International Journal of Scientific Research in Science, Engineering and Technology

Print ISSN - 2395-1990
Online ISSN : 2394-4099

OPEN ACCESS

Available Online at : www.ijsrset.com
doi : https://doi.org/10.32628/IJSRSET231069

A Study of Rule Extraction from Double Random Forest to Identify the Characteristics of Working Poor in Jakarta Province, Indonesia

Adlina Khairunnisa[1], Khairil Anwar Notodiputro*[2], Bagus Sartono[3]

[1,2,3]Department of Statistics, IPB University, Bogor, Indonesia

adlina.khairunnisa@bps.go.id[1]

ARTICLEINFO

Article History :

Accepted: 10 Nov 2023
Published: 30 Nov 2023

Publication Issue :
Volume 10, Issue 6
November-December-2023
Page Number :
258-266

ABSTRACT

Double Random Forest (DRF) outperforms Random Forest (RF) models, particularly when the RF model is underfitting. DRF generates more diverse and larger trees that significantly improve prediction accuracy. By applying association rule technique, the extracted rules from the DRF model provide an easily understandable interpretation of the characteristics of individuals identified as the working poor in Jakarta. The findings show that DRF performs good predictive performance in classifying poor workers in Jakarta, achieving an AUC value of 79.02%. The extracted rules from this model highlights interactions between education levels, working household member proportion, and job stability that significantly affect the classification of working poor. Specifically, worker with lower education levels, particularly high school or below, show a higher probability of being classified as poor workers. In addition, households with fewer employed members, especially those involving worker in self-employed/employee/freelancer roles, face a greater risk of falling into the poor category due to job instability and limited workforce participation. This implies that the interaction between the low proportion of working household members and low education, the interaction between unstable job position and low proportion of working household members, and the interaction between low education and unstable job position are the most important characteristics of the working poor in Jakarta.

Keywords: Association Rules, Double Random Forest (DRF), Employment, Poverty, Rule Extraction

I. INTRODUCTION

Double Random Forest (DRF) represents a new ensemble tree model similar to the Random (RF). It outperforms RF in predictive performance, particularly when the RF model is underfitting [1]. Underfitting occurs when the relative test accuracy

value is less than one, which means the size of the RF tree might not be large enough to provide optimal performance. Relative test accuracy is the accuracy result when a certain nodesize is divided by the accuracy when the nodesize is 1.

Unlike RF, DRF uses the complete training dataset instead of bootstrapping. Leveraging the entire training data leads to many unique observations at the nodes which makes the tree larger compared to the RF. To obtain the optimal splitting rules, DRF uses bootstrap sampling and selects random variable subsets at each splitting node. These processes introduce randomness into the tree building process, resulting in diverse trees.

While DRF demonstrates good predictive performance, the resulting models tend to be challenging to interpret. Model interpretation is important in explaining the contribution of each variable in decision-making. Throughout the decision-making process, the model must prove its accuracy through explanations that are understandable to humans [2]. Clear explanations regarding the model's outcomes are necessary to increase confidence levels in the generated predictions. Additionally, model interpretability provides insight into its predictive mechanisms [3].

The use of association rules becomes instrumental in facilitating model interpretation. This data mining technique aims to identify relationships among variable combinations [4]. The inTrees approach [5] applies association rule techniques by extracting rules from all decision trees and identifying prevalent variable-value pairs. This method can be applied to issues related to the working poor.

The working poor refers to individuals who work and live in poor households. This definition encompasses two aspects: the individual's classification as either working or non-working, and the household's classification as poor or non-poor. The definition of working poor highlights the relationship between employment status and poverty status. It refers to someone who works and lives in a poor household [6]. According to [7], the working poor are individuals who meet the criteria as workers while also being in a state of poverty.

Factors contributing to the working poor status extend beyond low income. Cheung and Chou [8] identified three contributing factors in Hong Kong, focusing on individual-levels variables, employment-related variables, and household-related variables. Faharuddin and Endrawati [9] adapted these factors in the context of working poor in Indonesia. The results reveal characteristics such as low education levels, younger or older age brackets, unmarried status, limited internet access, rural residency, and engagement in unstable employment. From a household perspective, the working poor mostly originate from households with numerous members, but only a few are employed, and they do not receive economic support from external sources. Another study in Indonesia found that the working poor typically consist of males, rural residents, and individuals with lower education indicate that the working poor tend to have low levels of education [10].

The National Socioeconomic Survey (Susenas) in March 2022 indicated Jakarta has relatively low poverty rate at 4.69%, contrasting with a higher open unemployment rate of 7.18%. Although many unemployed people in Jakarta can still fulfill their basic needs, a high unemployment rate has the potential to increase the poverty rate. It is crucial to identify employed individuals struggling to meet daily basic needs to prevent further increase in poverty. Additionally, it can prevent unemployed individuals from falling into the poor category.

To identify the characteristics of the working poor in Jakarta Province, the association rule approach is used to extract rules. This analysis aims to understand the variables that affect the poverty status of workers in

Jakarta. By analyzing the characteristics of the working poor, this analysis is expected to provide insights into interpreting the model and identifying crucial factors contributing to the poverty status of workers in this province.

II. METHODS AND MATERIAL

Empirical Data

The empirical data used in this study is sourced from the results of the National Socioeconomic Survey (Susenas) 2022 conducted in DKI Jakarta. The working poor comprises two concepts: poverty and employment. Poverty is measured using the concept of the ability to meet basic needs, including both food and non-food necessities [11]. An individual with an average per capita expenditure below the poverty line is categorized as poor. The poverty line indicates the minimum monetary value required for an individual to fulfill their basic needs for one month, covering both food and non-food necessities. Meanwhile, individuals considered as working are those aged 15 years and above, engaged in activities to earn or assist in earning income for at least one uninterrupted hour (continuously) per week, or who have a job but did not work due to holidays, leave, illness, and similar reasons [12]. Therefore, working poor denotes someone who is employed but lives in a household below the poverty line. The variables used are based on previous studies [8], [9].

Table 1. Variables used in this study

Code	Variables	Description	Scale
Dependent variable			
Y	Poverty status of worker	• Poor worker • Non-poor worker	Nominal
Individual-level variables			
X_1	Age	Age of worker	Ratio
X_2	Gender	1 : Male 2 : Female	Nominal

Table 1. Variables used in this study (cont.)

Code	Variables	Description	Scale
Individual-level variables			
X_3	Marital status	1 : Never married 2 : Married 3 : Divorce 4 : Widowed	Nominal
X_4	Educational level	1 : No education 2 : Primary school 3 : Secondary school 4 : High school 5 : University	Ordinal
X_5	Place of birth	1 : Jakarta 2 : others	Nominal
X_6	Residence of 5 years ago	1 : Jakarta 2 : others	Nominal
Individual-level variables			
X_7	Literacy ability	1 : Able 2 : Unable	Nominal
X_8	Functional disability	1 : Exists 2 : Not exist	Nominal
X_9	Internet use	1 : Use internet 2 : Not use internet	Nominal
Employment-related variables			
X_{10}	Job sector	1 : Agriculture 2 : Mining and quarrying 3 : Construction 4 : Industry 5 : Electricity, gas and water	Nominal

Table 1. Variables used in this study (cont.)

Code	Variables	Description	Scale
Employment-related variables			
X_{10}	Job sector	6 : Trade, accomodation and restaurants 7 : Transport and communicati-on 8 : Other services	Nominal
X_{11}	Working hours	Weekly working hours	Ratio
X_{12}	Employme nt status of worker	1 : Self-employed 2 : Employer with unpaid worker 3 : Employer with paid worker 4 : Employee 5 : Freelancer 6 : Family worker/unpa id worker	Nominal
Household-level variables			
X_{13}	Proportion of working household member	The proportion of working to the total household member	Ratio
X_{14}	Home ownership	1 : Own a home 2 : others	Nominal
X_{15}	Access to credit	1 : Has access 2 : No access	Nominal

Rule Extraction from DRF Model

The DRF procedure involves several steps. Initially, decision trees are built using all training dataset with p variables. During the splitting process, bootstrap sampling is performed from the training data if the observations within a node exceed 10% of the total observations, otherwise, the original data is used. At each node, a random variable subset of approximately $\sqrt{p}$ variables is selected from the bootstrap data to determine the best split. This iterative process creates k decision trees, with their predictions aggregated using majority voting to get the final prediction of response variable class.

Subsequently, rules are extracted from the decision trees within DRF by following the path from the root node to the leaf node. The number of rules depends on the number of leaf nodes in each tree. The rules are typically formed in the X ⇒ Y, which X represents the condition and Y represents the prediction. Figure 1 illustrates the rule extraction process from DRF.

Based on Figure 1, the rules extracted from the first decision tree are as follows:

1. The first leaf node extracts the rule {$X_{13} \leq 0.45$, $X_{14} = 2$, and $X_3 = 1$ ⇒ The worker's status is poor}
2. The second leaf node extracts the rule {$X_{13} \leq 0.45$, $X_{14} = 2$, and $X_3 = (2,3,4)$ ⇒ The worker's status is not poor}
3. The third leaf node extracts the rule {$X_{13} \leq 0.45$ and $X_{14} = 1$ ⇒ The worker's status is poor}
4. The fourth leaf node extracts the rule {$X_{13} > 0.45$ ⇒ The worker's status is not poor}

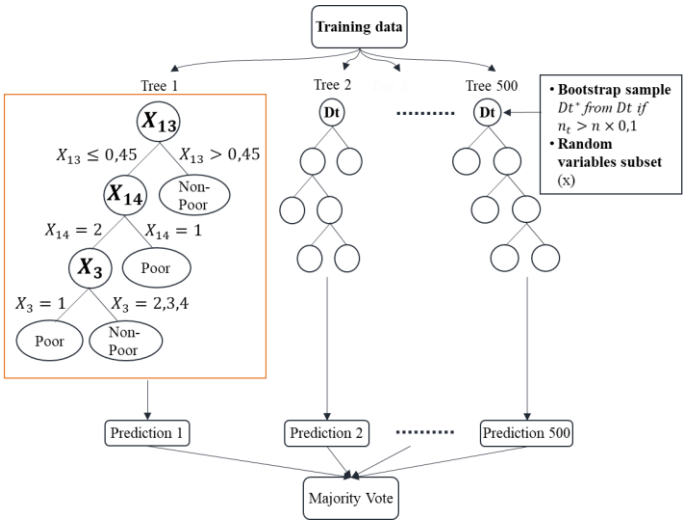

Figure 1: Rule extraction process of DRF

Interpreting DRF using association rules enhances its comprehensibility, which improves the credibility of its predictions. These rules are evaluated based on their support and confidence values. Support indicates the frequency a rule occurs among all generated rules, while confidence measures the certainty of a prediction outcome if the rule conditions are met. Furthermore, a rule's complexity is determined by its length, defined as the count of variable-value pairs within the rule's condition [5]. For example, a length value of 2 indicates frequent interaction between 2 variables. Figure 1 shows that rules 1 and 2 have the same length, which is 3. Rule 3 has a length of 2, while rule 4 has a length of 1.

III. RESULTS AND DISCUSSION

The empirical data used in this study focuses on the poverty status of workers in Jakarta Province. The dataset analyzed includes 7,773 workers. Before constructing the classification model, an identification was conducted to identify underfitting in the RF model concerning poverty status of workers. Findings revealed the presence of underfitting, prompting the adoption of DRF to enhance the RF model's performance. Optimal parameter values were determined through a 10-fold cross-validation technique, resulting in 1000 trees and a nodesize of 2.

Subsequently, the predictive performance of the DRF model was evaluated using test data, resulting in an AUC (Area Under the Curve) value of 79.02%. This AUC value indicates that the DRF model performs adequately in classifying poor workers in Jakarta [13]. In addition, sensitivity and specificity values are also considered in evaluating the predictive performance of a model. The sensitivity value indicates that 79.04% of the poor workers in Jakarta are predicted as poor. The specificity value indicates that 70.23% of non-poor workers in Jakarta are predicted as non-poor (Table 2).

Table 2. Performance evaluation of DRF model

AUC	Sensitivity	Specificity
79,02%	79,04%	70,23%

The rules extracted from this model resulted in 44,468 rules. Out of which 214 unique rules were identified to explain the variable interactions. Among these, 187 rules predicted the working poor, while 27 predicted non-working poor. Variable interactions in this study focused on high support values with $length \geq 2$. The analysis identified 125 rules predicting the working poor in Jakarta, as shown in Table 3. To characterize the working poor in Jakarta, rules with confidence values exceeding 95% were selected and sorted based on high support values.

Table 3. Number of unique rules to predict working poor in Jakarta

Length	Number of rules predicting working poor	Number of unique rules
1	57	70
2	113	131
3	12	13
Total	187	214

Table 4 shows the extracted rules from the DRF model based on the highest support values and confidence

values above 95%. These rules represent the interactions among different variables that indicate a worker's probability of being categorized as poor in the Jakarta Province, based on the support values and the confidence levels associated with each rule. The confidence value represents the probability of a worker being classified as poor based on the explanatory variable interactions. The support value indicates the quantity of evidence that validates the correctness of these variable interactions.

Table 4. Characteristics of poor workers in Jakarta based on rule extracted from DRF

Rule	Condition	*Sup*	*Conf*
1	$X_{13} \leq 0.45$ & $X_4 = (2,3,4)$	0.07	0.96
2	$X_{13} \leq 0.45$ & $X_8 = 1$	0.05	0.97
3	$X_{12} = (1,4,5)$ & $X_{13} \leq 0.45$	0.03	0.96
4	$X_{13} \leq 0.45$ & $X_{15} = 2$	0.03	0.97
5	$X_5 = 2$ & $X_9 = 2$	0.03	0.98
6	$X_{11} \leq 39.5$ & $X_4 = (2,3,4)$	0.03	0.98
7	$X_1 = (1,4,5)$ & $X_4 = (2,3,4)$	0.03	1.00
8	$X_4 = (2,3,4)$ & $X_8 = 1$	0.02	1.00
9	$X_{14} = 2$ & $X_4 = 2$	0.02	1.00
10	$X_{13} \leq 0.45$ & $X_4 = 3$	0.02	0.97
11	$X_{14} = 2$ & $X_3 = (1,2)$ & $X_4 = (2,3,4)$	0.02	0.97
12	$X_{11} \leq 39.5$ & $X_9 = 2$	0.02	0.97
13	$X_2 = 2$ & $X_4 = (2,3,4)$	0.02	0.97
14	$X_{12} = (1,4,5)$ & $X_{15} = 2$	0.01	0.96
15	$X_3 = 2$ & $X_8 = 1$	0.01	0.96
16	$X_{14} = 2$ & $X_3 = 2$ & $X_9 = 2$	0.01	0.96
17	$X_{13} \leq 0.45$ & $X_{15} = 2$ & $X_4 = (2,3,4)$	0.01	0.96
18	$X_{11} \leq 39{,}5$ & $X_{12} = (1,4,5)$ & $X_{13} \leq 0{,}45$	0.01	0.96
19	$X_3 = 2$ & $X_5 = 2$	0.01	0.95
20	$X_3 = 2$ & $X_4 = (2,3,4)$	0.01	0.95
21	$X_{10} = (1,7,8)$ & $X_{14} = 2$ & $X_4 = (2,3,4)$	0.01	0.95
22	$X_{13} \leq 0.45$ & $X_{14} = 1$ & $X_4 = (2,3,4)$	0.01	0.95

Here is the explanation of the rules presented in Table 4, which show the characteristics of workers categorized as poor in the Jakarta Province based on the highest support values with confidence above 95%:

1. If the proportion of working household members is less than or equal to 0.45 and the highest education level is primary/secondary/high school, the probability of worker being categorized as poor is 97% ($Confidence = 0.965$). This rule exists in 7% of the total extracted rules ($Support = 0.072$).
2. If the proportion of working household members is less than or equal to 0.45 and has functional disability, the probability of worker being categorized as poor is 97%. This rule exists in 5% of the total extracted rules.
3. If the worker's status is self-employed/employee/freelancer and the proportion of working household members is less than or equal to 0.45, the probability of worker being categorized as poor is 96%. This rule exists in 3% of the total extracted rules.
4. If the proportion of working household members is less than or equal to 0.45 and lacks access to credit, the probability of worker being categorized as poor is 97%. This rule exists in 3% of the total extracted rules.
5. If worker was not born in Jakarta and does not use the internet, the probability of worker being categorized as poor is 98%. This rule exists in 3% of the total extracted rules.
6. If someone works less than or equal to 39.5 hours per week and the highest education level is primary/secondary/high school, the probability of worker being categorized as poor is 98%. This rule exists in 3% of the total extracted rules.
7. If the worker's status is self-employed/employee/freelancer and the highest education level is primary/secondary/high school, the probability of worker being categorized as poor is 100%. This rule exists in 3% of the total extracted rules.

8. If the worker's highest education level is primary/secondary/high school and has functional disability, the probability of worker being categorized as poor is 100%. This rule exists in 2% of the total extracted rules.
9. If worker does not own a home and the highest education level is primary school, the probability of worker being categorized as poor is 100%. This rule exists in 2% of the total extracted rules.
10. If the proportion of working household members is less than or equal to 0.45 and the highest education level is secondary school, the probability of worker being categorized as poor is 97%. This rule exists in 2% of the total extracted rules.
11. If worker does not own a house, the marital status is never married/married, and the highest education level is primary/secondary/high school, the probability of worker being categorized as poor is 97%. This rule exists in 2% of the total extracted rules.
12. If someone works less than or equal to 39.5 hours per week and does not use internet access, the probability of worker being categorized as poor is 97%. This rule exists in 2% of the total extracted rules.
13. If the worker is female and the highest education level is primary/secondary/high school, the probability of worker being categorized as poor is 97%. This rule exists in 2% of the total extracted rules.
14. If the worker's status is self-employed/employee/freelancer and lacks access to credit, the probability of worker being categorized as poor is 96%. This rule exists in 1% of the total extracted rules.
15. If worker's marital status is married and has functional disability, the probability of worker being categorized as poor is 96%. This rule exists in 1% of the total extracted rules.
16. If someone who works does not own a house, the marital status is married, and does not use internet access, the probability of worker being categorized as poor is 96%. This rule exists in 1% of the total extracted rules.
17. If the proportion of working household members is less than or equal to 0.45, lack access to credit, and the highest education level is primary/secondary/high school, the probability of worker being categorized as poor is 96%. This rule exists in 2% of the total extracted rules.
18. If someone works less than or equal to 39.5 hours per week, the work status is self-employed/employee/freelancer, and the proportion of working household members is less than or equal to 0.45, the probability of worker being categorized as poor is 95%. This rule exists in 1% of the total extracted rules.
19. If the worker's status is married and lived outside Jakarta 5 years ago, the probability of worker being categorized as poor is 95%. This rule exists in 1% of the total extracted rules.
20. If worker's status is married and the highest education level is primary/secondary/high school, the probability of worker being categorized as poor is 95%. This rule exists in 1% of the total extracted rules.
21. If someone works in the agricultural/transportation and communication/other services sectors and the highest education level is primary/secondary/high school, the probability of worker being categorized as poor is 95%. This rule exists in 1% of the total extracted rules.
22. If the proportion of working household members is less than or equal to 0.45, own a house, and the highest education level is primary/secondary/high school, the probability of worker being categorized as poor is 95%. This rule exists in 1% of the total extracted rules.

According to the explanation of the extracted rules above, the most frequent interaction occurs between worker that has the proportion of household members

who work, which is less than 0.45 and the highest education is primary/secondary/high school. Education stands out as a significant factor in predicting the status of working poor in Jakarta. Analysis indicates that individuals with lower levels of education tend to fall into the category of working poor. Education is important to reduce the poverty rate due to its correlation with increased household income overall [14]. Moreover, a higher level of education, particularly at the university level, contributes to higher incomes compared to lower education. Thus, enhancing educational opportunities becomes crucial for improving income.

In Jakarta, a household proportion less than or equal to 0.45 indicates that in a household of 5 members, only about 2 are employed. This can contribute to the tendency of workers within such households to be categorized as working poor. Previous studies show that the risk of worker poverty increases in larger households with many dependents [15]. The more household members are employed, the greater the household's ability to meet needs and enhance overall welfare.

Furthermore, the interaction between workers in self-employed/employee/freelancer positions and the proportion of household members who work, less than or equal to 0.45, also frequently occurs. This might be related to the worker's status in these jobs, providing unstable income or high uncertainty in the jobs [9]. When associated with a low number of working household members, this adds to the burden on worker and increases the risk of worker falling into the poor category.

When the employment status interacts with education level, the probability of becoming a poor worker is significantly high for those in self-employed/employee/freelancer positions with an education level below university. The combination of unstable jobs with low education results in workers struggling to earn higher incomes, thereby posing a relatively high risk of falling into poverty

Therefore, the interaction between low proportion of working household members and low education, the interaction between unstable job positions and low proportion of working household members, as well as the interaction between low education and unstable job positions, play a significant role in determining the characteristics of individuals categorized as working poor in Jakarta.

IV. CONCLUSION

This study applies the Double Random Forest (DRF) model to identify the characteristics of the working poor in Jakarta, aiming to gain insights into variables affecting the poverty status of workers. The DRF proves beneficial in predicting poor worker classification by extracting various rules that explain the variable interactions. The rules extracted from the DRF highlight the substantial risk faced by workers with education levels below university, increasing their vulnerability to poverty due to unstable employment and lower education. In conclusion, the interaction of low proportion of working household members, lower education levels, and unstable job positions significantly defines characteristics of the working poor in Jakarta, emphasizing the need for targeted policy interventions to reduce working poverty and increase socio-economic resilience among vulnerable populations.

V. REFERENCES

[1]. S. Han, H. Kim, and Y.-S. Lee, "Double random forest", Machine Learning, vol. 109, no. 8, pp. 1569–1586, Aug. 2020, doi: 10.1007/s10994-020-05889-1.

[2]. M. Haddouchi and A. Berrado, "Assessing interpretation capacity in Machine Learning: A critical review", in Proceedings of the 12th

international conference on intelligent systems: theories and applications, 2018, pp. 1–6.

[3]. A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)", IEEE Access, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[4]. R. Agrawal, T. Imieliński, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", in Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, in SIGMOD "93. New York, NY, USA: Association for Computing Machinery, 1993, pp. 207–216. doi: 10.1145/170035.170072.

[5]. H. Deng, "Interpreting tree ensembles with inTrees", International Journal of Data Science and Analytics, vol. 7, no. 4, pp. 277–287, Jun. 2019, doi: 10.1007/s41060-018-0144-8.

[6]. N. Majid, "Working Poor in Developing Countries", Int"l Lab. Rev., vol. 140, p. 271, 2001.

[7]. J. Gautie and S. Ponthieux, Employment and the Working Poor. 2016.

[8]. K. C.-K. Cheung and K.-L. Chou, "Working Poor in Hong Kong", Social Indicators Research, vol. 129, no. 1, pp. 317–335, Oct. 2016, doi: 10.1007/s11205-015-1104-5.

[9]. F. Faharuddin and D. Endrawati, "Determinants of working poverty in Indonesia", Journal of Economics and Development, vol. 24, no. 3, pp. 230–246, Jan. 2022, doi: 10.1108/JED-09-2021-0151.

[10]. F. Ramadhani and F. S. Putra, "Having a Job Is Not Enough to Escape Poverty: Case of Indonesian Working Poors", IPTEK Journal of Proceedings Series, no. 6, pp. 58–64, 2019.

[11]. Statistics Indonesia, Indikator Kesejahteraan Rakyat 2022. Jakarta: [BPS] Statistics Indonesia, 2022.

[12]. Statistics Indonesia, Keadaan Angkatan Kerja di Indonesia Agustus 2022. Jakarta: [BPS] Statistics Indonesia, 2022.

[13]. F. Gorunescu, Data Mining: Concepts, models and techniques, vol. 12. Springer Science & Business Media, 2011.

[14]. A. M. Arsani, B. Ario, and A. F. Ramadhan, "Impact of education on poverty and health: Evidence from Indonesia", Economics Development Analysis Journal, vol. 9, no. 1, pp. 87–96, 2020.

[15]. J. Feder and D. Yu, "Employed yet poor: low-wage employment and working poverty in South Africa", Development Southern Africa, vol. 37, pp. 363–381, 2020.

Cite this article as :

Adlina Khairunnisa, Khairil Anwar Notodiputro, Bagus Sartono, "A Study of Rule Extraction from Double Random Forest to Identify the Characteristics of Working Poor in Jakarta Province, Indonesia", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 10 Issue 6, pp. 258-266, November-December 2023. Available at doi : https://doi.org/10.32628/IJSRSET231069
Journal URL : https://ijsrset.com/IJSRSET231069