Explainable Machine Learning for Predicting Solar Power Output

Pranav R

Student, Department of EEE
B.M.S. College of Engineering
Bangalore, India
pranav.ee18@bmsce.ac.in
shashanktk.me19@bmsce.ac.in

Shashank T K

Student, Department of ME
B.M.S. College of Engineering
Bangalore, India

Gururaja H S

Assistant Professor, Department of ISE,
B.M.S. College of Engineering
Bangalore, India
gururajhs.ise@bmsce.ac.in

Abstract.

In an everlasting effort to move away from fossil fuels, solar photovoltaic power has emerged as one of the most promising environmentally friendly, renewable power sources in the ongoing endeavor to move away from fossil fuels. However, lack of understanding optimum solar panel positioning would result in less optimal power output from the panels. This research provides a Machine Learning method for estimating solar panel power production. Three state-of-the-art tree based algorithms namely, LightGBM, XGBoost and Random Forests are considered in this study to predict the solar power output using historical data. Machine Learning algorithms, although extremely powerful, suffer from the "black-box" problem which is a condition where a user cannot explain the predictions of ML algorithms. The issue has been addressed in this paper using SHAP Explainable AI. Model explainability is extremely critical in determining which variables or features played the most important roles in models prediction. In this paper, model explainability helps understand which features are most important in predicting Solar Power Output. This helps industries set up solar panels in such a way that maximum electrical power can be extracted from the panels.

Keywords. Solar Power, Machine Learning, Explainable AI, SHAP Xai.

1. INTRODUCTION

Artificial Intelligence is turning out to be increasingly significant in deeply shaping our day to day routines. Moreover, as AI-based arrangements [1,2,3] multiply in fields like loaning, law enforcement, medical services, and schooling, the individual and expert outcomes of AI are wide-running. Because AI models assume such an overwhelming part in different spaces, there is rising stress over possible predisposition in these models, as well as a necessity for model interpretability. Model logic is also critical for building trust and reception of AI frameworks in high-stakes areas requiring unwavering quality and security, such as medical services and robotized transportation, as well as basic modern applications with significant monetary implications, such as predictive maintenance, regular asset investigation, and environmental change displaying.

Deep learning has contributed a substantial amount to artificial intelligence's recent progress. Deep learning algorithms, as proven in [4,5], outperformed traditional machine learning methods significantly. Deep Neural Networks (DNNs) on the other hand, are bad at describing their inference processes and final outcomes, and both developers and consumers consider them as a black box.

Solar energy, in its various structures, sunlight based heat, sun powered photovoltaic, sunlight based warm power, and solar oriented fuels gives humanity a perfect, environment well disposed, extremely plentiful, and endless energy asset. Sun oriented power is the transformation of daylight into power, either straightforwardly through Photovoltaic (PV) boards or by implication through Concentrated Solar Panels (CSP). Starting from the beginning, research has been led to create a modest, non-exhausting, and clean sun powered energy framework with long haul benefits. Therefore, the study in [6] inspects the advancement of sun oriented power age innovative work since its initiation. The existing and future problems associated in the creation of quality and reliable solar power technologies for future applications are also highlighted.

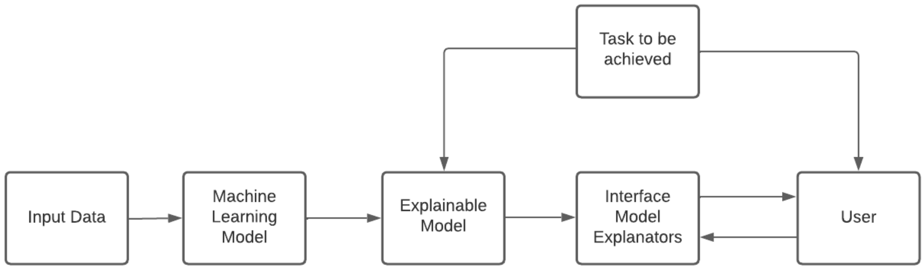


Figure 1. Explainable AI Architecture

Explainable AI is a collection of tools and frameworks that can aid in the understanding and interpretation of machine learning predictions. Figure 1. Represents the working of Explainable AI model for predicting the solar power output generation.

2. LITERATURE REVIEW

One study examines how to expect sunlight based photovoltaic (PV) energy utilizing Explainable AI (XAI) instruments like ELI5, SHAP and LIME, which can assist with the reception of XAI methods for shrewd lattice applications [8]. Understanding the internal functions of an AI-based forecast model can give knowledge into the field of use. Such information can assist with improving sun based PV forecasting models and distinguish significant qualities.

For short-term solar power prediction, [9] presents a least-square (LS) SVM-based model (SPP). The model accounts for variables such as wind speed, relative humidity, and sky cover, and provides verifiable information on air transmissivity in an original two-layered (2D) structure. The model's output is projected environmental transmissivity, which is then converted to solar oriented electricity based on the scope of the site and the time of day. Utilizing data from the NSRDB, which stands for **National Solar Radiation Database**, virtual simulations are performed to endorse the proposed model. The outcomes propose that the suggested model not simply just beats a reference autoregressive (AR) model concerning assumption accuracy, but moreover defeats a RBFNN, which is short for Radial Basis Function Neural Network - based model.

One of the barriers to their widespread use in power networks is the random character of solar energy supplies. Using meteorological data is the most feasible technique to forecast this renewable energy source. However, such information is frequently presented in a subjective configuration that can't be utilized with regular quantitative methodologies however can be demonstrated with fuzzy logic. Study [10] proposes type-1 and stretch sort 2 Takagi-Sugeno-Kang (TSK) fluffy frameworks for the demonstrating and expectation of sun based power plants. Type-2 TSK models with type2 predecessors and fresh outcomes supply the optimum presentation in light of the sun-powered plant information.

The research [11] proposes a method for forecasting generation of solar power based on the least absolute shrinkage and selection operator (LASSO). The model compares the proposed system to two example schemes using three real-world datasets. Using three real-world datasets, the model compares the proposed system to two sample plans (schemes). The work has additionally found that the algorithm based on LASSO accomplishes altogether higher accuracy than current techniques, while utilizing fundamentally less preparation information and being vigorous to inconsistency data of interest in the preparation information. Its variable determination capacity additionally gives an advantageous tradeoff among intricacy and accuracy, all of which join to make the proposed LASSO-based approach an exceptionally serious answer for forecasting solar power generation.

The study [12] tackles the issue of projecting renewable energy in the short term. Renewable energy sources' stochastic character affects power system arranging methods, influencing the reliability and security of force supply for extreme clients. Utilizing retroactive metering information and open source climate data offered by meteorological types of assistance, the creators propose an ML approach for effective day-ahead anticipating. The challenge of feature identification and acceptable error measures is addressed in this study. The user interface was built and tested on a genuine solar power plant in the Russian Federation's southern region in order to ensure effective testing.

Tree based State of the art models like LightGBM, XGBoost and Random Forests are being used in research work for many classification and regression problems nowadays. Study [13] proposed a LightGBM model to predict customer loyalty in the finance industry. It was found that LightGBM model performed better than XGBoost model in terms of RMSE score. A different study used LightGBM for Sales Forecasting. It was found that LightGBM outperformed classical methods like SVM and Logistic regression [14]. Another study used Random Forests and XGBoost algorithms to predict a student's behavior [15]. Model explainability is an important aspect to tackle the black-box problem of ML models. A particular study used SHAP method to explain reinforcement learning for power system frequency control [16]. In [17], SHAP and LIME explainable models were used to explain ML algorithms for identifying failure in a network of microwaves. Another study [18] implemented Dalex Xai for interpreting algorithms used for predicting credit defaulters. This paper focuses on implementing the SHAP model explainability.

3. METHODOLOGY

3.1. Random Forests

Random Forests is a decision tree-based ensemble machine learning model. It contains an implicit decision system, thus it can handle a variety of data restrictions without removing a few for reduced dimensionality [19]. The marginal function for a given classifier's group $d_1(a)$, $d_2(a)$ ….. $d_k(a)$ and the training set drawn aimlessly from the dispersion of the arbitrary vector B, A is characterized as shown :

$$mg(A, B\ ) = av_k I(h_k(A) = B\ ) - max\ jA\ av_k I(h_k(A) = j) \tag{3.1.1}$$

where *I(·)* is the indicator function. The generalization error is given by:

$$PE^* = P_{A,B}(mg(A, B) < 0) \tag{3.1.2}$$

3.2. XGBoost

XGBoost is a supervised learning algorithm based on ensemble trees. Numerous classification projects in different fields are being applied using this model [20]. Many Kaggle competitions that are won nowadays, use XGBoost. Crediting to its astounding performance, XGBoost opens door to a lot of research opportunities.
It targets enhancing a cost objective function made out of a loss function (*l*) and a regularization term (δ):

$$\mathrm{F}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \delta(f_k) \tag{3.2.1}$$

Where fk is a specific tree from the ensemble group trees, n is the number of cases in the preparation set, K is the number of trees to be provided, and is the predictive value. The term "regularization" is defined as follows:

$$\delta(f_t) = \gamma T + \frac{1}{2}\left[\alpha \sum_{j=1}^{T} | r_j | + \lambda \sum_{j=1}^{T} r_j^2\right] \tag{3.2.2}$$

Where γ is loss reduction factor, λ is regularization term and r is weight corresponding to a tree leaf.

3.3. LightGBM

LGBM additionally uses continuous feature histogram binning, which provides even greater speed than classic gradient boosting. Binding numeric values reduces the amount of split points in decision trees and eliminates the necessity for sorting methods, which are always computationally intensive. Light GBM builds trees upward, but other computations build trees equally, implying that Light GBM builds trees not level by level but rather leaf by leaf. It will grow the leaf that has the most delta-loss. While growing a similar leaf, leaf-wise calculations can restrict more loss than level-wise calculations [21].

3.4. SHapley Additive exPlanations (SHAP)

SHAP is a visualization tool that can be utilized to interpret the result of an AI/ML model to make it more reasonable. It can be used in making sense of any model's forecast by processing each variable's commitment to the prediction can be utilized. It combines a number of tools, including lime, SHAPley sample values, DeepLift, QII, and others. SHAPley values are one of the essential components of the SHAP tool since it allows SHAP to correlate optimal credit allocation with local explanations. The Figure represents the Pipeline of post hoc explainable tool SHAP [7].

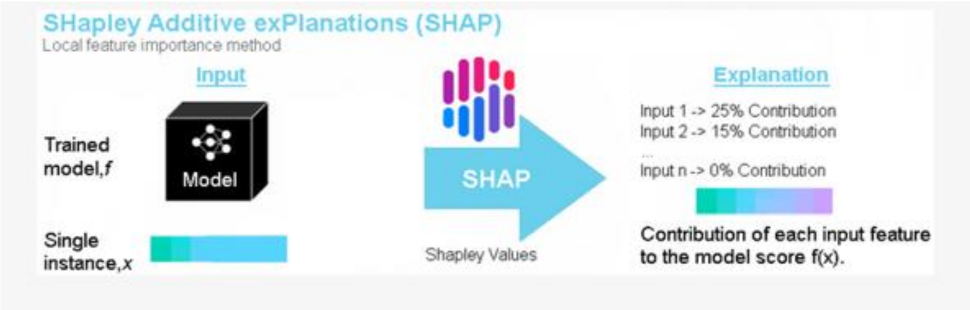


Figure 3.4. Visual of SHAP Explainable model

4. Results and Discussion

4.1. Data analysis and pre-processing

The dataset used for this study was taken from Kaggle. It has 21045 training examples consisting of weather data and power output acquired for 14 months. The dataset was analyzed and checked for missing values. The categorical variables were handled using Label Encoder. The dataset was split in the ratio 8:2(80% training and 20% testing) and was standardized using sklearn's StandardScaler.

4.2. Model evaluation

Models	Metrics		
	RMSE	MAE	R2 Score
LightGBM	4.029	2.632	0.681
XGBoost	4.134	2.639	0.664
Random Forests	4.150	2.833	0.661

For evaluating the models, the paper considers the following metrics:

Mean Absolute Error:

The Mean Absolute Error (MAE) is a regression model evaluation statistic. The difference between the true value and the anticipated value for each instance is the prediction error.

$$mae = \left(\sum_{i=1}^{n} abs\left(yi - \lambda(\mathrm{xi})\right) \right) \div \mathrm{n}$$

(4.2.1)

R2 score:

R-squared (R2) is a factual measure that evaluates how much variety made sense of by an independent variable or factors in a relapse model for a dependent variable

$$R^2 = 1 - \frac{\sum(y - \hat{y}_i)^2}{\sum(y - \bar{y}_i)^2}$$

(4.2.2)

Root Mean Squared Error:

Having a solitary number to quantify a model's exhibition is unquestionably valuable in AI, whether it's during preparing, cross-validation, or checking after organization. One of the most frequently involved estimations for this is root mean square mistake. It's a straightforward scoring decide that is additionally viable with a few of the most pervasive measurable presumptions.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y - \hat{y}_i)^2}{N}}$$

(4.2.3)

By considering all the parametric evaluation results, the results of each model is depicted in Fig 4.2. It can be seen that the LightGBM performed best on this dataset in terms of all the considered metrics followed by XGboost with Random Forest not too far behind.

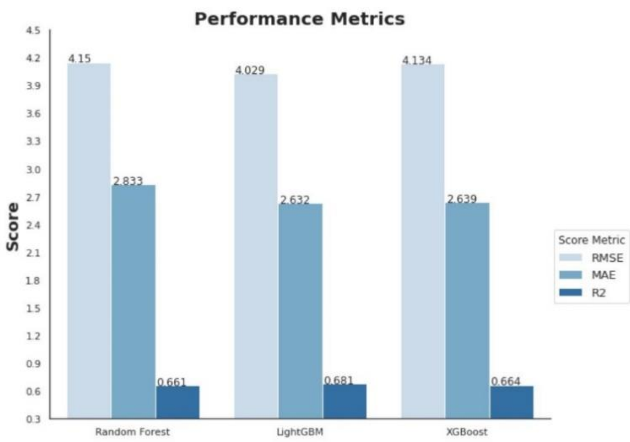
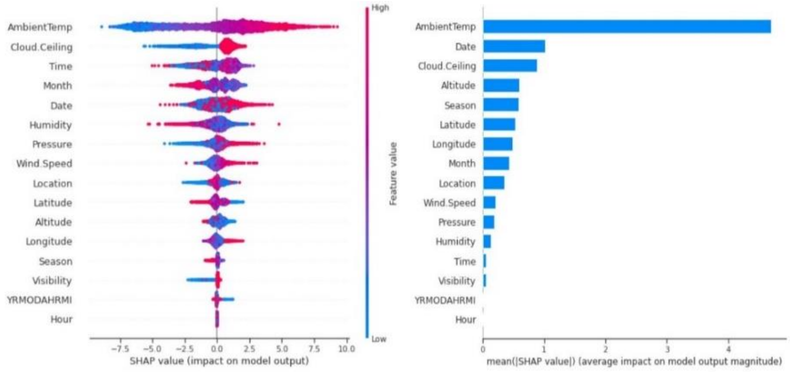


Figure 4.2. Bar Plot of performance metrics of the models

4.3. Model Explainability

4.3.1. Light Gradient Boosted Machine (LGBM):

The bee swarm plot depicts the overall performance of the model and discusses the features that led to model prediction. As can be seen, according to LGBM, the ambient temperature was the most crucial feature. Lower ambient temperatures (blue color) do not help the panels produce a good quantity of electricity production. For a single instance, the bar plot provides a good interpretation of features that contribute to model prediction. It should come as no surprise that the most essential feature was the ambient temperature.
The bee swarm plot depicts the overall performance of the model and discusses the features that led to model prediction. As can be seen, according to LGBM, the ambient temperature was the most crucial feature. Lower ambient temperatures (blue color) do not help the panels produce a good quantity of electricity production. For a single instance, the bar plot provides a good interpretation of features that contribute to model prediction. It should come as no surprise that the most essential feature was the ambient temperature.



(a) (b)

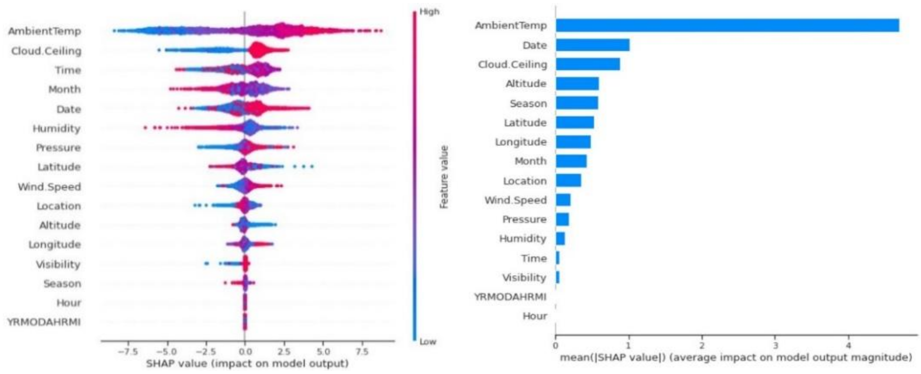
Figure 4.3.1. (a) Bee-Swarm plot for global interpretability. (b) Bar plot for local interpretability for LGBM model.

4.3.2. Extreme Gradient Boosting (XGBoost):

Tree boosting is a popular and successful machine learning technique. It additionally remembers data for store access designs, information compression, and sharing to assist with building a versatile tree boosting framework. XGBoost grows past billions of illustrations utilizing impressively less assets than earlier frameworks by coordinating these thoughts.

We can note that the XGBoost and LightGBM were dependent on similar features for model prediction. However, it is important to note that XGBoost gave more importance to Latitude than Wind speed and Location in contrast to the LGBM model.

It is interesting to note that when a single instance is considered, the same features, in the same order contributed for both XGBoost and LGBM models. This comes as no surprise as XGBoost and LGBM are always in fierce competition with each other and a single instance is just not enough to decide which features take precedence in model prediction.



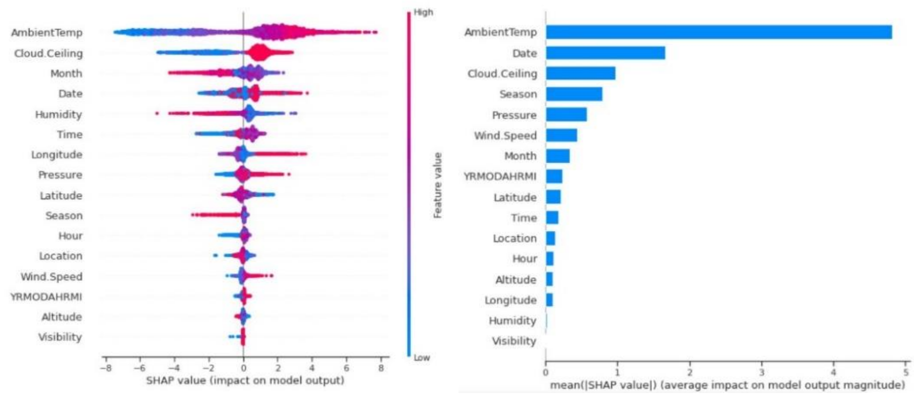
(a) (b)

Figure 4.3.2. (a) Bee Swarm plot. (b) Bar plot for XGBoost Model

4.3.3. Random Forests (RF)

It is no surprise that ambient temperature is the most important factor in random forests as well when the entire range of training examples is considered. Month and date, rather than time, take primacy. According to Random Forest, the hours of the day and the season have risen up the table and are now deemed fairly essential in comparison to altitude and visibility.

However, when a single instance is considered, date ends up as a significant feature after Ambient temperature for model prediction. Season and WindSpeed are also considered important which was not seen in LGBM or XGBoost. Another point to note is that, Humidity, which was a fairly important feature for model prediction in LGBM and XGBoost, has plummeted to the bottom. This can be attributed to Random Forests missing the key features of the data by a decent margin as it is the worst performing model of the three considered models.



(a) (b)

Figure 4.3.3. (a) Bee-Swarm plot for global interpretability. (b) Bar plot for local interpretability for LGBM model.

5. CONCLUSION AND FUTURE SCOPE

Based on input characteristics such as solar PV panel's temperature, ambient temperature, solar flux, relative humidity and time of the day, tree based ML methods such as Random forest, XGBoost, and LGBM, which support non-linearity, were explored to predict PV power. Regardless of the high performance of the suggested ML techniques, other factors such as the size of the solar PV panel, dust, and wind can have a significant impact on solar PV output.

With the help of SHAP Explainable AI, the features that most impacted for model prediction was interpreted, thus tackling the black-box issue of ML algorithms. Going forward, this research can be implemented with hybrid and stacked ensemble models to achieve better results. A comparative analysis of different XAI models can be done to visualize the behavioural aspects of different models. Explainable AI is a hot topic now and provides ubiquitous opportunities to conduct research.

6. REFERENCES

[1] Gade, K., Geyik, SC, Kenthapadi, K., Mithal, V., and Taly, A. (July 2019). Descriptive AI in the industry. Minutes of ACM SIGKDD International Conference on Understanding Discovery and Real Mining

[2] Samek - W. - Montavon - G. - Vedaldi - A. - Hansen - L. K, & Müller, K.R. (Edit). (2019). Descriptive AI: Explanation of in-depth research, explanation, and visualization (Volume 11700). Springer Nature.

[3] Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., and Zhu, J. (October 2019).

[4] Holzinger, A. (2018, August). From studying gadget to descriptive AI. In 2018 an international series on virtual intelligence of buildings and equipment (DISA) (pages 55-66). IEEE.

[5] Tjoa, E., & Guan, C. (2020). Survey on descriptive ingenuity (xai): Focus on xai science. Implementation of IEEE on neural networks and learning structures, 32 (11), 4793-4813.

[6] Singh, G.K. (2013). Photovoltaic power generation technology using PV (photovoltaic power generation) technology: a review. Power, 53, 113.

[7] Knapič S, Malhi A, Saluja R, Främling K. Descriptive AI for web decision-making in the field of medicine. Machine learning and extracting information. 2021; 3 (3): 740770. https://doi.org/10.3390/make3030037

[8] Kuzlu, M., Cali, U., Sharma, V., & Güler, Ö (2020). Use descriptive artificial intelligence tools to get PV technology prediction details. Access IEEE, 8, 187814187823.

[9] Zeng, J., and Qiao, W. (2013). Short-term forecast of sunset using assist vector gadgets. Renewable Energy, 52, 118127.

[10] Jafarzadeh, S., Fadali, M.S., and Evrenosoglu, C.Y. (2012). Solar intensity forecast using C Type2TSK programming language model

IEEE Transactions on Sustainable Energy, 4 (2), 333339.

[11] Tang N, Mao S, Wang Y, and Nelms R M (2018). Predicting solar intensity technology with a LASSO-based holistic approach. IEEE Journal of Online Objects

[12] Khalyasmaa, A. 、 Eroshenko S. A. 、 Chakravarthy T. P. 、 Gas 、 V

G., Bollu S. K. Y., Caire, R., ... ＆ Karrolla 、 S. (2019 年 10 Predictability of Renewable Energy Production Based on the Random Deforestation Model. 2019 International MultiConference in Engineering, Computer and data Sciences (SIBIRCON) (p.07800785) .IEEE.

[13] M.R. Machado, S. Karray unddassi. T. de Sousa, LightGBM: A Good Way to Develop a Consumer Decision Predictability in the Financial Industry, 14th International Conference on Applied Science and Education (ICCSE), 2019, S ...

[14] T. Deng 、 Y 8-Zhao 、 S.Wang 、 H。 Yu, "Sales Forecast Supports LightGBM", IEEE 2021 International Conference on Consumer Electronics and Computer Engineering (ICCE), 2021, S. 383386, doi: 101109 / ICCECE51280.2021.934244.

[15] H. T. T. Nguyen, L. H. Chen, V.S.Saravanarajan 、 H.Q Pham, Student Behavior Predicting Using XG Boost and Random Forest Classification Algorithms, New Trends in 2021 Industry 4.0 (ETI 4.0), 2021, S. 15,

[16] K. Zhang, P. Xu und J. Zhang, "Descriptive AI in Deep-Based Learning Models: SHAP Techniques Used for Institutional Emergency Management", 2020 IEEE 4th Conference on Energy Internet and Energy System Integration (EI2), 2020, S. 711716.

[17] O. Ayoub, F Musumeci, F. Ezzeddine, C. Passera, and M. Tornatore, "Using Explained Computer Diagnostics Computer in Microwave Networks," 2022, 2022, 25 Cloud Conference, Internet, and Network Innovation. (ICIN), 2022, p. .4855, doi: 10.1109 / ICIN53892.2022.9758095

[18] Tanmay Srinath, H.S. Gururaja, Descriptive Machine in Automated Credit Card Identification, Land Reform Processes, 2022, ISSN 2666285X

[19] Brayman, Leo "Zufällige Wälder" Machine Learning 45.1 (2001): 532.

[20] I. L. Cherif no-A. Kortebi, "Using the eXtreme Gradient Boosting (XGBoost) Machine Learning Algorithm for Home Network Traffic Classification", 2019 Wireless Days (WD)), 2019, S. 16.

[21] Machado, Marcos Roberto, Salma Karray, and Ivaldo Tributino de Sousa. "LightGBM: An effective way to grow a decision tree to predict customer loyalty in the financial industry." 14th International Conference on Computer Science and Education (ICCSE) 2019. IEEE, 2019.