DOI: 10.56892/bima.v7i3.487

A Decision Tree Model For Predicting Household Primary Cooking Energy Sources

*[1,2]Tabra, M. S., [1]Ahmadu, A. S. and [1]Baha, B. Y.

[1]Department of Computer Science, Modibbo Adama University, Yola, Adamawa, Nigeria

[2]Department of Computer Science, Gombe State University, Gombe, Nigeria

*Corresponding Author: munattabra@gsu.edu.ng

ABSTRACT

Access to clean and efficient cooking energy sources is a critical concern for sustainable development and public health. In Nigeria, a diverse range of cooking energy sources are used, including biomass, kerosene, electricity, and liquefied petroleum gas. This paper employs a decision tree algorithm to classify the primary cooking energy sources used in Nigerian households. The decision tree model is trained and evaluated using a dataset obtained from the National Bureau of Statistics (NBS) database which was collected through the Nigerian General Household Panel Survey (NGHPS) conducted across various regions in Nigeria. The dataset underwent pre-processing and Variance Threshold was used to select the most relevant features. A dataset representing 4,980 households was used to build the Decision Tree Model. The model was built using Gini Impurity Node-Splitting Criteria together with Pruning using Cost Complexity Pruning Alpha (ccp_alpha) parameter within a range [0.0001, 0.001, 0.01, 0.1, 1, 10]. The study shows that certain household attributes can be used by the model to predict households' Primary Cooking Fuel Sources in Nigeria with an accuracy of 96%, precision of 95%, recall and F1 Score of 93% and 94% respectively. In future, this study can be improved by exploring other Machine Learning Techniques.

Keywords: Cooking Energy Sources, Decision Tree Algorithm, classification, Nigeria

INTRODUCTION

Climate change presents a significant global challenge. Household fuel choices contribute to climate change through the emission of greenhouse gases (Smith and Haigler, 2008). CO2 emissions (and fossil fuel consumption) correlate with mortality rates from cardiovascular disease, diabetes mellitus, cancer, and chronic respiratory disease (Kapsalyamova, Mishra, Kerimray, Karymshakov, and Azhgaliyeva, 2021). The use of solid fuels for cooking and heating poses serious health risks, especially for women and children causing high levels of indoor air pollution and acute respiratory infections (ARI) (Mishra, Retherford, and Smith, 2005).

Nigeria, a country with a population exceeding 200 million, stands as one of the world's most densely populated countries. It possesses abundant natural resources, including various household energy sources such as Liquefied Petroleum Gas, electricity, firewood, charcoal, animal dung, and kerosene (Maina, Kyari, and Maina, 2019a). In Nigeria, firewood is a prevalent cooking fuel, while kerosene is used for lighting (Ogwumike, Ozughalu, and Abiona, 2014). Due to inadequate electricity supply and high cost of Liquefied Petroleum Gas, the demand for dirty energy sources surpasses that of clean energy sources and this exerts a substantial environmental burden due to the country's large population (Maina, Kyari, and Maina, 2019b).

DOI: 10.56892/bima.v7i3.487

Access to reliable and sustainable cooking energy is a fundamental aspect of modern living, influencing both health outcomes and environmental impact. The United Nations, in its 17 Sustainable Development Goals, includes Item 7, aiming to ensure access to affordable, reliable, sustainable, and modern energy for all by 2030 (Kapsalyamova *et al.*, 2021). Addressing the transition from dirty to clean energy necessitates a comprehensive understanding of the energy sources used by households across the country.

In recent times, machine learning (ML) has emerged as a potent tool for technological advancement (Rolnick, Donti, Kaack, Kochanski, Lacoste, and Sankaran, 2019). Several studies have utilized machine learning algorithms to develop models in the energy sector (Muhammad, Hernan, Saqib, Fahid, Shahzaib and Abrar, 2022; Lee, Kim and Gu, 2023; Vijendar, Lakshmi, Poojitha, Naga, Krithika, and Sai, 2023).

These algorithms can be used to develop classification models. “Decision tree classifiers are regarded to be a standout of the most well-known methods to data classification” (Bahzad and Adnan, 2021). The use of these classifiers have been employed in many studies from diverse fields (Namazkhan, Albers and Steg, 2020; Ghiasi, and Zendehboudi, 2021; Shafi , Ramli and Awalin, 2021; Micheal, Marion and Ayodele, 2021; Jade, Enrico, Ruaina, Robert, Renann and John, 2023; Gupta, Gaur, Vashishtha, Das, Singh, and Hemanth, 2023).

For example, a study explored predictors of charcoal and firewood usage for cooking using the 2014 Uganda Census dataset (Nzabona, Tuyiragize, Asiimwe, Kakuba, and Kisaakye, 2021). They employed a Multinomial Logistic Regression model to discern factors influencing household charcoal and firewood usage in comparison to electricity. Additionally, a data-driven model based on Support Vector Machines (SVM) was formulated to predict lighting energy consumption in office buildings in Philadelphia, PA, yielding promising outcomes (Amasyali and El-Gohary, 2016).

Furthermore, Madhusudanan, (2019) undertook a survey comparing the accuracy of machine learning-supported regression models with traditional regression models for forecasting energy consumption. Moreover, a Support Vector Regression-based model was introduced to forecast household electricity usage under multiple behavioural intervention strategies, assisting in decision-making for diverse households (Shen, Sun and Lu, 2017). Another study by Khan, Byun, Lee, Kang and Kang, (2020) proposed a hybrid machine learning model, integrating multi-layered perceptron, support vector regression, and CatBoost, to predict energy consumption by utilizing load data from renewable and non-renewable energy sources.

Additionally, a Neural Network-driven model was contrasted with a conditional demand analysis (CDA) approach for modelling residential end-use energy consumption in Canada's residential sector (Koksal, Ugursal, and Fung, 2002). The study found that Neural Networks outperformed CDA, especially in evaluating the influence of socioeconomic factors on energy consumption patterns.

A study by Sathiyanarayanan, Pavithra, Saranya, and Makeswari utilized the Decision Tree algorithm to detect breast cancer cases. Their research was focused on the identification of breast cancer and involved a comparison of the performance between the K-Nearest Neighbors (KNN) and Decision Tree (DT) algorithms. The findings revealed that KNN achieved an accuracy rate of 97%, while the Decision Tree algorithm achieved a maximum accuracy rate of 99%.

A novel Intrusion Detection System (IDS) was introduced by Ahmim, Maglaras, Ferrag, Derdour, and Janicke, (2019). The IDS combines multiple classification techniques based on Decision Tree (DT) and rule-based principles and utilizes the REP tree, JRip algorithm, and Forest PA to classify network traffic into either Attack or Benign categories.

In another study by Ramadhan, Sukarno and Nugroho (2020), a Decision Tree algorithm was compared with K-Nearest Neighbor (KNN) for detecting DDoS attacks. The Decision Tree algorithm outperformed KNN in detecting attaining an accuracy of 99.91%, as opposed to KNN's accuracy of 98.94%.

In another study by Ramadhan, Sukarno and Nugroho (2020), a Decision Tree algorithm was compared with K-Nearest Neighbor (KNN) for detecting DDoS attacks. The Decision Tree algorithm outperformed KNN in detecting attaining an accuracy of 99.91%, as opposed to KNN's accuracy of 98.94%.

Energy Sources

Energy sources encompass natural resources capable of providing heat, light, or power either directly or indirectly through conversion or transformation. These sources are classified as either Clean Energy Sources or Dirty Energy Sources. Clean energy sources encompass Liquefied Petroleum Gas (LPG) and electricity derived from renewable sources like hydro, solar, wind, and other consistently replenished sources. These energy alternatives exhibit minimal harmful carbon dioxide (CO2) emissions, which are recognized to have adverse environmental effects (Maina, Kyari, and Maina, 2019a).

In contrast, the term "dirty energy" pertains to energy production methods that contribute to climate change and adversely impact communities, particularly in regions of the global south (Friends of the Earth Africa, 2016). Numerous dirty energy projects involve the extraction or combustion of fossil fuels (such as oil, gas, and coal) for electricity generation, thereby releasing carbon dioxide into the atmosphere—a principal catalyst of climate change. Burning fossil fuels, however, sends greenhouse gases into the atmosphere, trapping the sun’s heat and contributing to global warming (National Renewable Energy Laboratory (NREL), 2001).

Decision Tree

Decision Tree is one of the state-of-the-art Supervised Machine Learning algorithms used for both classification and regression tasks. It is a hierarchical tree structure comprising the root node, the internal nodes, the branches and the leaf nodes as depicted in Figure 1.

Each path from the root node that passes through internal nodes and ends at a leaf node represents a classification decision rule. For a classification task, the key decision points in the tree algorithm pertain to the criteria for splitting nodes and pruning the tree. (Hastie, Tibshirani, and Friedman, 2008).

Decision trees work by selecting the most suitable features to divide the data at each node. Various criteria such as Gini impurity, entropy, and information gain are employed to assess the quality of these splits (Hastie, Tibshirani, and Friedman, 2008).

As a decision tree model becomes more complex, its reliability in predicting future records decreases. An alternative approach to constructing a decision tree model involves initially growing a large tree and subsequently pruning it to an optimal size by eliminating nodes that contribute less additional information (Song and Lu, 2015). Decision tree classifiers offer advanced tree pruning techniques, resulting in outcomes that are transparent and interpretable (Bahzad and Adnan, 2021).

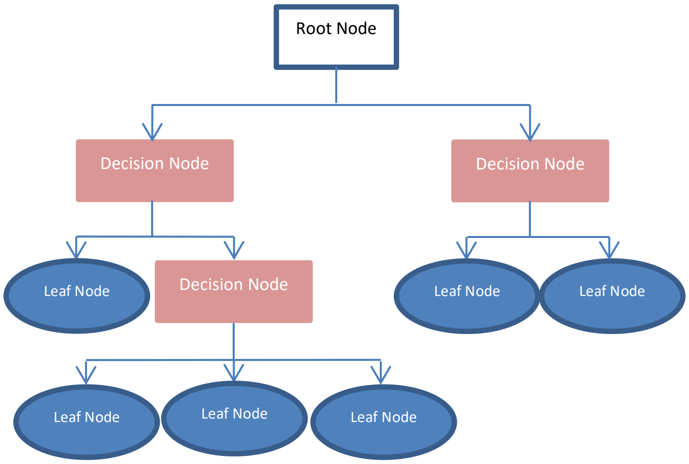

Figure 1: A decision Tree Structure

MATERIALS AND METHODS

Dataset

In this research, we used secondary Data acquired from the NBS database. The data was gathered via the Nigerian General Household Panel Survey (NGHPS), conducted by the NBS. Data collection involved face-to-face interviews with members of various households. The interview covered a diverse range of topics, including household demographics, socio-economic status, and fuel expenditure practice. The dataset comprises records from a total of 4,980 households.

Research Architecture

The study architecture is shown in Figure 2 representing both the training and testing phases. The training steps consist of data Pre-processing and Feature selection; later the selected set of features was used to train the Decision Tree classifier. The testing phase is the process of matching the actual target value with the value predicted by the trained classifier.

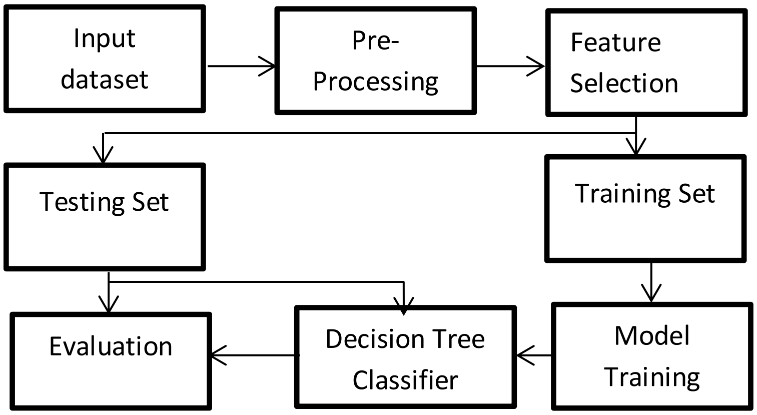

Figure 2: Research Architecture

DOI: 10.56892/bima.v7i3.487

Data Pre-Processing

The dataset underwent preprocessing to handle missing values, where imputation techniques were used to fill the missing values and later the dataset was standerdized within a scale range (0,1). Feature selection was conducted using the Variance Threshold Feature Selection Technique to select the most relevant features in the dataset to prepare the data for model training.

Feature Selection

Variance Threshold Feature Selection Technique was to select the most relevant set of features in the dataset. The selected Features using are Zone, State, Local Government Area (LGA), Enumeration Area (EA), Size, Age of Household Head (HHHAge), Age of Spouse (SpouseAge), Education Level of Household Head (EduHHH), Education of Spouse (EduSpouse), Occupation of Household Head (OccuHHH), Occupation of Spouse (OccuSpouse), Total Earning from Occupation (EarningOccu), Total Earning from Occupation Allowance (EarningOccuAllowc), Income Generating Activity (IncomeGenerate), Monetary Assistance from Nigeria (MonAssistNigeria), Monetary Assistance from Abroad (MonAssistAbroad), Other Income (OtherIncome), Total Income (TotalIncome), Expenditure on Charcoal (CharcoalAmount), Expenditure on Gas (GasAmount), Expenditure on Electricity (ElectAmount), Expenditure on Firewood (FWoodAmount) and Expenditure on Kerosene (KeroAmount) were considered totalling 23 features in the dataset.

Model Evaluation

The trained decision tree model was evaluated using a confusion Matrix and various performance metrics were used; namely, accuracy, precision, recall, and F1-score. The used metrics were estimated as in equations 1-4.

$$Accuracy\ =\ \left(TP+TN\right)/\left(PN\right) \tag{1}$$

$$Precision\ =\ TP/\left(TP+FP\right) \tag{2}$$

$$TPR=\ TP/P\ or\ TP/\left(TP+FN\right) \tag{3}$$

$$F1\,Score\ =\ 2*(Precision*Recall)\,/\,(Precision+Recall) \tag{4}$$

RESULTS AND DISCUSSION

The classification process started by generating the Decision Tree using two different node splitting criteria namely Entropy and Gini Impurity, and the results are represented and compared in Table 1 and in Figure 3 respectively.

Table 1: Performances of Decision Tree using Entropy and Gini Impurity Node Splitting

Criteria	Accuracy	Precision	Recall	F1 Score
Entropy	0.9639	0.9198	0.9338	0.9262
Gini	0.9639	0.9388	0.9454	0.9414

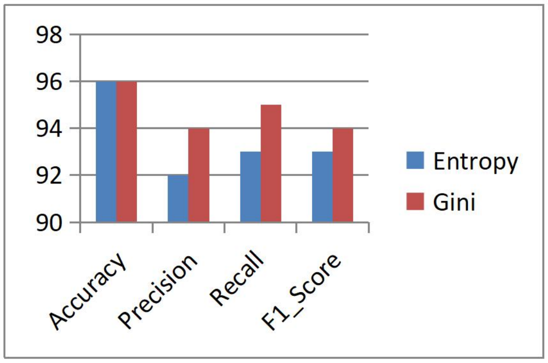

Figure 3: Comparison of Performances of Decision Tree Model using Entropy and Gini Impurity Node Splitting Criteria.

The two models attained the same accuracy of 96.39% while the model generated using the Gini Impurity node splitting criteria outperformed in Precision, recall and F1 Score.

Table 2: 5-Fold Crossed validated ccp_alpha values and their corresponding means scores.

ccp_alpha	Mean_Score
0.0001	0.8772 ± 0.1022
0.001	0.9159 ± 0.0924
0.01	0.9111 ± 0.0156
0.1	0.7969 ± 0.0080
1	0.5311 ± 0.0005
10	0.5311 ± 0.0005

The process went further to reduce the Decision Tree complexity to increase reliability by applying the pruning technique. To find the most appropriate Prunning "ccp_alpha" value for the problem, 5- Fold cross-validation approach was used to test the performances of the Decision Tree at varying ccp_alpha values [0.0001, 0.001, 0.01, 0.1, 1, 10]. The cross-validated result is shown in Table 2.

The highest mean score was obtained with the pruning ccp_alpha value 0.001 and therefore was used in addition to the Entropy and Gini Impurity Node splitting criteria to train the Decision Tree Classifier. The classifier based on Entropy criteria attained an accuracy of 96%, with a precision of 93%, recall of 93%, and F1-score of 93%; while an accuracy of 96%, with a precision of 95%, recall of 93%, and F1-score of 94% was attained using the Gini Impurity splitting criteria. The performances are represented in Table 3 and compared in Figures 4 and 5.

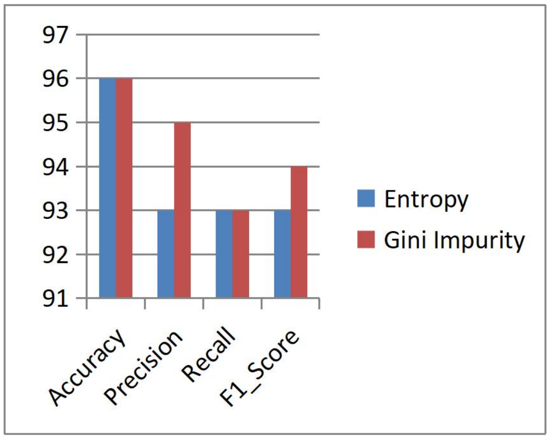

Figure 4: Comparison of the Performances of Decision Tree using Entropy and Gini Impurity Node Splitting using Pruning ccp_alpha value 0.001.

Table 3: Performances of DT using Entropy and Gini Impurity Node splitting criteria with Prunning ccp_alpha value 0.001.

	Pruning ccp_alpha value 0.001			
	Accuracy	Precision	Recall	F1_Score
Entropy	0.9649	0.9314	0.9337	0.9323
Gini	0.9639	0.95	0.9303	0.9394

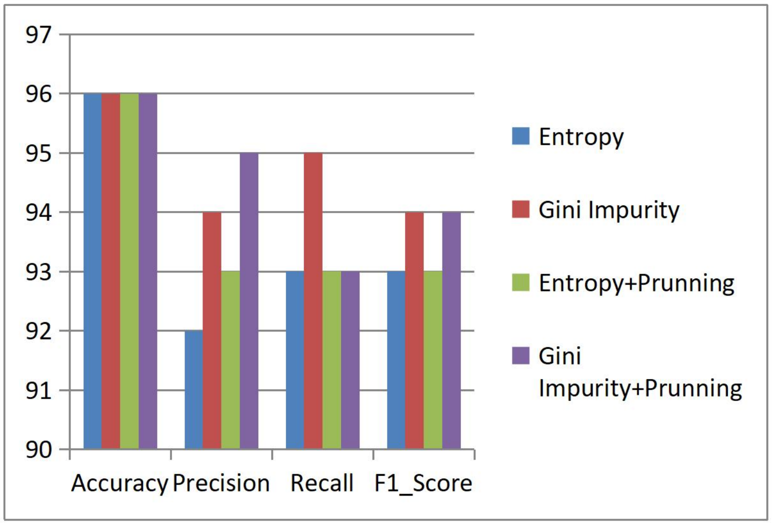

Figure 5: Comparison of the Performances of Decision Tree using Entropy and Gini Impurity Node Splitting without Prunning and using Pruning with ccp_alpha value 0.001.

This study used certain household characteristics from Nigerian households to develop a Decision Tree model for classification of households' primary cooking energy sources. From the experiment conducted in this study, the decision tree model trained using the pruning ccp_alpha value 0.001 together with the Entropy Splitting Criteria gives an accuracy of 96%, with a precision of 93%, recall of 93%, and F1-score of 93%; while accuracy of 96%, with a precision of 95%, recall of 93%, and F1-score of 94% was attained with the Gini Impurity Splitting Criteria. The highest model's performance was obtained using the Gini Impurity attribute Splitting Criteria with a pruning ccp_alpha value of 0.001. The model's accuracy was found to be 96%, with a precision of 95%, recall of 93%, and F1-score of 94%. The model shows that household family Size, Age of the household head and the fuel expenditure practice features are the major predictors of the Primary Cooking fuel.

The result from this study has a practical implication, as it will help the policymakers as well as energy companies to know the locations in the country where much attention is needed to provide effective strategies to promote the use of cleaner and more sustainable cooking fuels.

CONCLUSION

Knowledge of household cooking energy is imperative for an efficient switch from dirty to clean energy, to attain sustainable development goals. This study developed a Decision Tree model for prediction of households cooking energy sources. Based on the procedures employed, the Decision Tree Algorithm exhibited promising classification performance on the dataset. The model's accuracy was found to be 96%, with a precision of 95%, recall of 93%, and F1-score of 94%. This implies that given certain household attributes, the Decision tree model can classify the household's primary cooking energy source with an accuracy of 96%. Though there is still a need for a model with better performance, the model developed in this study has the potential to predict the households' primary cooking energy sources in Nigeria with a reasonable level of accuracy. In future, this study can be improved by exploring the use of other machine learning classification Algorithms.

REFERENCES

Ahmim, A., Maglaras, L., Ferrag, M. A., Derdour, M. and Janicke, H. (2019). A novel hierarchical intrusion detection system based on decision tree and rules-based models. *15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 228–233.

Amasyali, K. and El-Gohary, N. (2016). Building lighting energy consumption prediction for supporting energy data analytics. *International Conference on Sustainable Design, Engineering and Construction*, 145, 511-517.

Bahzad, T. J., and Adnan, M. A. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. *Journal of Applied Science and Technology Trends*. 2(1), 20-28.

Friends of the Earth Africa, (2016). *Dirty Energy in Africa.* Friends of the Earth International, Amsterdam. www.foei.org.

Ghiasi, M. M., Zendehboudi S. (2021). Application of Decision Tree-Based Ensemble Learning in the Classification of Breast Cancer. *Comput Biol Med. 128:104089 PMID: 33338982.*

Gupta, V., Gaur, H., Vashishtha, S., Das, U., Singh, V.K. and Hemanth, D.J. (2023). A fuzzy rule-based system with decision tree for breast cancer detection. *IET Image Process*. 17, 2083–2096.

Hastie, T., Tibshirani, R. and Friedman, J. (2008). *The Elements of Statistical Learning, Data Mining, Inference and Prediction.* Stanford, California: Springer Series in Statistics.

Jade V. Y. C., Enrico R. C. C., Ruaina, L. H. G., Robert, K. C. B., Renann, G. B. and John, C. V. P. (2023). A Decision Tree-Based Classification of Fetal Health using Cardiotocograms. *AIP Conference Proceedings*. 2562(1), 020003.

Kapsalyamova, Z., Mishra, R., Kerimray, A., Karymshakov, K. and Azhgaliyeva, D. (2021). Why is Energy Access Not Enough for Choosing Clean Cooking Fuels? Sustainable Development Goals and Beyond. *Asian Development Bank Institute,* 1234, 1-30.

Khan, P. W., Byun, Y., Lee, S., Kang, D. and Kang, J. (2020). Machine Learning-Based Approach to Predict Energy Consumption of Renewable and Non-renewable Power Sources. *Energies,* 13(18) 4870, 1-16.

Koksal, M., Ugursal, V. and Fung, A. (2002). Modeling of the appliance, lighting, and space-cooling energy consumptions in the residential sector using neural networks. *Applied Energy*, 71(2), 87-110.

Lee H., Kim, D., and Gu, J. (2023). Prediction of Food Factory Energy Consumption Using MLP and SVR Algorithms. *Energies*, *16*(3), 1550.

Madhusudanan, C. R. (2019). *A Machine Learning Framework for Energy Consumption Prediction.* All Theses, Tiger Prints Clemson University, Clemson, South Carolina. 3184. https://tigerprints.clemson.edu/all_theses/3184.

Maina, Y. B., Kyari, B. G. and Maina, M. B. (2019a). Households' Clean Energy Demand in Nigeria and Its Implication on the Environment. *Journal of Agricultural Economics, Environment and Social Sciences*, 5(1and2), 35-45.

Maina, Y. B., Kyari, B. G. and Maina, M. B. (2019b). Socio-Economic Determinants of Household Dirty Energy Use in Nigeria. *Journal of Agricultural Economics, Environment and Social Sciences, 5*(1 and 2), 64 - 71.

Micheal, O. A., Marion, O. A. and Ayodele, A. A.(2021). A Genetic Algorithm Approach for Predicting Ribonucleic Acid Sequencing Data Classification using KNN and Decision Tree. *TELKOMNIKA Telecommunication, Computing, Electronics and Control. 19(1) 310-316.*

Mishra, V., Retherford, R. D. and Smith, K. R., (2005). Cooking Smoke and Tobacco Smoke as Risk Factors for Stillbirth. *International Journal of Environmental Health Research,* 15(6), 397-410.

Muhammad, R. A., Hernan, A. G. R. , Saqib. H., Fahid, R., Shahzaib, H. and Abrar, H. (2022). Machine Learning-Based Prediction of Specific Energy Consumption for Cut-Off Grinding. *Sensors*, *22*(19), 7152.

Namazkhan, M., Albers, C., and Steg, L. (2020). A Decision Tree Method for Explaining Household Gas Consumption: The Role of Building Characteristics, Socio-Demographic Variables, Psychological Factors and Household Behaviour. *Renewable and Sustainable Energy Reviews, 119, 109542.*

National Renewable Energy Laboratory (NREL), (2001). *Renewable Energy: An Overview.* U.S. Department of Energy (DOE), Golden, Colorado. 1-8. https://www.nrel.gov/docs/fy01osti/27955.pdf

Nzabona, A., Tuyiragize, R., Asiimwe, J. B., Kakuba, C. and Kisaakye, P. (2021). Urban Household Energy Use: Analyzing Correlates of Charcoal and Firewood Consumption in Kampala City, Uganda. *Journal of Environmental and Public Health,* Vol. 2021 1-8.

Ogwumike, F. O., Ozughalu, U. and Abiona A. (2014). Household Energy Use and Determinants: Evidence from Nigeria. *International Journal of Energy Economics and Policy*, 4(2), 248-262.

Ramadhan, I., Sukarno, P. and M. A. Nugroho, M. A. (2020). Comparative Analysis of K-Nearest Neighbor and Decision Tree in Detecting Distributed Denial of Service. *8th International Conference on Information and Communication Technology (ICoICT), Yogyakarta, Indonesia,* 1-4.

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A. and Sankaran, K. (2019). Tackling Climate Change with Machine Learning, Cornell University, Ithaca, New York, 1-111. *arXiv:1906.05433v2 [cs.CY] 5.*

Sathiyanarayanan, P., Pavithra, S., SARANYA, M. S., and Makeswari. M. (2019). Identification of Breast Cancer Using The Decision Tree Algorithm. *IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), Pondicherry, India*, 1-6.

Shafi, M. K. M., Ramli, N. A. and Awalin, L. J. (2021). Energy consumption prediction by using machine learning for smart building: Case study in Malaysia. *Developments in the Built Environment 5 (2021) 100037*.

Shen, M., Sun, H. and Lu, Y. (2017). Household Electricity Consumption Prediction Under Multiple Behavioural Intervention Strategies Using Support Vector Regression. *9th International Conference on Applied Energy, ICAE2017.* Cardiff, UK. 142, 2734-2739.

Smith, K. R. and Haigler, E. (2008). Co-Benefits of Climate Mitigation and Health Protection in Energy Systems: Scoping Methods. *Annual Review of Public Health,* 29(1), 11-25.

Song, Y. and Lu. Y. (2015). Decision Tree Methods: Applications for Classification and Prediction. *Shanghai Archives of Psychiatry*, 27(2), 130-135.

Vijendar, G. R., Lakshmi, J. A., Poojitha, C., Naga, A. S., Krithika D. R. , and Sai, G. M. (2023). Electricity Consumption Prediction using Machine Learning. *E3S Web of Conferences 391, 048 ICMED-ICMPC.*