International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

The Review of Attributes Influencing Housing Prices using Data Mining Methods

Pelin Kasap[a*], Burçin Şeyda Çorba[b]

[a,b]*Department of Statistics, Ondokuz Mayıs University, Samsun, Turkey*

[a]*Email: pelin.kasap@omu.edu.tr*

[b]*Email: burcinseyda.corba@omu.edu.tr*

Abstract

Prices of housing show an alteration throughout the world. The reason of this is characteristics of housing such as age, tax, number of rooms, per capita crime rate by town and so on. In this study, we carry out the stages of the CRISP-DM process to investigate attributes influencing prices of housing. We use CART, C5 decision tree algorithms and Neural Networks model in modeling phase. Also, we use cross-validation method in modeling evaluation phase. It is shown that C5 model is the most appropriate model with the highest validation rate.

Keywords: Data mining; CRISP-DM; CART algorithm; C5 algorithm; Neural Networks model.

1. Introduction

Data Mining (DM) is the exploration and analysis of large quantities of data so as to explore novel, valid, useful and comprehensible patterns in data [19, 22]. However, there are various definitions of DM. These definitions depend on the point of views of the describer. Here are some definitions in the literature as follows:

- DM is the nontrivial process of describing valid, novel, useful and understandable patterns in data - *Fayyad* [4,5].
- DM is the process of extracting previously unknown and understandable information from large databases –*Zekulin* [13,25].

---

* Corresponding author.

- DM is a set of methods used in the knowledge discovery process to distinguish previously unknown relationships and patterns within data –*Ferruzza* [7].
- DM is the process of discovering advantageous patterns in data- *John* [10].
- DM is a decision support process, where we look at large databases for unknown and unexpected patterns of information- *Parsaye* [6, 13, 16].

DM is a *process* of discovering various models, summaries, and derived values from a given collection of data. The general experimental procedure adapted to DM problems involves the following steps:

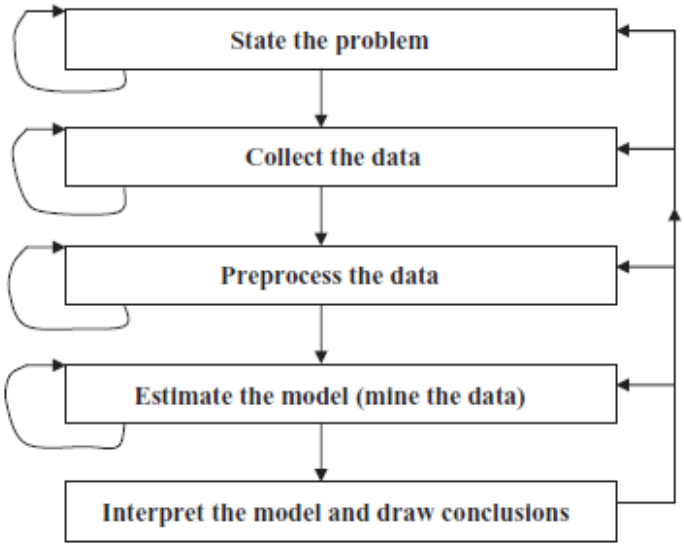

Figure 1: The data mining process [11]

Introducing a DM application into an organization is essentially not very different from any other software application project, and the following conditions have to be satisfied:

- There must be a well-defined problem.
- The data must be available.
- The data must be relevant, adequate, and clean.
- The problem should not be solvable by means of ordinary query or online analytical processing (OLAP) tools only.
- The results must be actionable [11].

The induction of decision rule sets that are accurate, simple and understandable is a fundamental goal of data mining. Decision tree algorithms have proven to be a relatively quick and effective method of induction [14].

In this study, we fulfil the stages of the CRoss Industry Standard Process for Data Mining (CRISP-DM) process to investigate attributes influencing prices of housing. We use decision tree algorithms such as the Classification and Regression Trees (CART) [1], C5 and Neural networks model in modeling phase. Also, we use cross-validation method in modeling evaluation phase.

The rest of this paper is organized as follows: in Section 2, we present CRISP-DM process. In Section 3, we introduce decision tree algorithms such as CART, C5 and in Section 4, we also introduce Neural networks model. In Section 5, we give an application of Boston Housing Data. Finally, we give conclusions in Section 6.

2. CRISP-DM Process

DM is an iterative process that compose of several stages. CRISP-DM method is a popular method for increasing the success of DM process [2]. This method is described in terms of a hierarchical process model. The CRISP-DM process purposes to make large DM projects, less costly, faster, more repeatable, more manageable and more reliable [24].

It is seen that Figure 2, according to CRISP-DM, DM process has a life cycle composed of six stages: Business understanding, data understanding, data preparation, modeling, evaluation and deployment [3]. The order of these stages can be changed, i.e. not fixed.

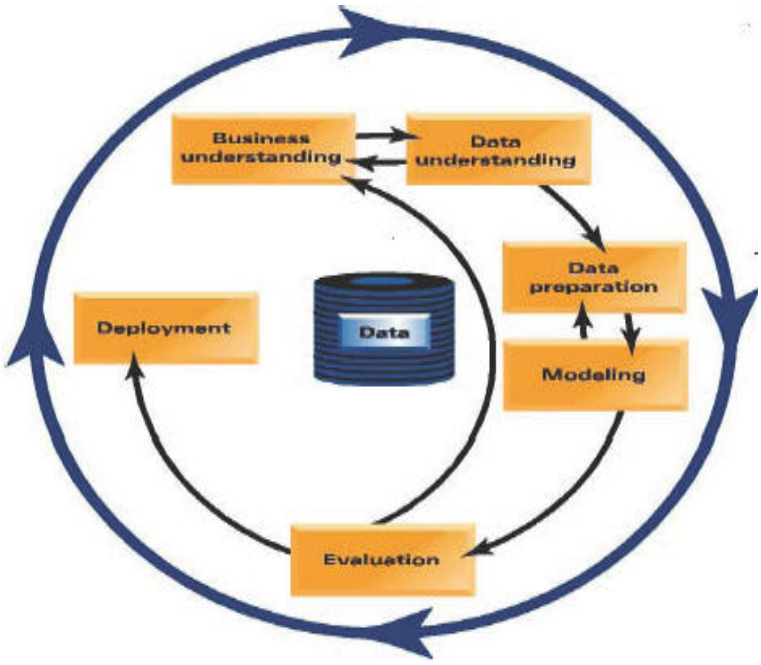

Figure 2: The CRISP-DM Process Model [2].

3. Decision Tree Algorithms

A decision tree is a set of circumstances arranged in a hierarchical structure [15]. It is a predictive model in which an instance is classified by following the path of satisfied conditions from the root of the tree until reaching a leaf, which will correspond to a class label. A decision tree can easily be converted to a set of classification rules [21]. There are two prominent algorithms in the area of classification: CART algorithm and C5 algorithm. These two algorithms result in a decision tree, which is, basically, a collection of decision nodes,

connected by branches, extending downward from the root node until terminating in leaf nodes [12].

3.1. CART Algorithm

CART is a nonrigid method to represent how the variable *Y* distributes after assigning the forecast vector *X*. This model uses the binary tree to divide the forecast space into certain subsets on which *Y* distribution is continuously even. Tree's leaf nodes correspond to different division areas which are designated by Splitting Rules relating to each internal node. By moving from the tree root to the leaf node, a forecast sample will be given an only leaf node, and *Y* distribution on this node also be designated [17]. The binary feature of the CART algorithm produces a very readable decision tree [12]. CART uses GINI Index to determine in which attribute the branch should be generated. The line is to choose the characteristic whose GINI Index is minimum after splitting [17].

3.2. C5 Algorithm

C5 algorithm is the classification algorithm which applies in big data set. C5 model works by splitting the sample based on the field that provides the maximum information gain. The C5 model can split samples on basis of the biggest information gain field. The sample subset that is get from the former split will be split afterward. The process will continue until the sample subset cannot be split and is usually according to another field. Finally, examine the lowest level split, those sample subsets that don't have remarkable contribution to the model will be rejected. [17].

C5 algorithm has following features: The large decision tree can be viewing as a set of rules which is clear. It gives acknowledge on noise and missing data. Problem of over fitting and error pruning is solved by the C5 algorithm. In classification technique the C5 classifier can anticipate which attributes are relevant and which are not relevant in classification [18]. While the CART algorithm produces a binary tree, the C5 algorithm is not restricted to binary splits [12]. C5 algorithm uses Information Gain or Entropy as its criteria.

4. Neural Networks Model

Neural networks are used for pattern recognition, clustering, prediction, classification and outlier detection [20]. In a wide range of applications, neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data. Using neural networks as a tool, data storing firms are harvesting information from datasets in the process known as data mining. The difference between these data storages and ordinary databases is that there is actual manipulation and cross-fertilization of the data helping users makes more informed decisions [Singh and Chauhan, 2009]. Neural networks are programmed to solve combinatorial optimization problems, to control nebulation problems and to filter noise from measurement data [23]. They basically consist three parts: the architecture, the learning algorithm and the activation functions, see also [20,23]. The data mining based on neural network is consist of three stages such as data preparation, rules extracting and rules assessment [8].

5. Application

This study was conducted to examine the prices of houses in Boston according to the conditions of environmental which they are in. The data used in this study were taken from the Statlib Library of Carnegie Mellon University [9]. This data set is well known a data set and taken in 1993. There are 506 observations in this data set. Also, there are 14 features which consist of 13 continuous and 1 categorical. All of these attributes are given in Table 1 (As CHAS is categorical variable, it is not given in Table 1). In data analysis, SPSS-Clementine program is used. Firstly, 6 stages of the CRISP-DM cycle are applied to this data respectively. The first stage of the cycle is *"business understanding phase"*. In this phase, it is aimed to determine the prices of houses in Boston according to the conditions of environmental which they are in. The second phase after business understanding phase is "*data understanding phase*". In this phase, descriptive statistics for 14 variables, which are going to be used to calculate the median value of house prices in Boston, are obtained like in Table 1. The explanation of attributes is as follows: CRIME: per capita crime rate by town; ZN: proportion of residential land zoned for lots over 25,000 sq.ft; INDUS: proportion of non-retail business acres per town; CHAS: Charles River dummy variable (=1 if tract bounds river; 0 othervise); NOX: nitric oxides concentration (parts per 10 million); RM: average number of rooms per dwelling; AGE: proportion of owner-occupied units built prior to 1940; DIS: weighted distances to five Boston employment centres; RAD: index of accessibility to radial highways; TAX: full-value property-tax rate per $10,000; PTRATIO: pupil-teacher ratio bu town; B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town; LSTAT: % lower status of the population; MEDV: Median value of owner-occupied homes in $1000's.

Table 1: Descriptive Statistics for Variables

	Features	Minimum	Maximum	Mean	Standard Deviation	Skewness
1	CRIME	0,006	88,976	3,614	8,602	5,223
2	ZN	0,000	100,000	11,364	23,322	2,226
3	INDUS	0,460	27,740	11,137	6,860	0,295
4	NOX	0,385	0,871	0,555	0,116	0,729
5	RM	3,561	8,780	6,285	0,703	0,404
6	AGE	2,900	100,000	68,575	28,149	-0,599
7	DIS	1,130	12,126	3,795	2,106	1,012
8	RAD	1,000	24,000	9,549	8,707	1,005
9	TAX	187,000	711,000	408,237	168,537	0,670
10	PTRATIO	12,600	22,000	18,456	2,165	-0,802
11	B	0,320	396,900	356,674	91,295	-2,890
12	LSTAT(%)	1,730	37,970	12,653	7,141	0,906
13	MEDV ($)	5,000	50,000	22,533	9,197	1,108

When Table 1 is examined, it is seen that minimum value of crime value is 0.6% and the maximum value is 88,976. It is seen that median values of houses vary between 5000$ and 50.000$. Nitric oxides rate in water is maximum 0,871 and minimum 0,385 and the standard deviation is 0.116. Houses have minimum 4, maximum 8 rooms. Ages of houses in Boston area varies between 3 and 100. The average price of houses is 69. Pupil-teacher ratio according to town is 18%.

According to Figure 3 which shows the distribution of median values for houses in Boston, median values of 212 houses with the highest rate of 41.9% varies between 20.000$ and 30.000$. Median values of 24 houses with the lowest rate of 4.74% varies between $0 and $10.000.

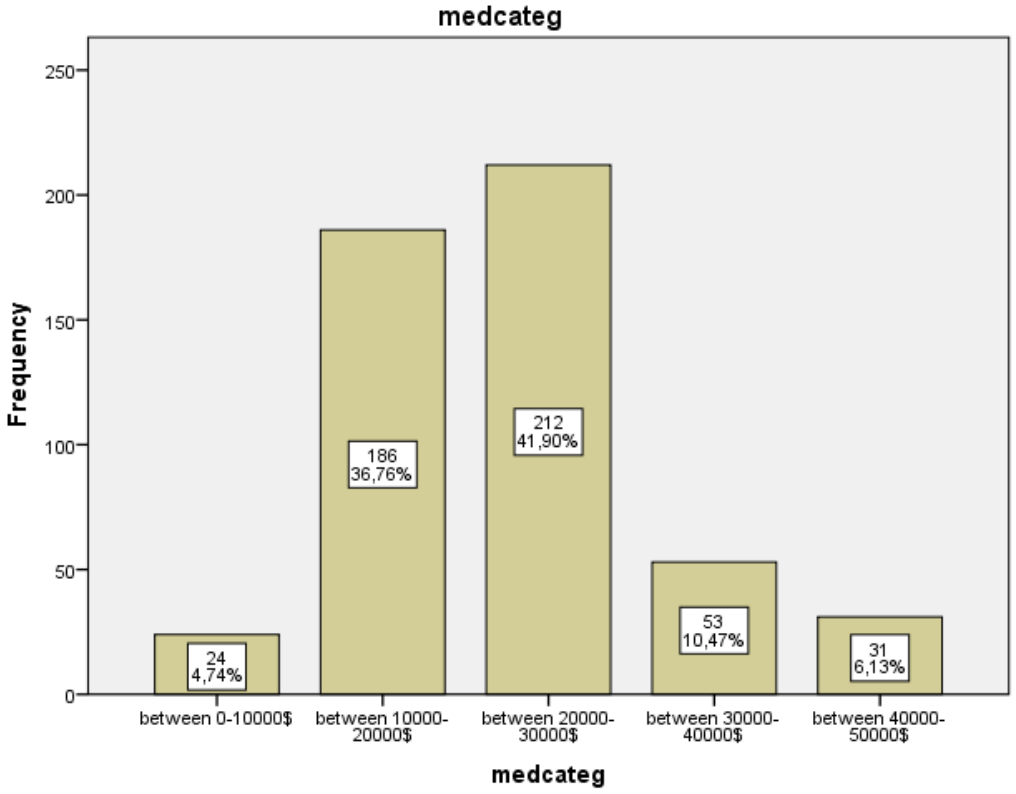

Figure 3: Distribution of Median Values for Boston Housing

Crosstab according to tax classes of price ranges of houses are given in Table 2. There are 69 observations with a tax of $250 or less, 300 observations with taxes ranging from $250 to $500 and 137 observations with a tax of $500 or more.

59% of the observations are in the tax class between $250 and $500. From Figure 5, it is seen that the number of houses with a tax of between $250 and $500 is the maximum and the number of houses with a tax of $250 or less is the least.

Table 2: Crosstab of Median Values for Taxes and Prices

		median					
		Between 0-10000$	Between 10000-20000$	Between 20000-30000$	Between 30000-40000$	Between 40000-50000$	Total
Tax	250 ve altı	0	12	33	17	7	69
	250 ve 500 arası	0	94	151	36	19	300
	500 ve üzeri	24	80	28	0	5	137
Total		24	186	212	53	31	506

Distribution of houses according to their ages are shown in Figure 4. It is seen that the distribution of ages of houses is skewed to the left and houses between the age of 80 and 100 are intense.

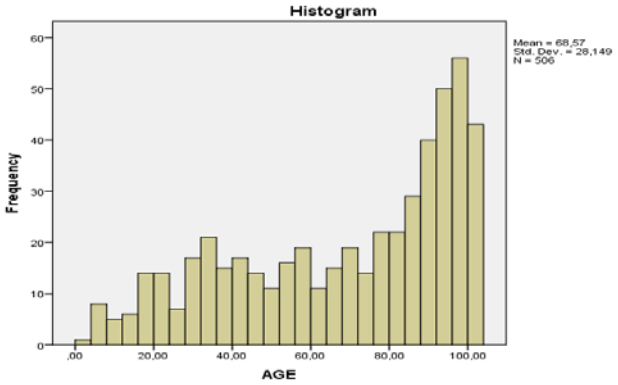

Figure 4: The Distribution of Houses According to Their Ages

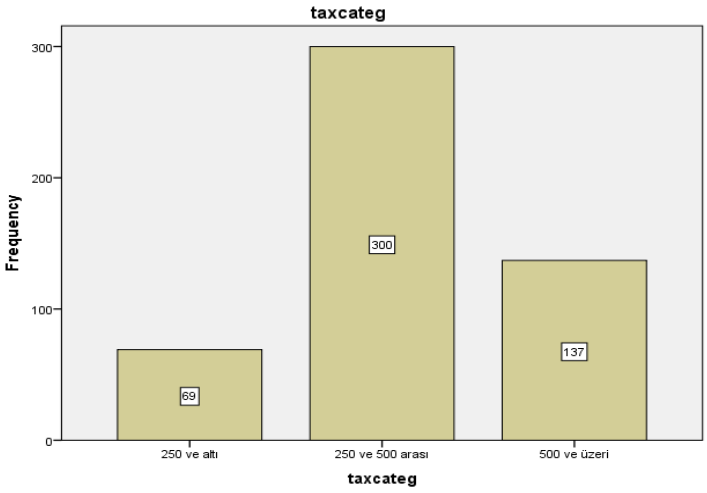

Figure 5: Distribution of Tax Classes

There are 35 houses riverside and there are 741 houses that are not riverside according to the distribution of houses riverside or not. Also the distribution of distances of houses to highways according to median values is examined. According to it, most of the houses, whose median values are between $40,000- $50,000 and between $30,000- $40,000, are near the 5th highway. Houses whose median values are between $20,000- $30,000 are closer to 4th and 5th highways than the other highways. Houses, whose median values are between $10,000- $20,000 and between $0- $10,000, are closer to the 24th highway. When houses riverside or not are examined according to tax classes, it is determined that houses, that are not riverside and have the tax class of between $250- $500, have more frequency. The phase after the understanding of the data phase is "*data preparation phase*". In this phase, there are no missing values in Boston housing data. In the data set, it is focused on median values of houses. The reason for this is that other 13 features have more or less effect on median values of houses. Therefore, in this study, median values are chosen as objective variable and are "*modeling*" according to median variables. In "*modeling phase*", firstly, taxes with a wide range, ages of houses and median values of prices of houses are categorized. As an objective variable, median values of prices of houses are taken and used to classify. While classifying, decision tree algorithms such as CART, C5 and Neural Network model are applied.

4.1. CART Algorithm

Root node for the phase of modeling in this model is stillness rates. Stillness rate is classified according to the number of rooms in the areas that stillness rate is less than or equal to than 14.4% and according to ages when stillness rate is higher than 14.4%. All other features are included in the problem as input variables. According to CART model, it is seen that median values of prices of houses, whose stillness rate is higher than 14.4%, the crime rate is lower than or equal to 11.369%, and ages are between 0 - 60, varies between $20.000 - $30.000.

4.2. C5 Algorithm

Root node for the phase of modeling in this model is the average number of rooms. Houses are classified according to stillness rates when the average number of rooms is 7 or lower and according to crime rates when the average number of rooms is higher than 7. All other variables are included as input values just like in CART algorithm. According to C5 model, median values of prices of houses, whose number of rooms is 7 or lower, stillness rate is 14.370% or lower, and distance to the business centers is 1.358 or lower, varies between $40.000 - $50.000.

4.3. Neural Networks Model

In this model, it is seen that feature with the most explanatoriness is stillness rate and with the least explanatoriness is crime rate.

"*Model evaluation phase*", which is the last phase of the study, is obtained through cross-validation rates. Since the data set we are examining is a small data set, when this phase is started, data set is divided into two parts, 50% of data set is as a test and remaining 50% is as a model set, and Cross-Validation method is applied. Validation rates for three models are obtained like in Table 3, 4 and 5.

Table 3: Tables for CART Algorithm Cross-Validation

Comparing median values		
Correct	182,5	72,11%
Wrong	70,5	27,89%
Total	253	

Table 4: Tables for C5 Algorithm Cross-Validation

Comparing median values		
Correct	184	72,74%
Wrong	69	27,26%
Total	253	

Table 5: Tables for Neural Network Cross-Validation

Comparing median values		
Correct	161	63,65%
Wrong	92	36,35%
Total	253	

When we examined validation rates in Table 3, 4 and 5, the C5 algorithm is chosen as the main model since it has the highest validation rate.

6. Conclusion

In this study, the aim is to investigate the underlying reasons for the change in the prices of housing. Relations between variables are examined by applying CRISP-DM process using these data. 6 stages of CRISP-DM, respectively “business understanding phase”, “data understanding phase”, “data preparation phase”, “modeling phase”, “model evaluating phase” and “deployment phase”, are completed for this data set. While classifying in modeling phase in the first three stages after the data are prepared, decision tree algorithms such as CART, C5 and Neural Network model are applied. While in CART model, classification is made based on stillness rates, in C5 model it is made based on the average number of rooms in houses. In Neural Network model, according to variables in interest for house prices, it is seen that feature with the highest explanatoriness is stillness rate and the feature with the lowest explanatoriness is crime rate. In the last phase of the study, model evaluating phase, cross-validation rates for CART, C5 and Neural Network models are obtained and C5 model is chosen as the

model with the highest validation rate of 72.74% valid classification.

7. Suggestion

In the future works in continuation of this study can be used different forecasting methods.

References

[1] Breiman, L. Friedman, J.H., Olshen, R.A., Stone, C.J. Classification and Regression Trees. Chapman & Hall, New York, 1984.

[2] Chapman P., Clinton J., Kerber R., Khabaz T., Reinartz T., Shearer C. and Wirth R. 2000. Step-by step Data Mining Guide, SPSS Inc..

[3] Cortez, P. 2006. Data mining with Neural Networks and Support Vector Machines using the R/rminer tool, FCT grant PTDC/EIA/64541/2006, http://www.r-project.org/.

[4] Fayyad, U.M. 1996. Advances in Knowledge Discovery and Data Mining, Menlo Park, CA: AAAI&MIT Press, USA.

[5] Fayyad, U., Shapiro, G.P. and Smyth, P. 1996. From Data Mining to Knowledge Discovery in Databases, Al Magazine, Volume 17, Number 3, pp. 37-54, AAAI.

[6] Friedman, J.H. 1997. Data Mining and Statistics: What is the Connection?, Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics, Houston, Texas, May 14-17, University of Huston.

[7] Gatnar E., Rozmus D. Data Mining-The Polish Experience. In Innovations in Classification, Data Science, and Information Systems, pp. 217-223, Springer Berlin Heidelberg, 2005.

[8] Gaur, P. 2012. Neural networks in data mining, International Journal of Electronic and Computer Science Engineering (IJECSE, ISSN-2277-1956), Vol.1, pp.1449-1453.

[9] Harrison, D. and Rubinfeld, D.L. 1978. 'Hedonic prices and the demand for clean air', J. Environ. Economics & Management, Vol. 5, 81-102.

[10] John GH., Enhancements to the data mining process (Doctoral dissertation, stanford university), 1997.

[11] Kantardzic, M. 2011. Data Mining Concepts, Models, Methods and Algorithms, Second Edition, IEEE Press, A John Wiley&Sons, Inc.

[12] Kasih, J., Ayub, M and Susanto, S. 2013. Predicting students' final passing results using the Classification and Regression Trees (CART) algorithm, World Transaction on Engineering and Technology Education, Vol.

11, No.1. pp.46-49.

[13] Kovalerchuk, B., Vityaev, E., 2000. Data Mining in Finance: Advances in Relational and Hybrid Methods. Springer Science & Business Media, 2000 edition, ISBN-10: 0792378040, ISBN-13: 978-0792378044

[14] Marmelstein, R.E., Hammack, L.P. and Lamont, G.B. 1999. "Concurrent approach for evolving compact decision rule sets", Proc. SPIE 3695, Data Mining and Knowledge Discovery: Theory, Tools, and Technology, 52 (February 25, 1999); doi:10.1117/12.339990; http://dx.doi.org/10.1117/12.339990

[15] Quinlan, J.R. C4.5: Programs for Machine Learning. Morgan Kaufman. 1993.

[16] Parsaye, K. 1997. OLAP and Data Mining: Bridging the Gap. Database Programming and Design, 10, pp.30-37.

[17] Patil, N., Lathi, R. and Chitre V. 2012. Comparison of C5 & CART Classification algorithms using pruning technique, International Journal of Engineering Research&Technology (IJERT), ISSN:2278-0181, Vol.1, Issue 4.

[18] Pnandya, R. and Pandya, J. 2015. C5 Algorithm to improved decision tree with feature selection and reduced error pruning, International Journal of Computer Applications, Vol. 117, No. 16.

[19] Ramakrishnan, R. and Gehrke, J. 2002. Database Management Systems, 3rd Edition, McGraw-Hill Professional.

[20] Ripundeep Singh Gill and Ashima, 2014. Neural networks in data mining, IOSR Journal of Engineering (IOSRJEN), Vol. 4, Issue 3, pp.1-6.

[21] Romero, C., Ventura, S., Espejo, P.G. and Hervas, C. 2008. Data Mining Algorithms to Classify Students, The 1st International Conference on Educational Data Mining, Montreal, Quebec, Canada, pp. 8-18. (June 20-21).

[22] Shamim, A., Shaikh, M.U. and Malik, S.R. 2010. Intelligent Data Mining in Autonomous Heterogeneous Distributed Bio Databases, Second International Conference on Computer Engineering and Applications.

[23] Singh Y. and Chauhan, A.S. 2009. Neural networks in data mining, Journal of Theoretical and Applied Information Technology, Vol.5, No.1, pp.37-42.

[24] Wirth, R. and Hipp, J. 2000. CRISP-DM: Towards a Standard Process Model for Data Mining, Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, pp.29-39, Manchester, UK.

[25] Zekulin AD, Busche FD, U.S. Patent No.6,430,547. Washington, DC: U.S. Patent and Trademark Office. 2002.