Prediction and Classification of Data Mining using Decision Tree

Sachin Kumar[1], Ashwani Kumar[2]

[1]CSE DEPT., HIERANK BUSINESS SCHOOL,NOIDA AFFILIATED TO CCS UNIVERSITY, MEERUT

[2]RESEARCH SCHOLAR, FACULTY OF CSE, JAGANNATH UNIVERSITY, NCR, HARYANA.

Abstract

Decision trees are graphs that are tree-like and model a decision. The questioners have to guess the object by asking as many as 20 questions as they can answer with yes, no, or perhaps. Asking questions of increasing specificity is an intuitive strategy for questioners; asking "is it a musical instrument?"As the first question, the number of possibilities will not be effectively reduced. A decision tree branches specify the shortest sequences of explanatory variables that can be examined to estimate a response variable's value. It is a commonly used data mining method for establishing multiple covariate-based classification systems or for developing target variable prediction algorithms. It classifies a population into a branch–like segments that build a root node inverted tree, internal nodes, and leaf nodes. It is non-parametric and can handle big, complex datasets effectively without imposing a complex parametric framework. When sample size is large enough, study data can be divided into training and validation datasets. Use the training dataset to create a decision tree model and a validation dataset to determine the appropriate tree size required to achieve the optimal final model. In order to obtain optimum medical result, this article presents diagnostic test of therapy approach. The prediction and classification of better information using datamining is discussed here.

Key words: decision tree; data mining; classification; prediction.

1. Introduction

Decision trees are frequently learned by repetitively dividing the set of training cases into sub-sets based on the explanatory variables ' instance values. A decision tree represented by boxes and the inner nodes of the explanatory variables of the decision tree experiment. These nodes are linked by edges specifying the possible test results. Based on the test results, the training cases are split into subsets. For example, a node could test whether or not a threshold exceeds the value of an explanatory variable. The instances that pass the test will follow an edge to the node's right child, and the instances that fail the test will follow an edge to the node's left child. The kids nodes test their training sub-sets similarly until a stop criterion is met. The leaf nodes of the decision tree depict classes in classification assignments. In prediction tasks, the response variable values may be averaged for instances contained in a leaf node to produce the response variable estimate.Following the construction of the decision tree, creating a forecast for a sample example involves only following the edges until reaching a leaf node. Data mining is used to obtain and show helpful data in easy-to-interpret visualizations from big datasets. Decision trees were first implemented in the 1960s and are one of the most efficient techniques for information mining; they have been commonly used in several fields because they are simple to use, free of ambiguity and robust even when missing values are present. It is possible to use both discrete and continuous variables either as target variables or as independent variables. More lately, in medical research, decision tree methodology has become popular. An instance of the medical use of decision trees is the diagnosis of a symptom-based medical condition in which the classes identified by the decision tree may be either distinct clinical subtypes or a situation, or patients with a situation that should receive distinct therapies.

Tree decision tree can be used to predict and classify information mining to select the most appropriate variables to be used to create decision tree models that can then be used to formulate clinical hypothesis and inform subsequent studies. Variables play a significant role in determining information precision when removing the variable.The more conditions the more a variable record will affect the higher the variable's significance.Decision tree can handle missing information in two ways, either classifying missing values as distinct categories that can be analyzed with other categories or using a built-in decision tree model that sets the variable with lots of missing value as a target variable to predict and replace those missing with the expected value.The outcome for future records must be obtained from historical information and simple to predict. It can help in deciding how categorical variables can best collapse into a more manageable number of categories or how to subdivide heavily skewed variable into ranges.

2. Literature Review

Chye Koh et.al, claimed that the current concern is a fundamental concept in accounting and auditing, and it is not an easy task to determine the status of a company's current concern. Several prediction models are proposed in the literature based on statistical methods to assist auditors. This research examines and contrasts the utility of neural networks, decision trees, and logistic regression in predicting a company's status of concern. The survey data includes financial ratios of 165 current concerns and 165 non-going concerns compared. The results of the classification suggest the potential utility of data mining techniques in a sense of ongoing prediction.In contrast, the decision tree that affects the system of prediction outperforms the models of logistic regression and neural networks. Data mining techniques such as neural networks and decision trees are effective in analyzing complex non-linear and interaction relationships and can therefore supplement and complement conventional statistical methods in the production of predictive models [1].

Abu-Salih et. al, aims to acquire the textual content domain created by online social network (OSN) platform users. Knowing thedomain(s) of interest of a client is an important step towards resolving their domain-based trustworthiness by accurately knowing their OSN material.This study uses a Twitter mining approach to identify domain-based users and their textual content. The proposed solution requires modules for machine learning. The approach comprises two analysis phases: the time-aware semantic analysis of users' historical content incorporating five commonly used machine learning classifiers.The system classifies users into two major categories: categories related to politics and categories related to non-politics. In the second stage, the probability predictions obtained in the first phase are used to forecast the domain of tweets for future users [2].

Sohrabi, B. et. al, claimed that pharmaceutical industry, marketing and sales managers frequently deal with massive amounts of marketing and sales information. One of their major concerns is to consider the effect of actions taken on sold-out goods.Data mining finds and removes valuable patterns to find hidden and worthy patterns for decision-making from such large data sets. This paper also seeks to demonstrate the ability of the method of data-mining to improve the quality of decision-making in the pharmaceutical industry. This work is informative in terms of the tool used, as well as analyzing the current situation and using and explaining real data. In reality, from the perspective of its data type and process, the analysis is qualitative and descriptive. In terms of intent, this work is also relevant.Data from a pharmaceutical company in Iran are the target population of this study. For data mining and data processing, the cross-industry standard system for data mining methodology was used [3].

Trivedi, S. et. al believed that the classification of email spam is now becoming a challenging field in the text classification domain. Correct and reliable classifiers are judged not only by accuracy of classification, but also by sensitivity and specificity to correct classification, captured by both false positive and false negative levels.This paper aims to present a comparative study with / without different boosting algorithms (bagging, boosting with re-sampleand AdaBoost) between different decision tree classifiers (such as AD tree, decision stump and REP tree). This research combined artificial intelligence and text mining approaches.Each decision tree classifier in this study is tested on informative words / features selected from the two publically available data sets (SpamAssassin and LingSpam) using a greedy step-wise feature search method [4].

You, y et. al, claimed that the goal is to implement a web-based data mining application that incorporates online data collection and data mining into online auction sales strategies. This research aims to demonstrate the process of collecting online spider information from eBay and applying the Classification and Regression Tree (CART) to create successful sales strategies.The four stages of online spider data collection and CART data mining were demonstrated after creating a model for web-based data mining. The spiders can effectively and efficiently gather online auction information from the internet in the web-based data mining software, and the CART model offers successful sales strategies for sellers. Sellers can incorporate their two primary goals, i.e. auction performance and anticipated prices, into their online auction sale strategies by using projected auction prices with the classification and regression trees. Practical Implications–This study provides sellers with a useful tool by taking advantage of web-based data mining to develop successful sales strategies.Such successful selling strategies can help improve the performance of their online auction. This study contributes to the literature by offering a groundbreaking method for online data collection and effective selling strategies that are critical for the growth of electronic marketplaces[5].

Kuzey, C. et. al, claimed that the goal was to define and critically analyze factors affecting cost system functionality (CSF) using several machine learning techniques including decision trees, supporting vector machines and logistical regression. The study used a self-administered survey method to gather the data needed from businesses in Turkey. Several predictive models are developed and tested; a series of sensitivity analyzes are carried out on the predictive models developed to determine the factors / variables ranked in importance [6].

Lee,S. et. al stated that to suggest important determinants of the usefulness of product data, review characteristics and textual characteristics of the reviews and to identify the most important factors among these determinants using statistical methods. In addition, this study aims to suggest a recommender for classification-based evaluation using a decision tree (CRDT) that uses a decision tree to classify and recommend reviews that have a high level of helpfulness. This research used Amazon.com's publicly available data to construct determinants and helpfulness steps.The authors collected economic transaction data on Amazon.com to examine this and analyzed the associated review system. The final sample included 10,000 reviews with 4,799 helpful and 5,201 unsuitable reviews [7].

Yan-yan SONG et.al explained that businesses manage their most valuable assets–information obtained from consumers and customers using data mining tools that sift through massive amounts of data and find hidden information–to help them better understand customers and predict their behaviour.The aim of this paper is to address data mining approaches in Oracle, widely used for large corporate companies, and applications for Microsoft data mining, commonly used in SMEs. It discusses Oracle9i and Microsoft Data Mining algorithms that provide an efficient, scalable application-building infrastructure that automates business intelligence extraction and integration into other applications.It addresses the strengths and weaknesses of data mining tools within Oracle9i and Microsoft, demonstrating how smart tools support various business and industry sectors and scales [8].

3. Decision Tree Concepts

Decision trees are commonly learned by dividing the set of training instances into subsets on the basis of the values of the instance for the explanatory variables recursively. A decision tree is shown in the following diagram. Represented by boxes, the interior nodes of the decision tree test best explanatory variables.Such nodes are linked by edges defining the possible test results.The instances passing the test follow an edge to the right child of the node and the instances failing the test follow an edge to the left child of the node. The kids nodes evaluate their learning sub-sets similarly until a stop criterion is met. The leaf nodes of the decision tree represent groups in classification tasks.In regression tasks, the response variable values may be multiplied for instances found in a leaf node to generate the response variable estimate. After constructing the decision tree, predicting a test instance involves only following the edges before hitting a leaf node.Three forms of nodes exist. A root node also called a decision node is an option that will result in all records being subdivided into two or more mutually exclusive subsets. Internal nodes, also known as chance nodes, are one of the possible choices in the tree structure at that level, the top edge of the node is connected to the parent node and the

bottom edge is connected Leaf nodes also called end nodes, represent the final result of a combination of decisions or events. Branches represent chance outcomes or occurrences that emanate from root nodes and internal nodes. Using a branch hierarchy, a decision tree is created. That path from the root node to a leaf node by internal nodes is a rule of classification. It is also possible to describe such decision tree structures as' if-then' laws. For example, "if condition 1 and condition 2 and condition..... and condition k occur, then outcome j will occur."

Only target-related input variables are used to divide parent nodes into purer target-variable child nodes.The most important input variables must first be defined when constructing the template, and then the root node and subsequent internal nodes must be divided into two or more classes or' bins' dependingon the status of these variables. Features correlated with the degree of' purity' of the resulting child nodes

Used to choose from various potential input variables: these include entropy, Gini index, grouping, benefit ratio, and requirements.In most cases, not all possible input variables will be used to construct the decision tree, and a similar input variable may be used several times at different decision tree rates in some cases. Stopping is the complexity and robustness of competing models that must be considered concurrently whenever a numerical model is constructed.When used to predict future data, the more complex a model is, the less accurate it will be. An extreme situation is to create a very complex decision tree that spreads wisely enough to make the records 100% pure in each leaf node. Such a decision tree would be excessively suited to the current findings and have few records in each branch, so it could not predict future cases accurately and would therefore be poorly generalized. Stoping must be implemented when constructing a decision tree to prevent the design from becoming overly complex in order to prevent this from happeningStopping rules doesn't fit well in some cases. An alternative way of building a decision tree is to first grow a large tree and then prune it to optimum size by removing nodes that provide less data. Including the proportion of records with error estimation is a common method of selecting the best possible subtree from several candidates.Certain approaches to select the best

alternative are using a test database and cross-validation of small samples. Pruning, pre-pruning and post-pruning are two styles. Pre-pruning uses Chi-square or multiple tests–methods of comparison adjustment to prevent non-significant branches from being produced.After creating a complete decision tree, post-pruning is used to extract branches in a way that increases the overall classification accuracy when applied to the validation dataset.

4. Algorithms used for building a decision tree and their analysis

Several statistical algorithms for building decision trees are here

CHID: CHAID (CHi-squared Automatic Interaction Detector) is a simple learning algorithm for the decision tree. Gordon V Kass[4] established it in 1980. CHAID is easy to interpret, easy to handle, and can be used to define and identify variables interaction. CHID is an extension of the AID and THAID procedures (Theta Automatic Interaction Detector). It operates on changed sense screening principal. After detecting interaction between variables, it selects the best attribute to divide the node that made a child node as a set of the selected attribute's homogeneous values. The system can accommodate values that are absent. It does not require any form of pruning.

CART: Classification and regression tree (CART) introduced by Breimanet al.[5] constructs binary trees often referred to as HODA. CART is a non-parametric tree learning technique that either generates classification or regression trees, depending on whether the dependent variable is either categorical or numerical.The word binary means that a node can only be split into two classes in a decision tree. CART uses the Gini index as a metric of impurity for attribute collection. The attribute with the greatest impurity reduction is used to break the records of the node. CART embraces numerical and categorical data and treats incomplete values of attributes. This uses pruning of cost-complexity and also produces regression trees.

ID3: Quinlan introduces the decision tree algorithm ID3 (Iterative Dichotomiser 3) [6]. The knowledge gain approach is usually used in the decision tree method to determine the appropriate property for each node of a decision tree created.Therefore, as the test attribute of the current node, we can pick the attribute with the highest information gain (entropy reduction in the maximum level). In this way, it will be the smallest information needed to identify the training sample subset obtained from subsequent partitioningThat is, using this property to partition the sample set in the current node will reduce to a minimum the mixing degree of different types for all sample subsets generated. Therefore, using such an approach to information theory would effectively reduce the necessary dividing number of classification of objects.

C4.5: C4.5 is an algorithm used by Ross Quinlan to construct a decision tree. C4.5 is an expansion of the earlier ID3 algorithm from Quinlan. It is possible to use the decision trees created by C4.5 for classification, which is why C4.5 is often referred to as a statistical classifier[7]. The C4.5 algorithm uses gathering data as criterion for splitting. With categorical or numerical values, it can accept data. This produces threshold to accommodate continuous values and then separates attributes above the threshold with values equal to or below the threshold. Missing values can be easily handled by C4.5algorithm. As missing values of attributes are not used inC4.5 gain calculations.

C5.0/Sec 5: C5.0 algorithm is a C4.5 algorithm variant that is also an ID3 extension. It is the algorithm of classification that applies to the big data set. On speed, memory and performance, it's better than C4.5. Model C5.0 works by splitting the sample based on the field providing the full gain of data.Centered on the largest information gain region, the C5.0 model will break samples. The test subset derived from the previous split would subsequently be split. The process will continue until it is not possible to split the sample subset and is usually based on a different field.Eventually, analyze the lowest split point, refusing certain sample subsets that have no significant contribution to the model. The multi value attribute and missing attribute from the data set can be easily handled by C5.0 [8]

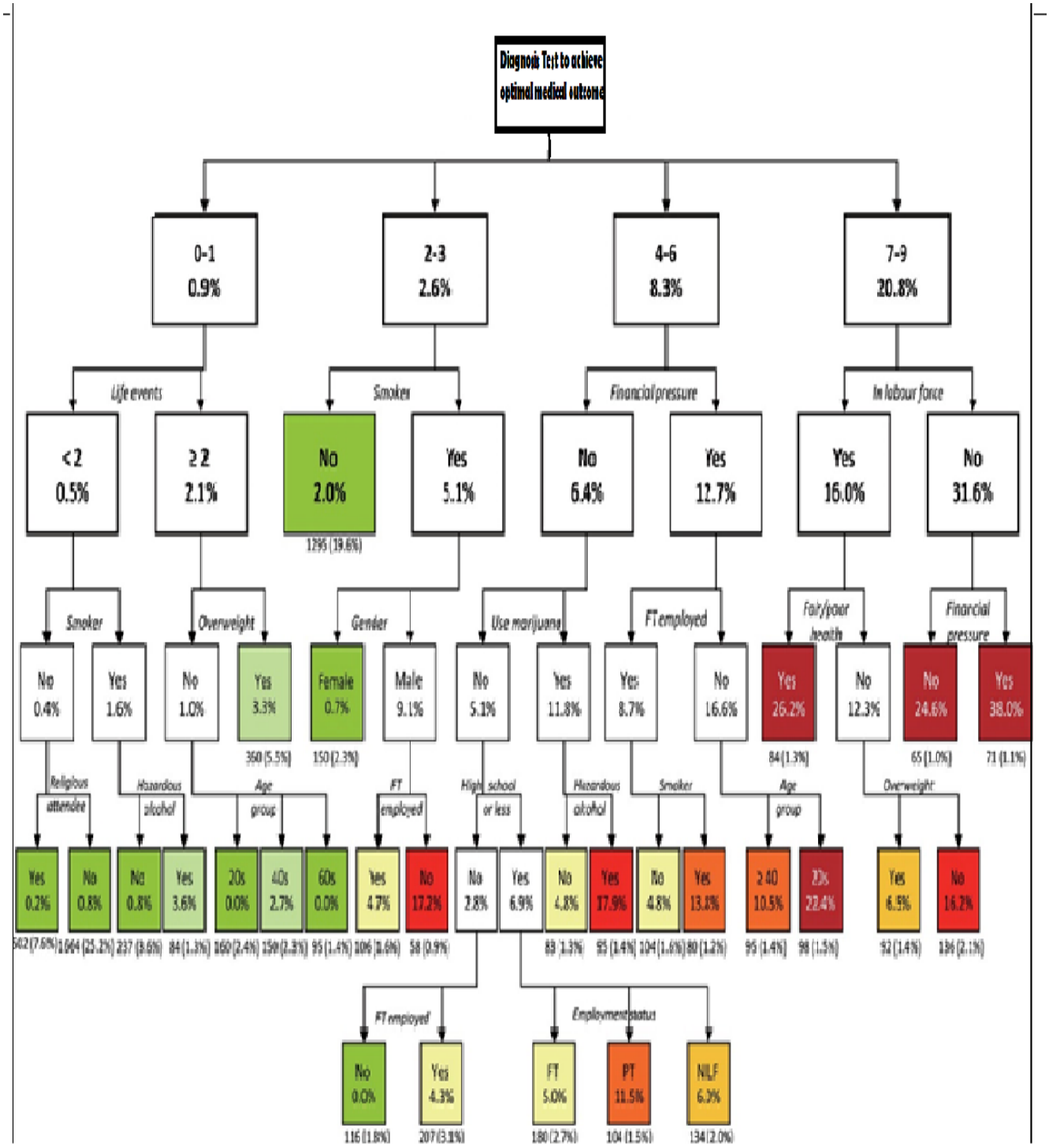
Hunt's Algorithm: Hunt's algorithm builds a tree of judgment by top-down method and divides and conquers. The data from the sample / row includes more than one group, use a check attribute to break the data into smaller subsets. Hunt's algorithm preserves optimal splitting for each stage based on a certain threshold value as greedy fashion [9]

Table 1 Comparisons between different Decision Tree Algorithm

	ID3	C4.5	C5.0	CART
Type of data	Categorical	Continuous and Categorical	Categorical Continuous and Categorical, dates, times, timestamps	Continuous and nominal attributes data
Speed	Low	Faster than ID3	Highest	Average
Pruning	No	Pre-pruning	Pre-pruning	Post-Pruning
Boosting	Not Supported	Not supported	Supported	Supported
Missing Values	Can't deal with	Can't deal with	Can't deal with	Can't deal with
Formula	Use information entropy and information Gain	Use split info and gain ratio	Use split info and gain ratio	Use Gini Diversity Index

5. Supported Example

To demonstrate the development of a decision tree model, risk factors relevant to major diagnostic test or treatment plan are evaluated here to achieve optimal health outcome. The purpose of the study was to determine the most significant risk factors from a pool of 17 potential risk factors, including life events, smoker, financial pressure, labor force, smoker, overweight, race, use of cannabis, IT employed, fair / poor health, financial pressure, religious attendance, hazardous alcohol, age group, FT employed, high school or less, smoker, overweight, employment status and so forth. The decision tree generated from the dataset is shown in fig. 2



6. **The advantages and disadvantages of decision trees**

Decision trees are user friendly. In contrast to many learning algorithms, decision trees do not require zero mean and unit variance from the results. While for explanatory variables, decision trees can tolerate missing values. Decision trees can even learn to ignore non-relevant explanatory variables.With the export graphviz feature from the scikit-learn tree module, small decision trees can be easily interpreted and visualized. A decision tree's branches are combinations of logical assumptions and are easily visualized as flowcharts. Decision trees accept multi-output tasks and a single decision tree can be used to define multi-class without using a one-versus-all approach.

Trees of decision-making are willing students. Willing learners need to build an input-independent model from the training data before they can be used to estimate test instance values, but they can predict relatively quickly once the model is built. In comparison, lazy learners like the algorithm of k-nearest neighbors delay all generalization until they have to make a prediction. Although lazy learners do not spend time learning, they often gradually predict compared to eager learners. Decision trees are more likely to overfit than many models because their learning algorithms can produce large, complicated decision trees that perfectly model that training instance but fail to generalize the actual relationship. Pruning is a common strategy that eliminates some of a decision tree's highest nodes and leaves.Similar effects can, however, be accomplished by setting a maximum tree depth and generating child nodes only when the number of training instances they must contain reaches a threshold. Through making locally optimal decisions, they learn effectively, but are not guaranteed to generate the worldwide optimal tree. ID3 builds a tree by choosing an explanatory variables sequence to be evaluated. Every explanatory variable is chosen as it reduces the node uncertainty more than the other variables. However, in order to find the globally optimal tree, it is possible that locally suboptimal tests are needed.Some more advantages: -

- Simplifies complex relationships between input variables and target variables by dividing original input variables into significant subgroups.
- Easy to understand and interpret.
- Non-parametric approach without distributional assumptions.
- Easy to handle missing values without needing to resort to imputation.
- Easy to handle heavy skewed data without needing to resort to data transformation.
- Robust to outliers.

The main disadvantage is that it can be subject to overfitting and underfitting, particularly when using a small data set. This problem can limit the generalization and robustness of the resultant model.

7. Conclusion

One implication is to be careful in evaluating decision tree models and in forming casual hypotheses using the effects of these models. In this case, the choice of input variables is based on statistical properties, but the selection of input variables in real-world can be based on the relative cost of obtaining the variables or on the clinical significance of the variables. Another application of the decision tree approach is to create a decision tree that distinguishes patient subgroups that should have specific diagnostic tests and treatment approaches in order to achieve optimal medical results.

8. References

[1] Chye Koh, H. and Kee Low, C. (2004), "Going concern prediction using data mining techniques", Managerial Auditing Journal, Vol. 19 No. 3, pp. 462-476. https://doi.org/10.1108/02686900410524436

[2] Abu-Salih, B., Wongthongtham, P. and Chan, K. (2018), "Twitter mining for ontology-based domain discovery incorporating machine learning", Journal of Knowledge Management, Vol. 22 No. 5, pp. 949-981. https://doi.org/10.1108/JKM-11-2016-0489

[3] Sohrabi, B., RaeesiVanani, I., Nikaein, N. and Kakavand, S. (2019), "A predictive analytics of physicians prescription and pharmacies sales correlation using data mining", International Journal of Pharmaceutical and Healthcare Marketing, Vol. 13 No. 3, pp. 346-363. https://doi.org/10.1108/IJPHM-11-2017-0066

[4]Trivedi, S. and Panigrahi, P. (2018), "Spam classification: a comparative analysis of different boosted decision tree approaches", Journal of Systems and Information Technology, Vol. 20 No. 3, pp. 298-105. https://doi.org/10.1108/JSIT-11-2017-0105

[5] Tu, Y. (2008), "An application of web-based data mining: selling strategies for online auctions", Online Information Review, Vol. 32 No. 2, pp. 147-162. https://doi.org/10.1108/14684520810879791

[6] Kuzey, C., Uyar, A. and Delen, D. (2019), "An investigation of the factors influencing cost system functionality using decision trees, support vector machines and logistic regression", International Journal of Accounting & Information Management, Vol. 27 No. 1, pp. 27-55. https://doi.org/10.1108/IJAIM-04-2017-0052

[7] Lee, S. and Choeh, J. (2017), "Exploring the determinants of and predicting the helpfulness of online user reviews using decision trees", Management Decision, Vol. 55 No. 4, pp. 681-700. https://doi.org/10.1108/MD-06-2016-0398

[8] Hanna, M. (2004), "Data-mining algorithms in Oracle9i and Microsoft SQL Server", Campus-Wide Information Systems, Vol. 21 No. 3, pp. 132-138. https://doi.org/10.1108/10650740410544036

[9] Yan-yan SONG, Ying LU (2015), "Decision tree methods: applications for classification and prediction", Shanghai Archives of Psychiatry, Vol. 27, No. 2, pp. 130-135.

[10] Surjeet Kumar Yadav, Saurabh Pal(2012), " Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification ", World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 2, 51-56

[11] Bhaskar N. Patel, Satish G. Prajapati and Dr. Kamaljit I. Lakhtaria (2012), " Efficient Classification of Data Using Decision Tree ", Bonfring International Journal of Data Mining, Vol. 2, No. 1

[12] Lakshmishree J, K Paramesha(2017), " Prediction of Heart Disease Based on Decision Trees ", International Journal for Research in Applied Science & Engineering Technology (IJRASET), Volume 5 Issue V, May 2017, ISSN: 2321-9653

[13] Brijain R Patel, Kushik K Rana (2014), " A Survey on Decision Tree Algorithm For Classification ", IJEDR | Volume 2, Issue 1 | ISSN: 2321-9939

[14] A.Rajeshkanna, K.Arunesh (2018), " Role of Decision Tree Classification in Data Mining", International Journal of Pure and Applied Mathematics, Volume 119 No. 15 2018, 2533-2543

Book

[1] Gavin Hackeling (2014)," Mastering Machine Learning with scikit-learn", Apply effective learning algorithms to real-world problems using scikit-learn, Packt Publishing Ltd.