A Novel Feature Probability Based Decision Tree (FPBDT) Algorithm for Data Classification and Analysis

1st O.Yamini, 2nd Dr G.V.Ramesh Babu

[1]Research Scholar, [2] Research Supervisor

Department of Computer Science, S V University, Tirupati A.P., India

Abstract - *Generation of huge data by applications has resulted in explosive growth of data and databases which requires efficient Techniques and intelligent tools to process data for useful insights for helping managers and decision makers etc. One of the important technique of data mining is Classification which is used for finding a model (or function) that describes and distinguishes data classes or concepts or mapping of data item into one of known label which are predefined as per rules or conditions. Decision trees are used in various classification learning systems like ID3,C4.5,CART etc., One of the simple decision tree learning systems are ID-3 (Iterative Dichotomiser3) is the simple and effective way of classification but have disadvantages like creates a complex tree may not be able to backtrack, not suitable for large datasets etc. An extensions to ID-3, C4.5, also done such as the domain of classification from categorical attributes to numerical ones. So by considering above factors we have proposed a modified novel ID3 technique which not only reduces logarithmic computations in information, entropy calculations but with simple probabilistic calculations and table based feature selection is done node for constructing decision tree. Bankers has to make various permutations and combinations and to decide the probable rule set for sanctioning vehicle loans based critical feature value present in customer data set which are to be evaluated before the loan to be sanctioned. So in this research paper above facts are considered and designed feature probability based ID3 classifier prediction algorithm which will be helpful for stake holders under consideration of problem on hand*

IndexTerms - Data mining, KDD, *Dataset, features, Classification, Decision Tree, ID3, CART*

I. INTRODUCTION

Recent developments in terms of generating and consuming data had make Data processing and analysing a very important task in various sectors like corporate dealing with health care, financial, marketing, Artificial intelligence, scientific explorations etc. Generation of huge data by applications has resulted in explosive growth of data and databases which requires efficient Techniques and intelligent tools to process data for useful insights for helping managers and decision makers etc. Hence data mining task has become prominent in every area of data analysis. Data mining or KDD (Knowledge Discovery in Databases) is mechanism of extracting an implicit unknown and useful information such as associative rules, constraints, and patterns from databases.[1][2][3] KDD also known as knowledge mining from databases, knowledge extraction, data archaeology, data dredging, data analysis, etc. Investigation of data from data ware houses are done in varied angles for getting interesting facts, regularities, or high-level information which can serve as rich and reliable sources for knowledge generation and verification.

One of the important techniques of data mining is Classification which is used for finding a model (or function) that describes and distinguishes data classes or concepts or mapping of data item into one of known label which are predefined as per rules or conditions. In most of the classification techniques uses a classifier also known as an algorithm that learns from the training set and then assigns new data point to a particular class. A decision tree is a diagrammatic tool for classification and prediction. It is a flow-chart-like tree structure more than one leaf nodes from root node, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and leaves at the end of tree represents classes or class distributions. Decision trees are used in various classification learning systems like ID3,C4.5,CART etc., One of the simple decision tree learning systems are ID-3 (Iterative Dichotomiser 3), implements a top-down immutable strategy that moves on down and searches only part of the search space. ID3 algorithm's computation of Entropy factor and information gain is modified with many weight factors for better efficiency. It is the simple and effective way of classification but have disadvantages like creates a complex tree may not be able to backtrack, not suitable for large datasets etc. An extensions to ID-3, C4.5, also done such as the domain of classification from categorical attributes to numerical ones.So by considering above factors we have proposed a modified novel ID3 technique which not only reduces logarithmic computations in information, entropy calculations but with simple probabilistic calculations and table based feature selection is done node for constructing decision tree. Bankers has to make various permutations and combinations and to decide the probable rule set for sanctioning vehicle loans based critical feature value present in customer data set which are to be evaluated before the loan to be sanctioned. The proposed algorithm was designed with feature probability based ID3 classifier prediction algorithm for classification.

II. LITERATURE REVIEW

Analyzing data is the main theme of Data mining which may adopt supervised or unsupervised way for better results basing on the problem on hand. In case of supervised data analysis method entails defining information in advance and attempting to explore the problem in consideration. The principle of classification is to identify grouping attributes and establish grouping rules or patterns on the basis of the categories of known or available target data. The data are divided into corresponding target categories which is essential for reviewing data characteristics. The classification of data is computed or classified into various categories as per the rules which are predefined which gives us a predictable classification model.Decision rules or classification processes are defined on basis of determining the characteristics of specific groups in data set. (Chou et al. 2016) [19].

Most popular algorithm used in classification using the decision tree generation is ID3, which is widely applied to many fields now. But it has some of the short comings like time taken to generate decision tree, error in classification and it is applied to static data set containing classification rules only. In [20] authors has optimized the decision tree generation by shortened the time of generating decision tree, and at the same time, compensated the error brought by optimization process.Authors have modified ID3 scheme effectively to overcome the shortcoming of ID3 which usually chooses attributes which has more value instead selecting attribute with minimum values such as low entropy as given by authors.

In paper [21] an improved ID3 algorithm is proposed where information gain is combined with important of attribute with Association function (AF) for selecting an attribute for decision tree construction. Proposed function (AF) overcomes the deficiency of ID3 in considering data with more attributes but also give relation between all elements and its attributes. Improved ID3 will create decision tree by considering importance of attribute in data set and new information gain function is given. In proposed paper Gain'(A) is used as a new measure for attribute selection mechanism for constructing decision tree according to the procedures of ID3 algorithm.

A proposed an improved ID3 algorithm by Yi-bin et al [22], which computes information entropy based on different weights combined with coordination degree in rough set theory. In ID3, logarithmic computations in information gain are complex for selecting the optimal feature for decision tree. So authors have used new formula replacing the logarithmic equation of information gain by the four basic operations like addition, subtraction, multiplication, and division) which improves the running speed of the decision tree building process.

In ID3 algorithm, computational complexity is an issue apart from the known problem of bias towards multi-value attributes. In an attempt to solving these problems, Wang et al, 2017 proposed a novel approach to select the splitting attribute in a decision tree. The authors in place of information gain used rough set to implement the algorithm by introducing a concept of consistency, which forms as the criteria for splitting the data. By doing so the authors has improved algorithm and solved the issue of feature bias towards multi-valued attributes and reduced the computational complexity of the standard ID3 algorithm [23].

In Paper [24] decision tree creation is done by using algorithm with mutual information rather than information gain. Mutual information is used for selecting the splitting attribute for decision tree. As the decision tree creation by ID3 is complex and slow authors used mutual information concept. The results show that the new technique gave better accuracy and the performance than the traditional ID3 algorithm.

The CART algorithm is induction based which recursively partitions the measurement space, displays the resulting partitions as decision trees. Even though CART uses cross-validation (cv) for selecting tree of reasonable sizes but there is a possibility of over fitting the trees to the data. In [25] authors presented the incremental CART algorithm as the currently available implementations of CART methodology are not capable of incremental learning as new dataset values can be incorporated into the decision tree only via the brute force method to the existing tree T and the induction procedure is executed from scratch on the expanded tree T. The newly designed algorithm offers considerable potential as a means by which accurate and parsimonious trees can be maintained, with minimal computational burden, in situations in which new training data is not stable.

In (Chen et al, 2013), the researchers in their paper has proposed an improved C4.5 decision tree algorithm based on best sample selection of data in order to improve the classification accuracy, to minimize the training time of large sample of data on hand, and finding the best training set. In this paper by authors has have modified C 4.5 algorithm by building the decision tree with nodes giving only get local optimal solution and which shows the bigger relativity with initial sample in considered. When evaluating the results authors used simple iteration process to find the best training set to get a better sample. The algorithm proposed has experimented on large datasets had given better accuracy and time consumption has reduced when compared with standard C4.5 algorithm [26].

III. THE KDD PROCESS

The processing of data with a well-defined model is known as Knowledge Discovery Databases (KDD) is an iterative and interactive model which has phases given in figure-1. Each phase uses data mining method extracts patterns in finding knowledge in data.Hence Data mining is the application of specific algorithms for extracting patterns from data at every stage. Practical view of KDD process is given by [4], which is iterative in nature. Various stages in KDD is outlined below

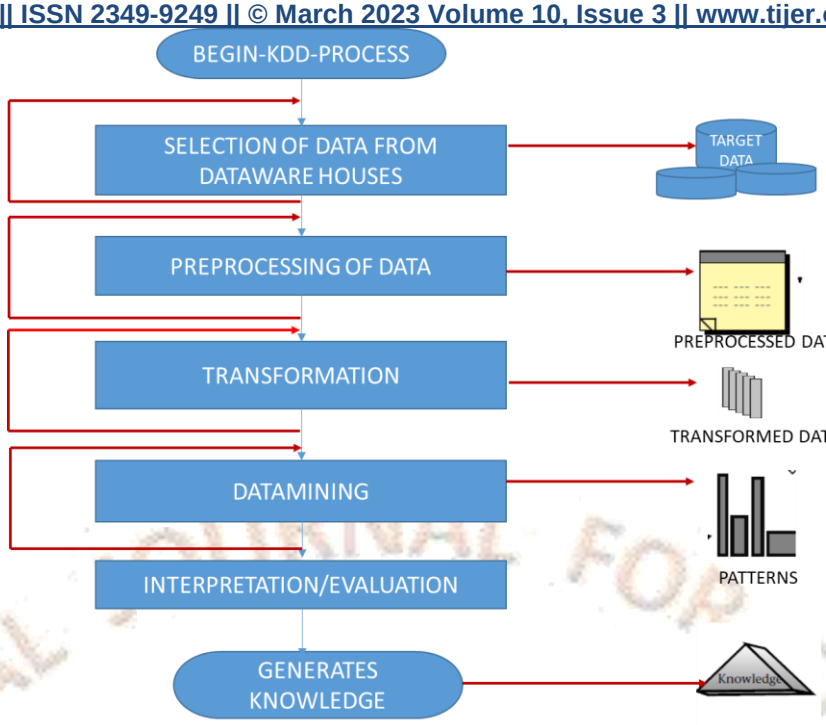

Figure:-1 KDD Process Scenario

In *Stage-1* according to customers view point, primary goal of KDD is formulated by understanding of the application domain and the relevant prior knowledge.

In *Stage-2* focus is on to creating target data, by extracting a target data set: by selecting appropriate subset of variables or data samples, on which final result is to be deduced.

At *Stage-3* data cleaning and pre-processing is performed. It means data may contain missing values, null values or any noise data, which should be removed with various strategies before data to be used in next stage.

Stage-4 uses dimensionality reduction or transformation methods, for finding useful features to represent the data depending on the goal of the task. Effective features can be identified or invariant representations for the data can be found.

In *Stage-5,* includes the process of incorporating the goals of the KDD process (step 1) a particular data-mining method such as summarization, classification, regression, clustering, and so on are chosen. [5]

Exploratory analysis at Stage-6 is a hypothesis selection model, for extracting patterns by searching in data with an efficient data mining algorithms are to be selected. Here decision is done for selecting models and parameters which might be appropriate data mining technique for reaching our goal in KDD process as data may be categorical or numerical.

Stage-7 in KDD is application of data mining technique such as classification rules or trees, regression, and clustering for searching for patterns of interest in a particular representational form. Success of this stage depends upon cautious selection of methodologies in above stages.

Stage-8 involves in analysing mined patterns, which may lead to repetition of any of steps 1 through 7 for further iteration and accuracy. Visualization of the extracted patterns and models with data is used for analysis or comparisons.

At final stage-9 after discovering useful knowledge, it can be used or taken as input into another system for further action, or simply documenting it and reporting it to interested parties. This process may also include verification and validation for and resolving impending conflicts with already derived (or extracted) knowledge. As described above data mining is used in various models like SEMMA, CRISP-DM and with KDD one describe above.

SEMMA process was developed by the SAS Institute. The acronym SEMMA stands for Sample, Explore, Modify, Model, Assess, and refers to the process of conducting a data mining project. It has five different phases for discovering knowledge. Cross-Industry Standard Process for Data Mining (CRISP-DM) [4] was launched in late 1996 by Daimler Chrysler (then Daimler-Benz), SPSS (then ISL) and NCR. This models the refines over the years. It contains six steps or phases.[6]

IV. CLASSIFICATION

Classification is one of the data mining technique for finding a model (or function) that describes and distinguishes data classes or concepts or mapping of data item into one of known label which are predefined as per rules or conditions. The model learns to classify data item by analysis set of training data (i.e., data objects whose class label is known). This technique takes new input instances and maps it to given label or group. The input instances can be data set which contains unstructured or structured data.[9]

In most of the classification algorithms uses a classifier also known as an algorithm that learns from the training set and then assigns new data point to a particular class. Classification uses mapping function that maps new data entry to class label with help of training dataset which used by mapping function for prediction. Classifications not only maps to single class label but also classifies to more than one. In case of Binary classification there will be two possible outcomes. For example, weather forecast (it will rain or not), spam or fraud detection (predict whether an email is spam or not).In case of Multi-label classification from data set, results in more than two possible Outcomes. For example, classify academic performance of students as excellent or good or average or poor. Classification techniques are also applied in financial markets as part of knowledge discovery for classifying trends of various shares and the automated identification of objects of interest in large image databases.[7][8]

V. DECISION TREE

A decision tree is a diagrammatic tool for classification and prediction. It is a flow-chart-like tree structure more than one leaf nodes from root node, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and leaves at the end of tree represents classes or class distributions. Decision trees can easily be understood or can be converted to classification rules [9]. A Decision tree is built in two phases such as: the growth phase and the pruning phase. In the growth phase a recursive partitioning of the training data set which results a Decision tree such that either each leaf node is associated with a single class label or further partitioning if results from given leaf would result in at least its child nodes which further partitioned for some specified threshold value [10].The pruning phase aims in avoid over-fitting by generalize the Decision tree that was generated in the growth phase by generating a sub-tree from training data. The pruning phase initially known as Pre-pruning occurs during growth phase of which avoids splits and also avoids data which does not meet criteria such as minimum number of observations for a split search, minimum number of observations for a leaf and post-Pruning that occurs after tree construction when it is having more depth and possibility of over fitting.[11]

A.ID3 Algorithm:

Decision trees are used in various classification learning systems like ID3,C4.5,CART etc., One of the simple decision tree learning systems are ID-3 (Iterative Dichotomiser 3) [12], implements a top-down immutable strategy that moves on down and searches only part of the search space. In this process of reaching classification in ID3 entropy and information gain are calculated. ID3 algorithm starts with the dataset as the root node and for every attribute entropy and information gain is calculated. The default entropy used is Shannon entropy in ID3 algorithm. The attribute with the smallest entropy of the largest information gain is chosen for further split.[13]

The ID-3 system [12] uses information gain as the evaluation functions for classification, with the following evaluation function,

$$I = \sum(pi(\log(pi)))$$

Where p in equation given above, is the probability that an object is in class i. There are many other evaluation functions, such as *Gini index*, *chi-square test*, and so forth.

B. C.4.5 Algorithm

C4.5 is a decision tree based classification algorithm was proposed by J.R. Quinlan in 1993 mainly designed to overcome some of the drawbacks in ID3. Information Gain rate is calculated for test attribute selection [14]. In this algorithm pruning phase in decision tree construction eliminates doubtful branches by swapping them with leaf nodes by backtracking tree.C4.5 will deal with missing values in training set. In C4.5 algorithm, information gain rate is used as the basis of test attribute selection [14]. During construction of decision tree, pruning phase of C4.5 tries to eliminate the un-comfort branches by swapping them with leaf nodes by going back through the tree once it has been generated. The main advantage of C4.5 are it deals with training set with data having missing feature values, deals both discrete and continuous features that support for both pre and post pruning. Algorithm C4.5, in place of information gain a normalized method of "split information is used to overcome the bias in information gain, so we use Split information not information gain as

$$\text{SplitInfo}_A(D) = -\sum_{j=1}^{v} \frac{|Dj|}{D} \; x \; log2\left(\frac{|Dj|}{D}\right)$$

The above computed value will give possible information gained by dividing data set D, into *v* partitions, corresponding to the v outcomes of a test on an attribute A. For each v_i of outcome, it considers 'i' number of tuples having that outcome with respect to the total number of tuples in D. An attribute with high gain ratio is chosen as a splitting attribute which is given as

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

Even though C4.5 is used commonly , but has disadvantages like more logarithmic operations in computing Gain-Ratio which leads to more computation time , may lead to over fitting problem due to selection of attributes not have importance in classification and leads more time consuming if tree having more nodes.[14][15]

C. CART Algorithm

The CART algorithm used in one of the data mining task was proposed by Breiman et al. (1984) is known as Classification and Regression Tree algorithm which creates binary tree with exactly two outcomes from internal nodes. Node splitting is selected by using Gini index. In this algorithm uses binary approach by dividing data set into tow subsets and recursively splits subsets in binary fashion until no longer split is possible. After applying trained data set and tree is pruned, a smallest tree is selected for efficient classification which is the algorithm designed for. CART can be applied with target variable having with both categorical and continuous data, as the tree is known as regression tree if the target variable represents continuous data and otherwise known as classification tree if the target variable contains categorical data, a classification tree can be used. In the CART algorithm, at every root node split rule is applied based on dynamic threshold value entropy is used and whether node to be split or not is computed by using Gini Index. Lower value of Gini index indicates target variable has single category and vice-versa. [16][17][11]

Assume that a data set D contains n samples and that P_j is the relative probability that the sample of category 'j' appears in D. The Gini index is defined as follows: Gini index is defined as

$$Gini\ Index = 1 - \sum_{j=1}^{n} Pj^2$$

Decision tree classification involves selecting node and building tree, in case of ID3algorithm's some modification done with Entropy factor with some of the weight factors. Even though ID3 is simple and effective in classifying data set but it has some of the drawbacks like not possible able to backtrack tree, not suitable for large datasets etc. An extension to ID-3, C4.5, proposed to alter the domain of classification from categorical attributes to numerical ones. In case of C4.5 maximum information gain rate is chosen as the split attribute when selecting the split attribute for tree, but with many of logarithmic operation are involved which required many calls to library functions when comes to implementation which reduces efficiency, also computation of information gain rate increases frequently when continuous attributed are discretized, also generating more trees for non-considering attributed leads to over-fitting and algorithm possibility of generation more number of decision trees leads to inefficiency. While using CART algorithm less combinations are considered in each split of node in building tree, also tree is unstable while adding more independent attributes to data set. Due to data generation is very high by applications now-a-days most of the techniques developed in machine learning and statistics may encounter the problem of scaling-up. They may perform reasonably well with high efficiency in relatively small databases but they lack it in case of large databases with either poor performance or the reduction of classification accuracy when the training data set grows very large due to its usage.[18][26].

VI. PROPOSED FPBDTALGORITHM

Decision tree algorithm like ID3, C4.5, CART etc., has modified various ways due to disadvantages studied in literature survey like complexity in calculations, memory usage, unable to manage large datasets, decision tree may not be stable in some situations are some of the points considered and designed the proposed Feature Probability Based Decision Tree (FPBDT) Algorithm which uses probability regarding features of data set to decide the in node selection in decision tree. For each table Feature-Frequency-table is constructed to build Feature-Frequency-Table (FFT) with classification (yes/no) and node is selected based on the outcome in the algorithm given below.

Algorithm: Feature_Probabilty_Based_Decision_Tree_Algorithm (Dataset, DT)

1. Read Dataset D, with Rows R1...Rn with columns {C1, C2…Cn} with features {f_1, f_2….f_n} for each column determines class to which class a rows in table belongs.
2. Repeat for each column C_i
   - 2.1 For each feature F_i =1 to n from dataset D
     - a. Construct Feature-Frequency-Table (FFT) with classification (yes/no).
3. Repeat for each Row (R_i) in FFT create Probability-FFT table.
   - 3.1 Calculate Probability PF_i for each feature
   - 3.2 update Column C_{i+1} with PF_i
4. For each column C_i in Dataset with Feature F_i construct Feature-Selection-Tables (FST_i) as
   - 4.1 $FST_i[Fi,Yes]= P(Yes/F_i) = P(F_i/Yes) *P(Yes)/P(F_i)$
   - 4.2 $FST_i[Fi, No]= P(No/F_i) = P(F_i/No) *P(No)/P(F_i)$
5. From above generated tables FSTi in Step-4
   - 5.1 Select Maximum-Value (FSTi[Fi, Yes])
   - 5.2 if Similar ($FST_i[F_i,Yes],FST_{i+1}[F_i,Yes]$)=true)
     - 5.2.1 ComputeConfidence (such as EQR1, EQR2... are equal value rows in FST)
       - Compute EQR1 = $FST_i[F_i, Yes]$,*STD ($FST_i[F_i, Yes]$, $FST_i[F_i, No]$) *No of records
       - Compute EQR2 = $FST_i[F_i, Yes]$,*STD ($FST_i[F_i, Yes]$, $FST_i[F_i, No]$) *No of records
   - 5.3 select Root_Node= (Min(EQR1,EQR2)) for Decistion_Tree(DT);
6. Repeat Steps 3 to 6 for each Characteristic (Ci) in Dataset (D) with Features(Fi) for select of next node in Decision_Tree(DT).
7. Return (DT)

VII. ALGORITHM EVALUATION AND RESULTS

The proposed FPBDT Algorithm is evaluated against the following sample dataset and results are shown.

Data set: for Customer Car Purchase

Buying price	Maintenance_cost	Person capacity	Lug_Boot	Safety	Class
Vhigh	Vhigh	2	Small	Small	No
Vhigh	Vhigh	2	Small	Med	No
High	High	4	Small	High	Yes
High	High	4	Med	High	Yes
High	High	4	Big	Med	Yes
Med	Low	4	Big	Low	No
Med	Low	4	Big	Med	Yes
Med	Med	2	Small	Med	No
Low	Med	4	Small	Med	Yes
Low	Med	4	Small	High	Yes

Table: 1 showing Dataset for buying car to build decision tree.

For each feature above FFT is constructed as below for each of the possible outcomes.

Feature Frequency Table (FFT) for Buying Price:

Buying Price	NO	Yes	Probability
Vhigh	2	0	2/10=0.2
High	0	3	3/10=0.3
Med	2	1	3/10=0.3
Low	0	2	2/10=0.2
Total	4/10=0.4	6/10=0.6	

Table: 2 showing frequency of feature buying price

A). Vhigh :

1.P(yes/Vhigh)=P(Vhigh/yes)*P(Yes)/P(Vhigh)

P(Vhigh/yes)=0/6=0 Total Yes=6

P(Yes)=0.6 Total No=4

P(Vhigh)=0.2

P(yes/Vhigh)=0*0.6/0.2=0

2).P(No/Vhigh)=P(Vhigh/No)*P(NO)/P(Vhigh)

P(Vhigh/No)=2/4=0.5 Total Yes=6

P(No)=0.4 Total No=4

P(Vhigh)=0.2

P(No/Vhigh)=0.5*0.4/0.2=1

B). high :

1.P(yes/high)=P(high/yes)*P(Yes)/P(high)

P(high/yes)=3/6=0.5 Total Yes=6

P(Yes)=0.6 Total No=4

P(high)=0.3

P(yes/high)=0.5*0.6/0.3=1

2).P(No/high)=P(high/No)*P(NO)/P(high)

P(high/No)=0/4=0 Total Yes=6

P(No)=0.4 Total No=4

P(high)=0.3

P(No/high)=0*0.4/0.3=0

C). Med:

1.P(yes/Med)=P(Med /yes)*P(Yes)/P(Med)

P(Med /yes)=1/6=0.16 Total Yes=6

P(Yes)=1 Total No=4

P(Med)=0.3

P(yes/Med)=0.16*0.6/0.2=0.53

2). P(No/ Med)=P(Med **/No)*P(NO)/P(**Med**)**

P(Med /No)=2/4=0.5 Total Yes=6

P(No)=2 Total No=4

P(Med)=0.3

P(No/ Med)=0.5*2/0.3=3.33

D). Low:
1.P(yes/Low)=P(Low /yes)*P(Yes)/P(Low)
P(Low/yes)=2/6=0.3 Total Yes=6
P(Yes)=2 Total No=4
P(Low)=0.2
P(yes/Low)=0.3*0.6/0.2=0.09
2).P(No/ Low)=P(Low /No)*P(NO)/P(Low)
P(Low /No)=0/4=0 Total Yes=6
P(No)=0.4 Total No=4
P(Low)=0.2
P(No/ Low)=0*0.4/0.2=0

Feature selection Table (FST) of Buying Price:

Buying Price	**YES**	**NO**
Vhigh	0	1
High	1	0
Med	0.53	3.33
Low	*0.09*	*0*

Table-3: showing Feature Selection values for Buying Price.

We can observe that the above table has no common values if we observe any common values in the above table Yes having the same highest values so we can apply the Confidence values as given below:

Standard deviation:
Let High=0.4242 for High feature value (STD(0.6,0)) since 0.6 is same for High and Low.
Low=0.1909 for Low feature value (STD(0.6,0.33))

Confidence:
High=Yes value*STD*No of records=0.6*0.4242*10=0.703
Low=Yes value*STD*No of records=0.6*0.1909*10=0.1365

By comparing above two values and we take it as the smallest value in confidence which is eligible to the Node.
Similar computations are done for remaining features for each feature FFT, FST is constructed and next node in branch node is considered for decision tree construction. Following decision tree is built given in figure is constructed.

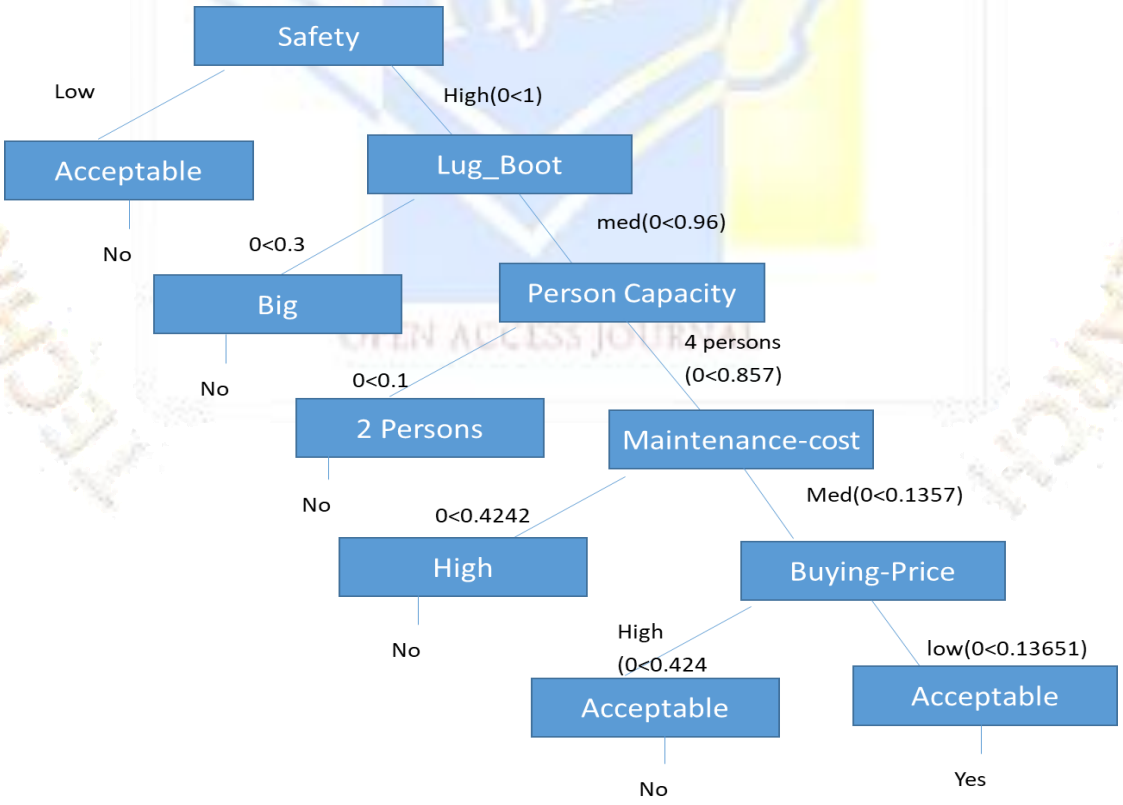

Figure: 2 Decision tree for Buying car Dataset.

FSBDT algorithm even though uses all features for computing with all possibilities, computations are simple without logarithmic or calls to library functions which is overhead in already decision tree algorithms and can be implemented with simple Array structures or collections in Java or Python etc.

VIII. CONCLUSION

Decision tree algorithm like ID3, C4.5,CART etc., has modified various ways due to disadvantages like complexity in calculations, memory usage, unable to manage large datasets, decision tree may not be stable in some situations are some of the points considered and designed the proposed Feature Probability Based Decision Tree (FPBDT) Algorithm which uses probability computations for features of data set to decide the in node selection in decision tree. Algorithm is evaluated with sample car data set to decide to purchase car or not and results are given. The implementation of algorithm is less complex in terms of computations will be published in next article.

IX. REFERENCES

[1] UM Fayyad, G Piatetshy-Shapiro, P Smyth, and R Uthurusamy, Advances in Knowledge Discovery and Data Mining AAAI/MIT Press, 1996
[2] G. Piatetsky-Shapiro and W.J. Frawley, Knowledge Discovery in Databases. AAAI/MIT Press, 1991.
[3] A. Silberschatz, M. Stonebraker, and J.D. Ullman, "Database Research: Achievements and Opportunities into the 21st Century," Report NSF Workshop Futuve of Database Systems Research, May 1995.
[4] Brachman, R., and Anand, T. 1996. The Process of Knowledge Discovery in Databases: A Human-Centered Approach. In Advances in Knowledge Discovery and Data Mining, 37–58, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Menlo Park, Calif.: AAAI Press.
[5] Fayyad, U. M.; Piatetsky-Shapiro, G.; and Smyth, P.1996. From Data Mining to Knowledge Discovery: An Overview. In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 1–30. Menlo Park, Calif.: AAAI Press.
[6] Azevedo, Ana, and Manuel Filipe Santos. "KDD, SEMMA and CRISP-DM: a parallel overview." IADS-DM (2008).
[7] Sen, P.C., Hajra, M., Ghosh, M. (2020). Supervised Classification Algorithms in Machine Learning: A Survey and Review. In: Mandal, J., Bhattacharya, D. (eds) Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing, vol 937. Springer, Singapore. https://doi.org/10.1007/978-981-13-7403-6_11
[8] Weiss, S. I., and Kulikowski, C. 1991. Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems. San Francisco, Calif.: Morgan Kaufmann.
[9] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition, Morgan Kaufmann Publishers is an imprint of Elsevier 2006.
[10] Kotsiantis, S.B. Decision trees: a recent overview. Artif Intell Rev 39, 261–283 (2013). https://doi.org/10.1007/s10462-011-9272-4
[11] Quinlan, J. 1992. C4.5: Programs for Machine Learning.
[12] J.R.Quinlan, "Induction of Decision Tree", Machine Learning Vol -1 ,pp, 81-106, 1986
[13] A. Rajeshkanna, 2K. Arunesh, ID3 Decision Tree Classification: An Algorithmic Perspective based on Error rate, Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020)IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281-4108-4
[14].Jin Jing. Optimization of sorting algorithm based on Map Reduce model [J]. Computer science, 2017,41 (12): 155-159
[15] ZHU M, SHEN D, YU G, et al. Computing the Split Points for Learning Decision Tree in Map Reduce. Database Systems for Advanced Applications, Lecture Notes in Computer Science 7826, 339–353 (2013)
[16].Chien-Liang Lin & Ching-Lung Fan (2019) Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan, Journal of Asian Architecture and Building Engineering, 18:6, 539-553, DOI: 10.1080/13467581.2019.1696203
[17] B.R. Gains, "Transforming Rules and Tree into comprehensive knowledge structures" U.M. Fayyad, G. Platetsky-shapiro, p.syth and R.Uthurusamy, eds., Advances in knowledge discovery and data mining , pp205-228, AAAI/MIT Press, 1996
[18].R.Agrawal, S.Ghosh,T. Imielinski, B.Iyer and A.Swami, "An Interval classifier for Data mining Applications", proc, 18th Intl conf, Very Large Databases pp.560-573, Aug,1992.
[19].Jui-Sheng Chou, Shu-Chien Hsu, Chih-Wei Lin, Yu-Chen Chang, Classifying Influential Information to Discover Rule Sets for Project Disputes and Possible solutions, International Journal of Project Management, Volume 34, Issue 8,2016,Pages 1706-1716,ISSN 0263-7863
[20].R.m Chai and M. Wang, "A more efficient classification scheme for ID3," 2010 2nd International Conference on Computer Engineering and Technology, Chengdu, China, 2010, pp. V1-329-V1-332, doi: 10.1109/ICCET.2010.5486128.
[21] Chen Jin, Luo De-lin and Mu Fen-xiang, "An improved ID3 decision tree algorithm," 2009 4th International Conference on Computer Science & Education, Nanning, 2009, pp. 127-130, doi: 10.1109/ICCSE.2009.5228509.
[22].L. Yi-bin, W. Ying-ying, and R. Xue-wen, Improvement of ID3 algorithm based on simplified information entropy and coordination degree, in 2017 Chinese Automation Congress (CAC), 2017, pp. 1526–1530.
[23].Ibomoiye Domor Mienye, Yanxia Sun, Zenghui Wang, Prediction performance of improved decision tree-based algorithms: a review, Procedia Manufacturing, Volume 35, 2019, Pages 698-703,ISSN 2351-9789,https://doi.org/10.1016/j.promfg.2019.06.011.
[24].L. Fang, H. Jiang, and S. Cui, An improved decision tree algorithm based on mutual information, in 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), 2017, pp. 1615–1620.
[25].Stuart L. Crawford, Extensions to the CART algorithm, International Journal of Man-Machine Studies, Volume 31, Issue 2, 1989, Pages 197-217, ISSN 0020-7373
[26].F. Chen, X. Li, and L. Liu, Improved C4.5 decision tree algorithm based on sample selection, in 2013 IEEE 4th International Conference on Software Engineering and Service Science, 2013, pp. 779–782.