Classification and regression trees with gini index

Article *in* Advances in Mathematics Scientific Journal · October 2020

DOI: 10.37418/amsj.9.10.53

Dr T. Daniya
GMR Institute of Technology
46 PUBLICATIONS **516** CITATIONS
SEE PROFILE

Suresh Kumar K Dr
SAVEETHA ENGINEERING COLLEGE
76 PUBLICATIONS **383** CITATIONS
SEE PROFILE

ADV MATH
SCI JOURNAL

Advances in Mathematics: Scientific Journal **9** (2020), no.10, 8237–8247
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.10.53 Spec. Iss. on AOAOCEP-2020

CLASSIFICATION AND REGRESSION TREES WITH GINI INDEX

T. DANIYA[1], M. GEETHA, AND K. SURESH KUMAR

ABSTRACT. Classification and Regression Trees (CART) is applied for classifying and predicting regression problems. The CART model is represented using a binary tree, which uses split rule at each root node. Starting from the root node the split condition is applied and the decision process is continued for each sub root node. ‘x’ represents a single input variable at each root node including a split on that variable. ‘y’ represents an output variable, placed at the leaves which predicts the output. The output variable can be defined with a continuous target or a categorical target. The continuous target uses a sum of square errors and the categorical target uses the choice of entropy. Gini measure is a splitting rule. In this paper, CART uses the Gini Index for classifying the decision points. The choice of applying splitting rule improves the performance of the CART classifier algorithm.

1. INTRODUCTION

Classification and regression trees are used for classification or regression predictive working problem. Each node of the classification and regression tree model has two children only, which is a binary tree condition. The input variable ‘x’ is stored at the root node and output variable ‘y’ is at the leaf node. Splitting rule is applied to input variable and output is predicted with the output variable ‘y’. A process of dividing the input space is binary decision tree.

[1] *corresponding author*

2020 *Mathematics Subject Classification.* 68R10, 68T35.

Key words and phrases. Gini Index, Regression, Classification, CART.

The Various measures [9] like entropy, Gini Index and Information Gain are developed for attribute selection in order to place a particular attribute in an appropriate position of the decision tree. For doing this, an average is calculated between Gini Index, Information Gain and Diversity Index and based on this attributes are assigned with a weight, finally for classification, the attribute that has highest average value is selected. For example, given a dataset with two inputs x_1 -> height in centimeter and x_2 -> weight in kilogram. Output is to predict whether the person is male or female. The decision tree can be represented as in Fig. 1. In Classification and regression trees we use the greedy approach. Using this approach we divide the input space called recursive binary splitting.

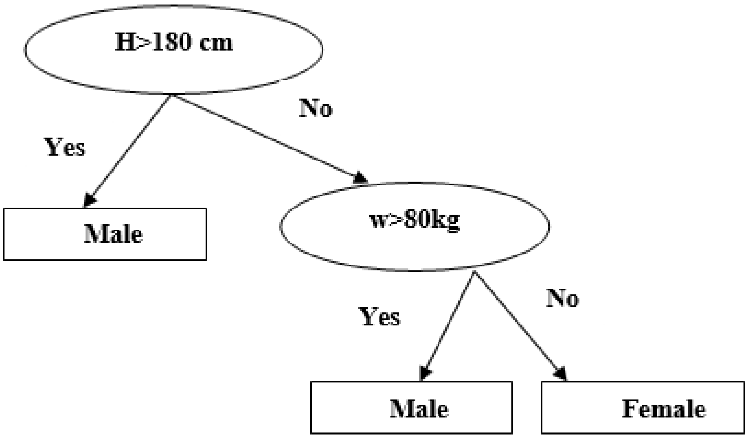


FIGURE 1. Decision Tree Representation

Splitting Rules. The greatest separation in the target variable 'y' is achieved by selecting and input variable (x=t1).

Regression Trees. "Use sum of square of errors"

Classification Trees. Use the choice of "Entropy", "Gini measure", "twoing" Splitting rules. In this paper, section 2 comprises of the related work where several related articles are referred and the work related to the references are expressed in detail. In section 3, the detail methodology for calculating the Gini Index measure for each attributes is mentioned along with algorithmic perception. In section 4, the results are discussed and the paper is concluded in section 5.

2. Related Work

The administered learning procedures, the preparation information is named. It implies every perception in the informational collection has both spellbinding factors and a named result variable. Marks can be either classes or persistent qualities [1]. In contrast to managed learning, with solo learning the information isn't marked. For accurate future predictions this learning model maps the input and output variables. In Unsupervised learning the training set consists of descriptive variables and not the outcome variables Perhaps the model itself has to predict the interesting pattern and structure present in the data [2]. The aim of classification algorithm is to develop a model by analyzing the training data and predicting the future characteristics. Decision tree is an efficient classification algorithm which classifies the data based on the input values to number of classes, which are categorical in future [4].The classification in decision tree starts at the root node and progressed till the leaf node which predicts the class labels of an output variables [5]. The split condition at each node decides a traversal towards the leaf nodes and results in homogeneous subsets.

Each homogeneous subset should hold same class label using relative data to achieve homogeneous subset is not possible which results in split criterion to ensure lowest impurity in chosen nodes [6]. In research field there are different applications like image forgery detection [8], face detection [7], Plant disease detection [3] using Machine Learning Algorithms. One of the main advantage [10] of Gini Index is that, it can distinguish any two distributions that has same entropy measure. This is because, any distribution cannot be summarized with a single measure. Most of the government agencies uses Gini Index for summarizing the income inequality.

3. Methodology

How classification and regression trees selects the optimal trees. Use Cross Validation to select the optimal decision tree. Basic idea is "grow the tree" out of far as you can and then "prune back". Then this cross validation tells you when you stop "pruning". The Classification and regression trees use a new metric named as Gini Index to create decision points for classification tasks. The Gini Index in-equality can be expressed as Eq. 1,

$$\text{Inequality} = \sum_j X_j f(r_j). \tag{3.1}$$

For uniform population,the Gini Index can be expressed as Eq. 2,

$$G = \frac{2\sum_{i=1}^{n} iY_i}{n\sum_{i=1}^{n} Y_i} - \frac{n+1}{n}. \tag{3.2}$$

The Gini Index for calculating discrete probability distribution depicted in Eq. 3.

$$G = 1 - \frac{\sum_{i=1}^{n} fY_i(S_{i-1} - S_i)}{S_n}. \tag{3.3}$$

The Gini Index for calculating continuous probability distribution expressed as Eq. (4).

$$G = \frac{1}{2\mu}\iint |Q(F_1) - Q(F_2)|dF_1 dF_2. \tag{3.4}$$

Gini Impurity can be expressed as Eq. 5.

$$\text{Gini Impurity} = 1 - \sum_{i=1}^{c} P_i^2. \tag{3.5}$$

Steps to Pick a Decision Node.

Step 1: Calculate the Gini Index for each attribute.

Step 2: Weighted sum of Gini indexes is calculated for the feature.

Step 3: Pick the attribute with lowest Gini index value.

Step 4: Repeat 1,2,3 until a generalized tree has been created.

For the dataset in Table 1, the decision tree can be represented in Fig. 2.

Consider the outlook in the Table 1. Outlook is a nominal feature. It can be Daylight, Cloudy or Rainfall. Now from Table 1, we summarize the final decision for outlook feature in Table 2. Now put these values in the formula for Gini Index. So we get,

- Gini(Outlook=Daylight)=1-(2/3)2-(3/5)2=0.48
- Gini(Outlook=Cloudy)=1-(4/4)2-(0/4)2=0
- Gini(Outlook=Rainfall)=1-(3/5)2-(2/5)2=0.48

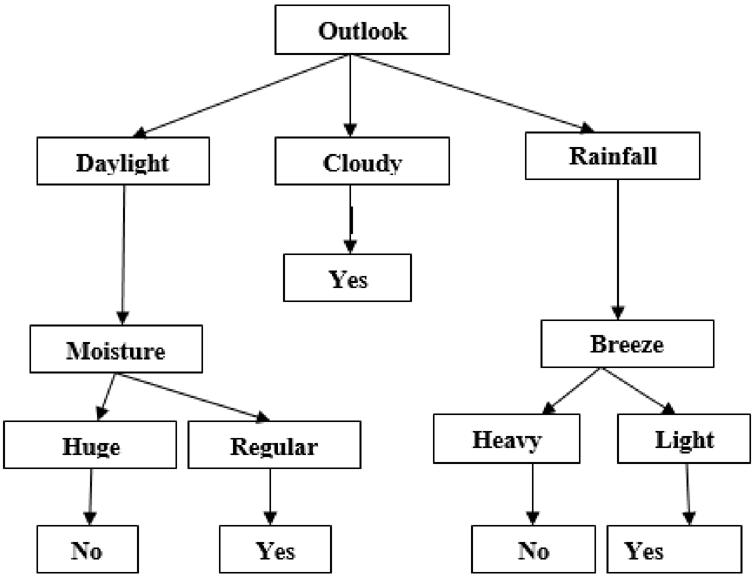


FIGURE 2. Decision tree for Table 1

TABLE 1. Dataset

Day	Outlook	Temperature	Moisture	Breeze	Play Game
D1	Daylight	Hot	Huge	Light	No
D2	Daylight	Hot	Huge	Heavy	No
D3	Cloudy	Hot	Huge	Light	Yes
D4	Rainfall	Warm	Huge	Light	Yes
D5	Rainfall	Cold	Regular	Light	Yes
D6	Rainfall	Cold	Regular	Heavy	No
D7	Cloudy	Cold	Regular	Heavy	Yes
D8	Daylight	Warm	Huge	Light	No
D9	Daylight	Cold	Regular	Light	Yes
D10	Rainfall	Warm	Regular	Light	Yes
D11	Daylight	Warm	Regular	Heavy	Yes
D12	Cloudy	Warm	Huge	Heavy	Yes
D13	Cloudy	Hot	Regular	Light	Yes
D14	Rainfall	Warm	Huge	Heavy	No

Then we calculate the weighted sum of gini indexes for outlook feature.

Gini(Outlook)=(5/14)*0.48+(4/14)*0+(5/14)*0.48=0.342

TABLE 2. Final decisions for outlook feature

Outlook	Yes	No	Number of instances
Daylight	2	3	5
Cloudy	4	0	4
Rainfall	3	2	5

Repeat the same steps for other attributes temperature, Moisture and Breeze. The Weighted sum of Gini Indexes mentioned in Table 3. From above feature values, we have to pick the lowest one as the root. So from above table we get outlook at the root which has less Gini Index Value. Now the Decision tree look like in Fig. 3.

TABLE 3. Weighted sum of Gini indexes

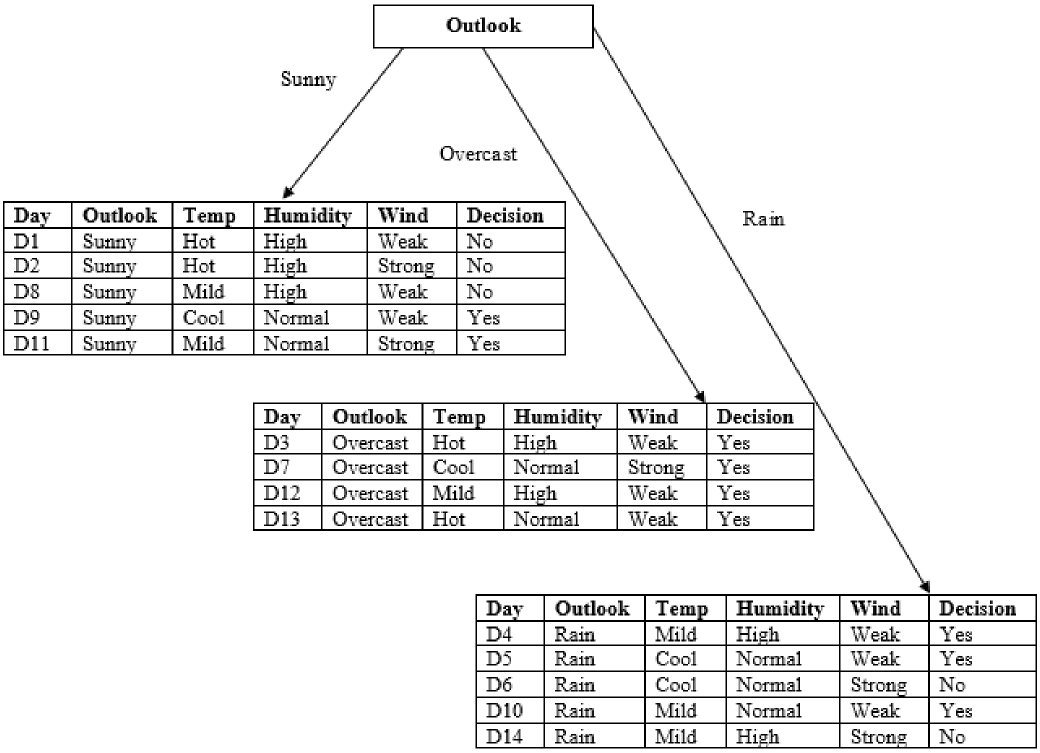
Feature	Gini index
Outlook	0.342
Temperature	0.349
Moisture	0.367
Breeze	0.428

If you see the Cloudy, all the decision is "yes" so we can directly put "yes". This is the first decision we made. Now look on Daylight, here we have multiple decision 'yes' and 'No'. Already we used outlook, So consider the other three features and calculate the Gini Index from above tables and weighted sum of Gini Indexes mentioned in Table 4. The Gini Indexes from Table 4 is calculated from the Table 5, Table 6 and Table 7 target values.

TABLE 4. Weighted Sum of Gini indexes

Feature	Gini index
Temperature	0.2
Moisture	0
Breeze	0.466

- Gini(Temp=Hot)=1-(0/2)2-(2/2)2=0
- Gini(Temp=Warm)=1-(1/2)2-(1/2)2=0.5
- Gini(Temp=Cold)=1-(1/1)2-(0/2)2=0



Day	Outlook	Temp	Humidity	Wind	Decision
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes

Day	Outlook	Temp	Humidity	Wind	Decision
D3	Overcast	Hot	High	Weak	Yes
D7	Overcast	Cool	Normal	Strong	Yes
D12	Overcast	Mild	High	Weak	Yes
D13	Overcast	Hot	Normal	Weak	Yes

Day	Outlook	Temp	Humidity	Wind	Decision
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D10	Rain	Mild	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

FIGURE 3. Decision Tree based on split conditions

TABLE 5. Predicted target value for Daylight

Temperature	Yes	No	Number of instances
Hot	0	2	2
Warm	1	1	2
Cold	1	0	1

Then we calculate the weighted sum of Gini indexes for Temperature feature.

Gini(Temperature)=(2/5)*0+(2/5)*0.5+(1/5)*0=0.2.

Similarly below table we calculate, Gini(Moisture)=0 and Gini(Breeze)=0.466. Now Moisture has value ‘0’ that is less value. So, it is splitted as shown in Fig.

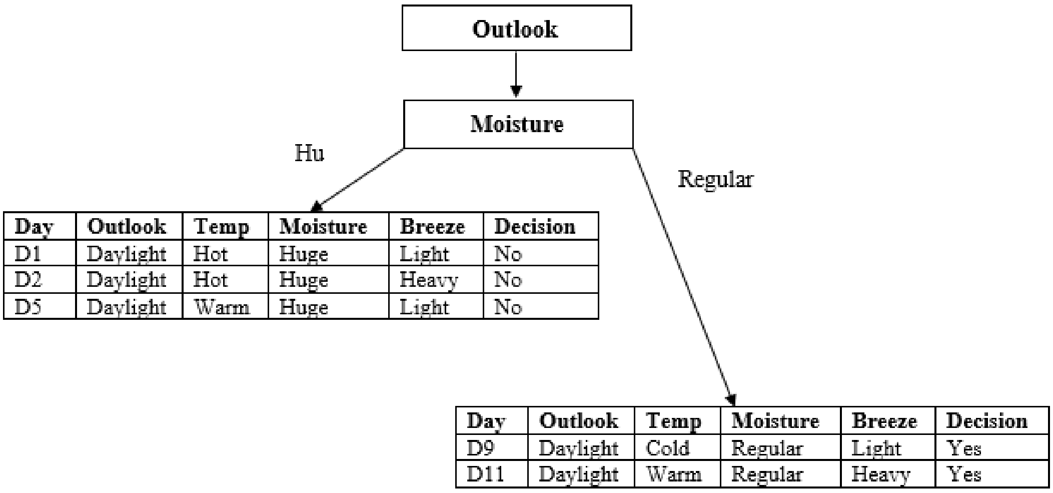
TABLE 6. Predicted target value for Moisture

Moisture	Yes	No	Number of instances
Huge	0	3	3
Regular	2	0	0

TABLE 7. Predicted target value for Breeze

Breeze	Yes	No	Number of instances
Light	1	2	3
Heavy	1	1	2

4. In above decision tree, all the output decisions for Huge is "No" and Regular is "Yes". So we do not want to calculate further, we directly place Moisture as 'No' and 'Huge'. So we get the decision tree as Fig. 5.



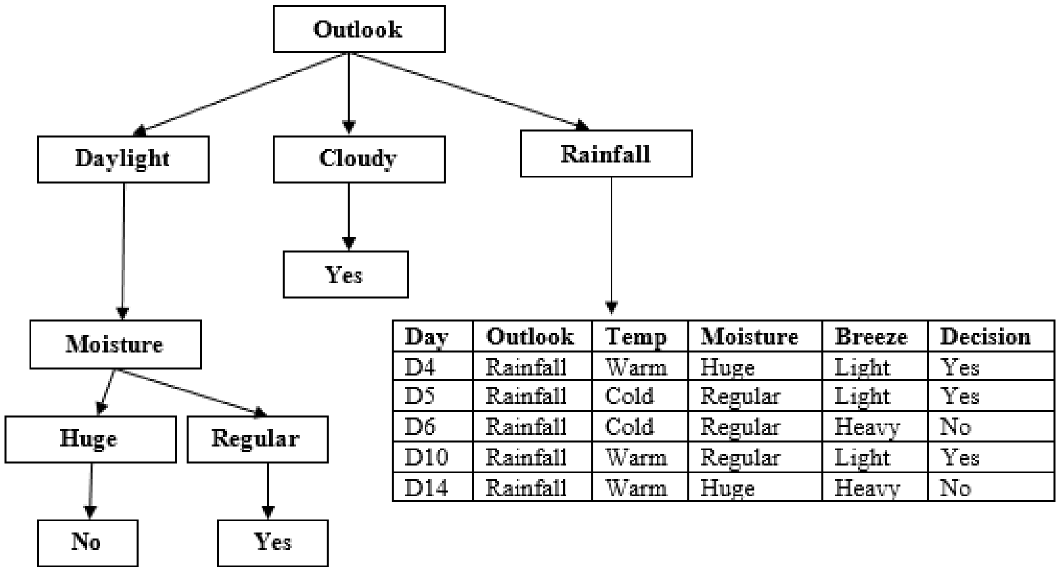
Day	Outlook	Temp	Moisture	Breeze	Decision
D1	Daylight	Hot	Huge	Light	No
D2	Daylight	Hot	Huge	Heavy	No
D5	Daylight	Warm	Huge	Light	No

Day	Outlook	Temp	Moisture	Breeze	Decision
D9	Daylight	Cold	Regular	Light	Yes
D11	Daylight	Warm	Regular	Heavy	Yes

FIGURE 4. Decision Tree on splitting 'Moisture'

4. RESULTS

Since the output decision of "Rainfall" has multiple classes that is 'Yes' and 'No', again we have to calculate the values. By following the previous steps, we get the final decision tree as Fig. 6. In Table 8, the weighted sum of Gini index for each attribute is listed which has been evaluated with the help of splitting rule on decision nodes. The graphical representation of the Gini index is depicted in Fig. 8.



Day	Outlook	Temp	Moisture	Breeze	Decision
D4	Rainfall	Warm	Huge	Light	Yes
D5	Rainfall	Cold	Regular	Light	Yes
D6	Rainfall	Cold	Regular	Heavy	No
D10	Rainfall	Warm	Regular	Light	Yes
D14	Rainfall	Warm	Huge	Heavy	No

FIGURE 5. Decision Tree on splitting 'Rainfall'

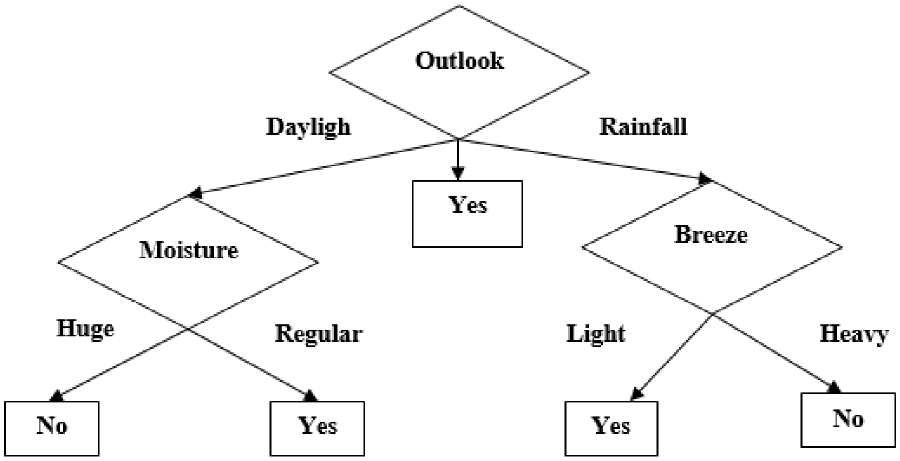


FIGURE 6. Final decision tree

5. CONCLUSION

Gini Index is popularly used in economics and distribution theory. Some of the other applications are biodiversity measurement, healthcare, education, chemistry, credit risk management so on. In this paper, several related articles are referred with respect to decision tree and for measuring the attribute the Gini

TABLE 8. Final Gini Index for the given attributes

Attributes	Gini Index
Outlook	0.342
Temperature	0.2
Moisture	0
Breeze	0.466

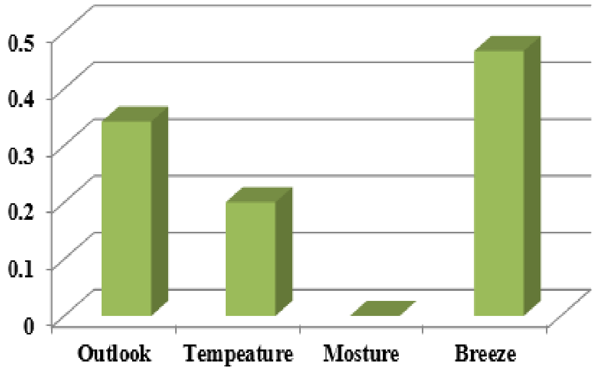


FIGURE 7. Graphical representation of final Gini index

Index measure is calculated for each root node. Also, Gini Coefficient measures for in-equality condition, Uniform distribution, Discrete probability distribution. Continuous probability distribution along with Gini Impurity has been employed. The CART used Gini index for classifying the decision points for the given dataset and applied splitting rule that improves the performance of CART classifier algorithm which results in an optimal tree.

REFERENCES

[1] G. JAMES, D. WITTEN, T. HASTIE, R. TIBSHIRANI: *An Introduction to Statistical Learning*, 1st ed., Springer, New York, 2013.

[2] C. DOHERTY, S. CAMINA, K. WHITE, G. ORENSTEIN: *The Path to Predictive Analytics and Machine Learning*, 1st ed., O'Reilly Media, Inc., California, 2016.

[3] R. CRISTIN, B.S. KUMAR, C. PRIYA: *Deep Neural Network based Rider-Cuckoo Search Algorithm for Plant Disease Detection*, Artif Intell Rev., (2020), 1–12, DOI: 10.1007/s10462-020-09813-w.

[4] E. TURBAN, R. SHARDA, D. DELEN: *Business Intelligence and Analytics: Systems for Decision Support*, 10th ed., Pearson, London, 2015.

[5] W.Y. LOH, Y.S. SHIH: *Split Selection Methods for Classification Trees*, Statistica Sinica. **7** (1997), 815–840.

[6] L. BREIMAN, J.H. FRIEDMAN, R.A. OLSHEN, C.J. STONE: *Classification and Regression Trees*, 1st ed., Chapman & Hall/CRC, London, 1984.

[7] R. CRISTIN, J.P. ANANTH, V. CYRIL RAJ: *Illumination Based Texture Descriptor and Fruitfly Support Vector Neural Network for Image Forgery Detection*, IET Image Proc., **12**(8) (2018), 1439–1449.

[8] R. CRISTIN, V. CYRIL RAJ: *Consistency Features and Fuzzy Based Segmentation for Shadow and Reflection Detection in Digital Image Forgery*, Sci China Infor Sci., **65**(1) (2017), 43–66.

[9] S. MUHAMMAD, Z. TANVEER, S. ALI KHAN: *Decision Tree Classification: Ranking Journals using IGIDI*, J Inform Sci., **46**(2) (2020), 325–339.

[10] Y. LIU, J.L. GASTWIRTH: *On the Capacity of the Gini Index to Represent Income Distributions*, METRON., **78** (2020), 61–69.

DEPARTMENT OF INFORMATION TECHNOLOGY
GMR INSTITUTE OF TECHNOLOGY
RAJAM, ANDHRA PRADESH, INDIA
Email address: daniya.t@gmrit.edu.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING AND TECHNOLOGY
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
VADAPALANI CAMPUS, CHENNAI, TAMIL NADU, INDIA
Email address: geethapandian@yahoo.com

DEPARTMENT OF INFORMATION TECHNOLOGY
SAVEETHA ENGINEERING COLLEGE
CHENNAI, TAMIL NADU, INDIA
Email address: ksureshmtech@gmail.com