted viewership numbers enable channels to negotiate advertising rates and optimize advertising revenues. Furthermore, for channels employing a subscription-based model, audience prediction aids in estimating potential subscription revenues. Another crucial application lies in content development. By understanding audience preferences and behavior through prediction models, TV channels can tailor their content offerings to cater to the interests of their target audience. This targeted approach enhances viewer engagement, fostersviewer loyalty, and has the potential to attract new viewers.

Theoretical Background of the Study

The Random Forest Regressor functions as an ensemble learning method that merges numerous decision trees, aiming to improve accuracy and resilience. It is employed to predict audience engagement for a specific channel over 15 days, using historical data on audience numbers andtiming. The study aims to evaluate the effectiveness of the Random Forest Regressor in accurately forecasting audience engagement and its potential implications for decision-makingrelated to scheduling, advertising strategies, and resource allocation. The research focuses ona specific channel and its audience, utilizing secondary data sources. Every Decision Tree is made by randomly selecting the data from the available data, (Ali, J., et.al.,2012). The RandomForest is appropriate for high-dimensional data modeling because it can handle missing valuesand can handle continuous, categorical, and binary data, (Ali, J., et.al.,2012). The study aims to provide insights into audience behavior and improve the utilization of available data sourcesfor optimizing audience engagement. Random forests can be related to two main sources, regression trees, and bagging. Regression trees are constructed by recursive partitioning of the input space based on some criterion, dependent or independent of the data to estimate the regression function. (Benjamin Goehry 2020).

Each decision tree focuses on different aspects of the data, capturing diverse patterns and reducing the impact of individual outliers or noisy data points. Additionally, the Random ForestRegressor handles high-dimensional datasets effectively by randomly selecting subsets of features for each decision tree.

Review of Literature

Khan, R. S., & Bhuiyan, M. A. E. (2021) explore AI methodologies' effectiveness, like decision trees and neural networks, in improving rainfall estimates in the Upper Blue Nile Basin in their study titled "Artificial Intelligence-Based Techniques for Rainfall Estimation Integrating Multisource Precipitation Datasets." El Mrabet, Z., Sugunaraj, N., Ranganathan, P., & Abhyankar, S. (2022) introduce a data-driven method employing Random Forest Regressor (RFR) models for real-time detection of fault characteristics in power systems, simultaneously identifying fault location and duration using fault trajectory samples. Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020) focus on crafting a tailored text classification system for BBC news articles, emphasizing the critical role of various sections and preprocessing techniques in refining raw text data. Rodriguez-Galiano .V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015) conduct a detailed performance analysis of machine learning algorithms in mineral prospectivity modeling, concentrating on epithermal Au prospectivity mapping in the Rodalquilar district, Spain. Fawagreh, K., Gaber, M. M., & Elyan, E. (2014) present a comprehensive overview of the evolution and advancements of Random Forest (RF) in ensemble classification, focusing on its high accuracy and ongoing research for further enhancement.

Methodology

This study primarily focuses on forecasting the engagement levels of a specific channel over the forthcoming 15 days. Employing an explanatory research design, the methodology hinges on secondary data usage, utilizing the Random Forest Regressor as the analytical tool and Python as the software for analysis. Upon data collection, the subsequent phase involves data preprocessing to ensure its quality and suitability. This encompasses rectifying missing values, addressing outliers, and structuring the data suitably for input into the Random Forest Regressor algorithm.

The research design adopted here is explanatory, aiming to decipher the interrelation between variables: time, geographic regions, and audience engagement of the designated channel. The key objective is to predict audience engagement based on these specific variables. This design allows for data analysis and interpretation, offering insights into the factors influencing audience engagement patterns. The data utilized in this study originates from secondary sources, signifying its retrieval from previously amassed repositories. These sources likely comprise historical records of the channel's audience engagement, encompassing viewership statistics alongside other pertinent variables like timing and geographical viewer distribution. The Random Forest Regressor serves as a potent machine-learning algorithm in this study. Trained on historical data, the model aims to discern the correlation between time, geographic areas, and audience engagement. The study's outcomes can aid broadcasters and advertisers in decision-making processes related to scheduling, advertising strategies, and resource allocation. Upon training the model using the dataset, an accuracy score of 90% was achieved. This high accuracy signifies the model's proficiency in predicting channel audience engagement. The findings underscore the Random Forest Regressor's effectiveness in forecasting audience engagement levels for the channel.

Python, a widely used programming language renowned for data analysis and machine learning, was the primary software employed. Leveraging various libraries and tools within Python facilitated data manipulation, visualization, and modeling. Specifically, the analysis centered on the Random Forest Regressor algorithm, implemented through Python's machine learning library, like sci-kit-learn. Renowned for its versatility and robustness, this algorithm amalgamates multiple decision trees to ensure accurate predictions.

The analysis process involved multiple steps utilizing Python and the Random Forest Regressor algorithm. Initial stages encompassed data preprocessing, including tasks such as data cleansing, feature selection, and handling missing values. Subsequently, the data was partitioned into training and testing sets, with the former used to train the Random Forest Regressor model. The training phase involved fitting the model to the data and constructing numerous decision trees using random data subsets and features, thereby minimizing overfitting and enhancing the model's generalization ability.

Result and Discussion

Data Preprocessing

Data preprocessing is an important step in the data analysis process that involves cleaning and transforming raw data into a format suitable for analysis. It includes several techniques such as handling missing values, addressing

outliers, dealing with categorical variables, and normalizing or standardizing numerical data. One approach to handling such missing values is to eliminate the rows that contain them. Another alternative is to employ imputation techniques, such as mean imputation, which involves replacing the missing valueswith the mean value of the corresponding feature.

Furthermore, it is advised to conduct a comprehensive examination of the dataset to identify and handle any duplicates or irrelevant data that could potentially impact the accuracyand reliability of the analytical model. This enhances the accuracy and reliability of subsequentanalyses and modeling processes, enabling researchers to draw meaningful insights and makeinformed decisions based on clean and trustworthy data. One of the first steps in data preprocessing is to check for missing values. This involvesidentifying any missing data in the dataset and determining how to handle it, whether by fillingin the missing values with an appropriate estimate or removing the observations with missing values altogether. Outliers, which are extreme values in the dataset that do not conform to thegeneral pattern of the data, can also be detected during the preprocessing stage.

Finally, normalizing or standardizing numerical data is another important technique in data preprocessing. Scaling numerical data to achieve a mean of zero and a standard deviation of one is part of the process. This action can enhance the efficiency of certain machine-learning algorithms by guaranteeing uniform scales for all features. Data preprocessing stands as a pivotal phase within data analysis, ensuring data readiness for analysis and potentially augmenting result accuracy.

Data Exploring

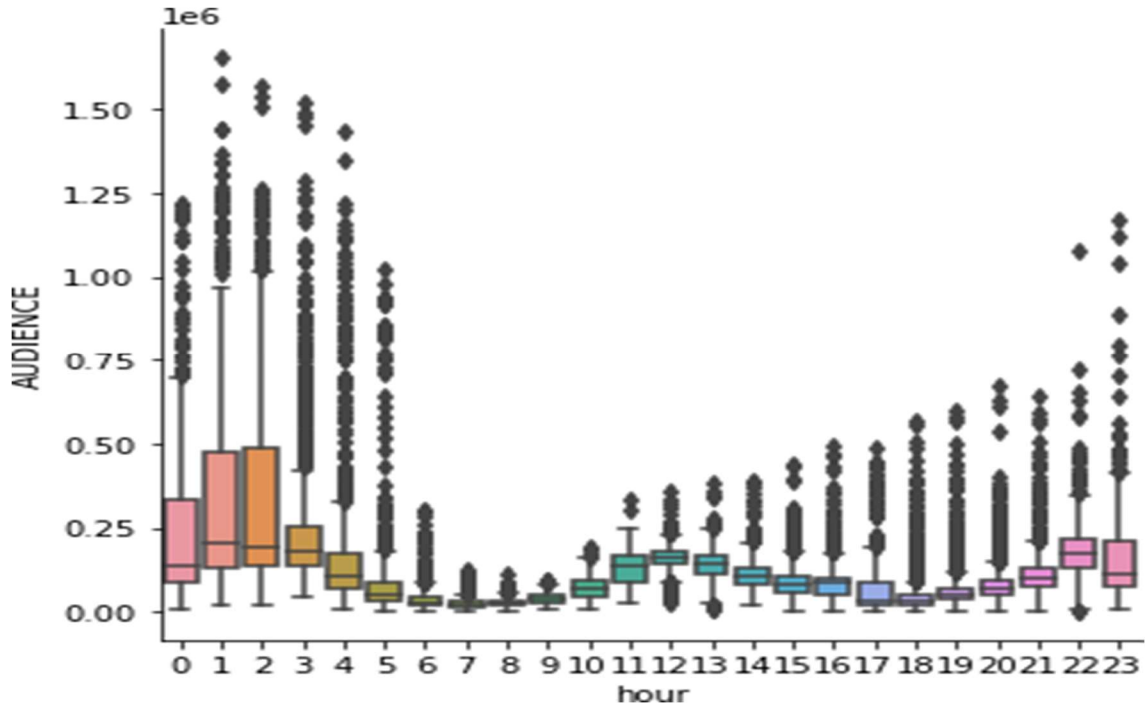

Chart 4.2 – Boxplot of the Audience on an Hourly basis

Chart 4.2 is a boxplot, a valuable visualization tool used to display the distribution of a dataset, particularly in terms of its central tendency, variability, and range. Based on the provided information regarding the number of audiences on an hourly basis, we can predict the characteristics of a boxplot for this dataset. In the boxplot, the median line is positioned at the center of the box, representing the dataset's central tendency. From the information given, we can expect the median number of viewers to be larger during the firstand second hours

Furthermore, the whiskers in the boxplot extend outward from the box and represent the data range, excluding outliers. In this case, we would expect the whisker on the right side of the box to be longer, indicating the presence of outliers at specific times of the day. These outliers could represent unusually high or low numbers of viewers during those particular hours.

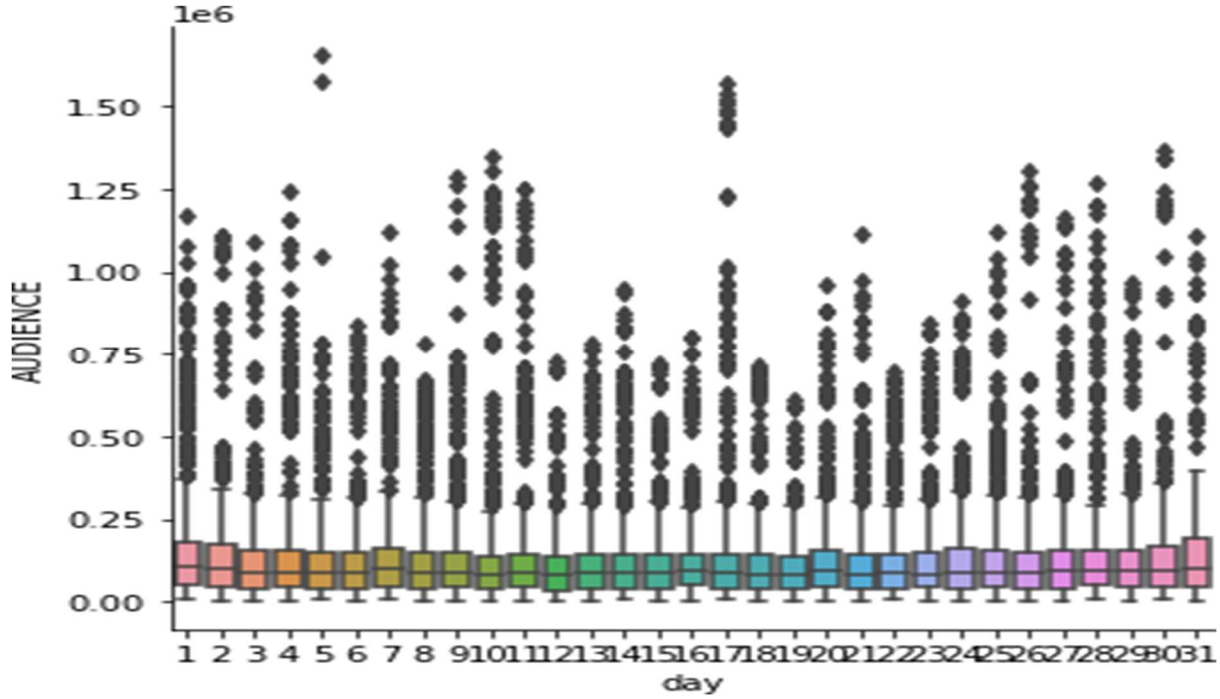

Chart 4.3 – Boxplot of the audience daily

Chart 4.3 presents a box plot of the number of viewers for each day, it is evident that there is minimal variation in the data across the different days. This is indicated by the consistency in the size and shape of the boxes in the box plot, which represents the interquartilerange (IQR). The IQR provides insights into the spread and variability of the data, with a largerIQR suggesting greater variation. In this case, the boxes remain relatively similar in size, suggesting that the number of viewers does not differ significantly from day to day. To gain amore comprehensive understanding of the data, further examination of the quartiles and outliersis necessary. The quartiles, specifically the lower quartile (25th percentile) and the upper quartile (75th percentile), provide additional information about the central tendency and distribution of the data. Comparing these quartiles across the days would help identify any potential differences or patterns.

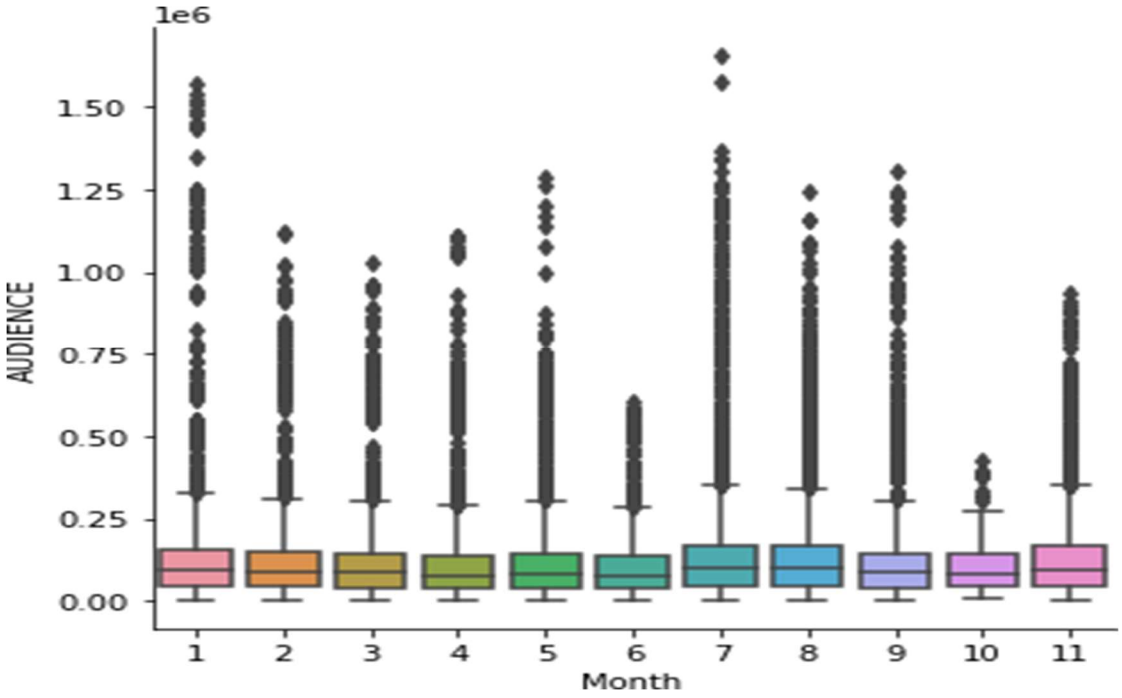

Chart 4.4 – Boxplot of the audience every month

Chart 4.4 box plot illustrates the number of audiences for each month, and it indicatesthat there is relatively limited variation in the data between the different months. This observation is supported by the consistent shape and positioning of the boxes in the box plot, suggesting that the distribution of the number of audiences remains fairly stable throughout themonths.

In summary, the analysis of the box plots indicates the level of variation and distributionpatterns in the number of audiences over time. Further examination of quartiles, and outliers, and employing feature engineering techniques can provide more detailed insights and enhance the model's accuracy.

Training and Testing of Data

After preparing the data, it is important to split it into training and testing sets. The training set is used to train the model, whereas the testing set is used to assess the model's performance. In a typical machine learning workflow, the available data is usually divided intotwo sets: training and testing sets. The training set is used to train the machine learning model, while the testing set is used to evaluate the performance of the trained model on new, unseen data.

In this case, the total size of the available data is 26700. This could represent the total number of samples or observations in the dataset. The training data size is 26000, which meansthat 26000 samples are used to train the machine learning model. The remaining 700 samples are used as the testing data set to evaluate the performance of the trained model.

```
Regression = RandomForestRegressor(n_estimators = 100, max_depth=10, random_state=0
```

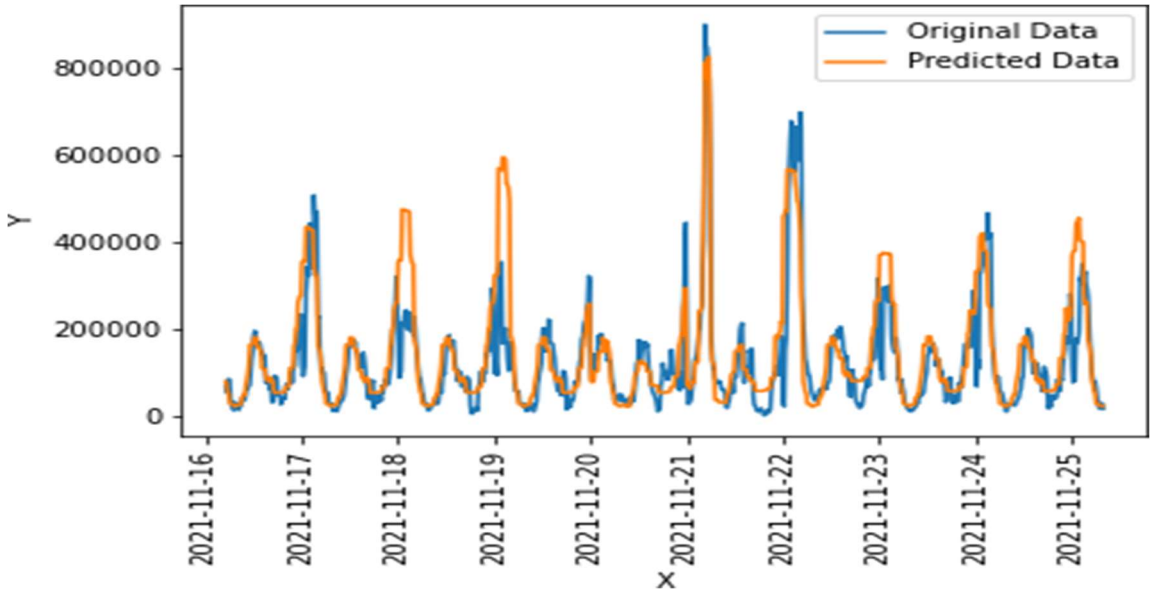

Chart 4.5 – Line graph of original and predicted data to check the accuracy

Chart 4.5 is the line chart that displays two distinct datasets: the original data and the predicted data, which are compared specifically for the time frame spanning from November 16th to November 25th. The close resemblance observed between these two datasets can be attributed to the utilization of a highly accurate model, as indicated by an accuracy score of 90.This score signifies that the predicted data aligns remarkably well with the original data, showcasing the model's capability to capture and replicate the underlying patterns and trends present in the original dataset. By achieving such a high accuracy score, the model demonstrates its effectiveness and reliability in accurately reproducing the behavior of the original data.

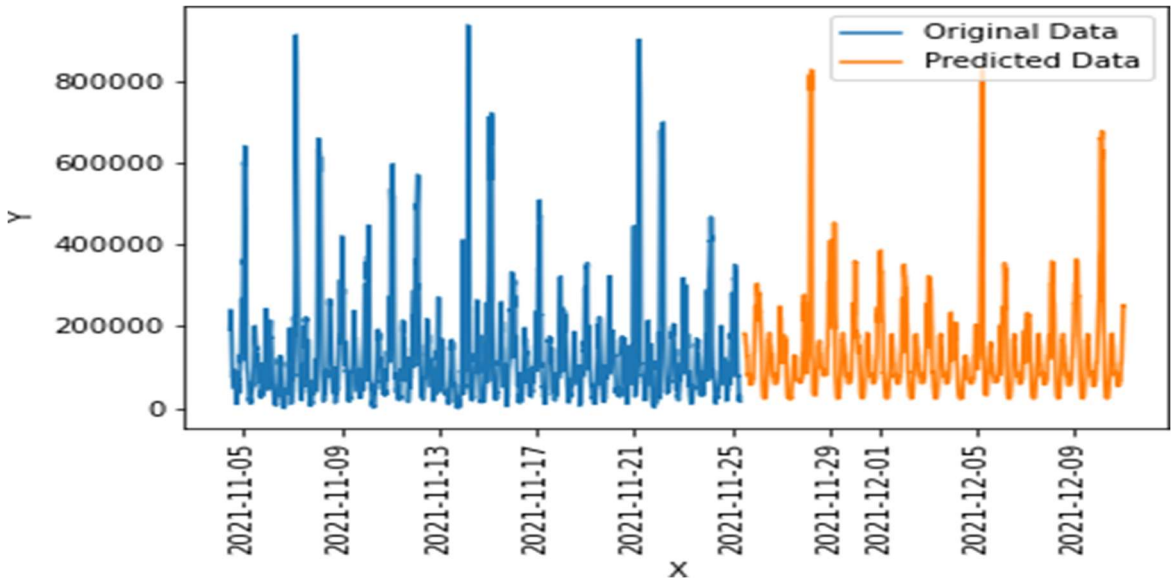

Chart 4.6 – Line Graph of the predicted data from November 26th to December 1st.

Chart 4.6 is the graph showcasing two distinct datasets: the original data, spanning fromNovember 5th to November 25th, and the predicted data, covering the period from November26th to December 1st. Despite the shift in the time frame covered by each dataset, a consistentpattern is observed in both sets of data, indicating a certain level of stability and minimal fluctuations. This consistency between the original and predicted data suggests that the modelemployed to generate the predicted data effectively captured the underlying patterns and trendspresent in the original dataset.

Conclusion

In conclusion, random forest regression models excel in capturing complex relationships within high-dimensional datasets. By training these models on the training set andfine-tuning their hyperparameters, their performance can be significantly improved. Evaluatingthe accuracy of the model involves comparing its predictions with the original data. A high accuracy score signifies the model's ability to effectively capture patterns and trends, instillingconfidence in the reliability of the predicted data for decision-making and future trend forecasting.

REFERENCES

• Khan, R. S., & Bhuiyan, M. A. E. (2021). “Artificial Intelligence-Based Techniques for Rainfall Estimation Integrating Multisource Precipitation Datasets.” Atmosphere, 12, 1239.

• Shah, K., Patel, H., Sanghvi, D., & Shah, M. (2020). “A Comparative Analysis of Logistic Regression, Random Forest, and KNN Models for Text Classification.” Augmented Human Research, volume 5, Article number: 12.

- Nogueira, J., Guardalben, L., Cardoso, B., & Sargento, S. (2018). "Catch-up TV forecasting: Enabling next-generation over-the-top multimedia TV services." Multimedia Tools and Applications, 77(12), 14527-14555. doi:10.1007/s11042-017-5043-9

- Chang, N., & Sheng, O. R. L. (2008). "Decision-Tree-Based Knowledge Discovery: Single- vs. Multi-Decision-Tree Induction". INFORMS Journal on Computing, 20(1), 46-54.

- Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). "Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees, and support vector machines." Ore Geology Reviews, volume 71, 804-818

- Moreta, C. E. G., Acosta, M. R. C., & Koo, I. (2019). "Prediction of Digital Terrestrial Television Coverage Using Machine Learning Regression. IEEE Transactions on Broadcasting, 65(4), 702-712. DOI: 10.1109/TBC.2019.2890065

- El Mrabet, Z., Sugunaraj, N., Ranganathan, P., & Abhyankar, S. (2022). "Random Forest Regressor-Based Approach for Detecting Fault Location and Duration in Power Systems". Sensors, 22(2), 458. doi:10.3390/s22020458

- Fawagreh, K., Gaber, M. M., & Elyan, E. (2014) "Random forests: From early developments to recent advancements." International Journal of Data Warehousing and Mining, 10(4), 602-609.

- Schonlau, M., & Zou, R. Y. (2020). "The Random Forest Algorithm for statistical learning." The Stata Journal, 20(1), 3-29. DOI: 10.1177/1536867X20909688.

- Wang, H., Wu, B., Yao, Y., & Qin, M. (2019). "Wideband Spectrum Sensing Method Based on Channels Clustering and Hidden Markov Model Prediction". Information, 10(11), 331. DOI: 10.3390/info10110331.