Pertanika J. Sci. & Technol. 29 (4): 2341- 2354 (2021)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Classification of Existing Health Model of India at the End of the Twelfth Plan using Enhanced Decision Tree Algorithm

Ashok Kumar*, Arun Lal Srivastav, Ishwar Dutt and Karan Bajaj

Chitkara University School of Engineering and Technology, Chitkara University, Himachal Pradesh, 174103 India

ABSTRACT

The high rate of urbanisation has increased the need for state-of-art health models that can meet the growing needs of society during any pandemic. Information-theoretic algorithms based on decision tree can mine the data to establish standards for the final decision by classifying the related data. Classification is an effective tool to analyse the existing health system in India's states and union territories. For this purpose, the data is categorised and then treated with the enhanced Shannon Entropy-based C4.5 decision tree algorithm to set some rules. These rules are capable of finding the major gaps in the health care systems after the analysis. Supposedly, these gaps are taken care of properly in the affected regions. In that case, the health care models will accomplish the endeavouring Sustainable Development Goals.

Keywords: C4.5 Algorithm, classification algorithms, decision tree, health model, Shannon entropy

ARTICLE INFO

Article history:
Received: 23 January 2021
Accepted: 24 May 2021
Published: 28 October 2021

DOI: https://doi.org/10.47836/pjst.29.4.06

E-mail addresses:
ashok.kumar@chitkarauniversity.edu.in (Ashok Kumar)
arun.srivastav@chitkarauniversity.edu.in (Arun Lal Srivastav)
ishwar.dutt1@chitkarauniversity.edu.in (Ishwar Dutt)
karan.bajaj@chitkarauniversity.edu.in (Karan Bajaj)
*Corresponding author

INTRODUCTION

There has been growth in unregulated private health service providers due to the high rate of urbanisation. Hence, it has become cumbersome for existing health models to meet high expectations and make a better choice from many service providers. Therefore, these primary health care services, which are easily accessible and the first point of contact, need to be improved to provide adequate care and achieve the expected health outcomes (Kruk et al., 2015; Mackintosh et al., 2016; Zeng et al., 2015).

ISSN: 0128-7680
e-ISSN: 2231-8526

Potential primary health care models can address the outbreak of the disease on time, before they become epidemic, through established and equipped delivery processes. Through such well-managed primary health care systems, there is a possible reduction in bad health in the worst-affected regions. Moreover, these systems can also support to get better health results (Kruk et al., 2010; Shi, 2012).

The importance of improving health systems is that the representatives of various regions recognised the United Nations' Millennium Development Goals during the Millennium Summit in September 2000. Their target is to monitor the development in the health sector and its outcome from 1990 to 2015 (Assembly, 2000).

Much remarkable work has been done to find the various issues in health care models that need to be taken care of. For example, some discussed the gaps within health care processes and their outcome by analysing the existing facilities and through several health care visits and found the low-quality standard in other domains like user experience, evidence-based care and population health management (Macarayan et al., 2018). Virus outbreaks in healthcare facilities for the elderly in Japan were analysed during the winter season of 2004 to 2005 (Okada et al., 2006). Regardless of many policies, women are not getting adequate quality maternity care; efforts are required to improve these care models (Alkema et al., 2016; Afulani et al., 2019).

Considerable efforts are required to address the issues and find the solution of developing an effective health care model. As an effective tool, data mining can deal with such issues by mining available health-related complex data sets. In data mining, various problems have been solved to extract potential knowledge from unorganised data, using various measures, like information-theoretic measures (Chen et al., 1996), local generalised quadratic distance metrics (Karim & Frank, 2017) and means like clustering of imbalanced high-dimensional media data (Sarka et al., 2018; Antonella & Mariangela, 2017; Panagiotis & Christos, 2016). Classification, which is one of the main objectives of data mining, is an effective tool to analyse the training set and study a classified model (Gondek & Hofmann, 2007; Zhang et al., 2006). Maria and Gunter (2016) found the decision tree algorithm as one of the most effective and commonly used key algorithms to build a predictive model among various classification algorithms. Initially, it is applied on a training set so that some classification criteria can be set and the unknown classes of the data are classified. The algorithm identifies, the appropriate property to every node by the gained information (Zhu & Wen, 2010). Refer Rokach and Maimon (2014) and Tzirakis and Tjortjis (2017) for more details about their advancements.

There are many studies on health care services to determine their effectiveness (Jamaludin et al., 2020; Jonsson et al., 2020). However, identifying the regions with a high density population, where the essential health services are more in demand, is another critical parameter that also requires immediate attention for the success of health care models. Furthermore, it will help in finding the sensitive areas during any pandemic. This purpose can be achieved by classifying the related data of the particular regions.

The present study has been motivated by the effectiveness of classification. So the data under consideration have been classified to analyse various parameters of health care models by bifurcating them in various ranges. The enhanced information theory-based classification algorithm C4.5 of data mining has been applied (Wu et al., 2008) to set some rules that are useful for reviewing and updating health care projects or models. The next part of the article includes the introductory part of the C4.5 algorithm compared to other classification methods. At the same time, in a later section, we categorise the attributes of data under consideration. Using the C4.5 algorithm, we have developed decision tree-based constraints. Finally, based on a set of rules obtained, a comparison of the resultant trees have been discussed, and the results with future scope have been concluded at last.

MATERIAL AND METHODS

C4.5 Algorithm

C4.5 algorithm plays a vital role in the development of the decision trees. Among the most commonly used decision tree algorithms, the ID3 algorithm (Quinlan, 1986) is mainly preferred for classification, while the C4.5 algorithm (Salzberg, 1994) is the modification or enhancement of the ID3 algorithm. C4.5 algorithm gets more popular when it is listed as the top 10 algorithms mentioned by Wu et al. (2008). The working procedure of this algorithm is the same as that of the ID3 algorithm. However, the only difference in the attribute selection criteria for a node is by using the split information.

Gained Information is evaluated using Shannon entropy (Shannon, 1948) and then gained ratio is calculated for each attribute. Finally, the attribute with the maximum gained ratio is used for classification. Then, the process is repeated further to develop its subtrees.

The C4.5 algorithm is preferred over other algorithms of classification due to the advantages mentioned in Table 1 (Sharma & Kumar, 2016):

Table 1

Perfomance of Different Decision Tree Algorithms

Features	**Varma Entropy (Varma, 1966)**	**ID3 (Quinlan, 1986)**	**C4.5 (Quinlan, 1994)**
Data Types	Discrete	Discrete	Discrete and Continuous
Speed	Average	Low	Fast
Pruning	Post Pruning	Post Pruning	Pre Pruning
Attribute Selection Criteria	Gained Information	Gained Information	Split info
Missing Values	Affects	Affects	No effect

Application in Health Model of India at the End of the Twelfth Plan

The data of India's different states and union territories that support the health system has been collected at the end of the Twelfth Plan (OGD, 2015) and is summarised in Table 3. This data is mainly categorised into two classes, C_1 and C_2, where C_1 is classified as 'g' indicating good health care services. In contrast, C_2 is categorised as 'ng' that indicates 'not good' health care services. Based on the importance in health care models, four attributes (*the number of Primary Health Centers, Sub Centers and Community Health Centers functioning at the end of Twelfth Plan* and *Rural Population covered by Primary Health Centers)* from the data source have been selected. These attributes are essential in the process of decision making while selecting the appropriate region for health-related projects to be implemented based on the classification of the health model.

Associated average of health care for a particular region is considered 'good' if it is more than the national average. Otherwise, it is considered 'not good'. Finally, the three attribute classes are combined with the fourth having discrete values without any intervals and further are concluded as 'g' or 'ng' based on the majority as summarised in Table 2. Here A_1, A_2, A_3 *and* A_4 represent the four attributes, while case numbers (Case 1, Case 2 and so on) represents the various possible cases during the process of a class assignment to a particular region.

Table 2
Class Assignment Criteria

Regions	A_1	A_2	A_3	A_4	Classes of Health Care Services
Case 1	g	g	g	g	g
Case 2	g	g	g	ng	g
Case 3	g	g	ng	ng	ng
Case 4	g	ng	ng	ng	ng
Case 5	ng	ng	ng	ng	ng

Using the Shannon entropy-based C4.5 algorithm (Shannon, 1948), we developed a decision tree and compared it with Kumar et al. (2016), in which modified ID3 algorithm based on the Varma entropy measure (Varma, 1966) was used for specific values of two parameters such as α and β the same data with different specifications attributes. In the data used by Kumar et al. (2016), all the attributes have intervals, while in the present study, we deal with discrete values without intervals for one particular attribute. During the construction of the decision tree, the attributes data with discrete values have been divided at the threshold into two groups. Hence, only two branches emerge from it. In contrast,

in other attributes, branches are equal to the number of intervals and consequently result in more rules.

The first three attributes (*Primary Health Centers, Sub Centers* and *Community Health Centers*), based upon their values either less than or greater than the national average per unit population, are divided into four different ranges. Following are the details of all four attributes.

PHC stands for average population covered by *Primary Health Centers* functioning at the end of the Twelfth Plan, with four ranges: less than 26000, 26000-47000, 47001-61000 and greater than 61000.

SBC stands for average population covered by *Sub Centers* functioning at the end of the Twelfth Plan, with four ranges: less than 3600, 3600-4500, 4501-7100 and greater than 7100;

CHC is representing the average population covered by *Community Health Centers* functioning at the end of the Twelfth Plan with four ranges: less than 140000, 140000-220000, 220001-320000 and greater than 320000; and

RPHC is represents the average *Rural Population covered by Primary Health Centers* with discrete data without any interval and varies from 0 to 83808.

Table 3

Health Model of INDIA

Sr. no.	PHC	SBC	CHC	RPHC	Classes of Health Care Services
1	<26000	<3600	<140000	1077	ng
2	47001-61000	4501-7100	220001-320000	32979	g
3	<26000	4501-7100	<140000	9114	ng
4	26000-47000	4501-7100	140000-220000	26437	ng
5	47001-61000	>7100	>320000	4900	g
6	<26000	>7100	>320000	0	ng
7	26000-47000	4501-7100	140000-220000	25042	ng
8	47001-61000	4501-7100	>320000	26159	g
9	>61000	>7100	<140000	20132	ng
10	>61000	>7100	<140000	83808	g
11	>61000	4501-7100	>320000	26273	g
12	47001-61000	>7100	140000-220000	29961	ng
13	47001-61000	>7100	220001-320000	36364	g

Table 3 *(Continued)*

Sr. no.	PHC	SBC	CHC	RPHC	Classes of Health Care Services
14	<26000	<3600	<140000	12630	ng
15	<26000	4501-7100	140000-220000	14298	ng
16	>61000	>7100	140000-220000	75924	g
17	26000-47000	4501-7100	>320000	16780	ng
18	26000-47000	>7100	140000-220000	21075	ng
19	<26000	4501-7100	<140000	3535	ng
20	>61000	>7100	140000-220000	45426	g
21	>61000	>7100	220001-320000	33990	g
22	26000-47000	4501-7100	140000-220000	23784	ng
23	26000-47000	4501-7100	<140000	21958	ng
24	<26000	<3600	<140000	9218	ng
25	<26000	4501-7100	<140000	11171	ng
26	26000-47000	4501-7100	<140000	26797	ng
27	47001-61000	>7100	>320000	16467	g
28	>61000	>7100	140000-220000	40619	g
29	26000-47000	4501-7100	<140000	24736	ng
30	<26000	3600-4500	220001-320000	19042	ng
31	47001-61000	>7100	140000-220000	27195	ng
32	26000-47000	3600-4500	140000-220000	32291	ng
33	47001-61000	>7100	220001-320000	44414	g
34	26000-47000	4501-7100	140000-220000	27381	ng
35	>61000	>7100	220001-320000	68408	g

Now, by using Shannon entropy (Shannon, 1948), the amount of the information required for the object data is measured, using the following Equation 1:

$$I(C_1, C_2, C_3, \cdots, C_c) = -\sum_{i=1}^{c} p_i \log_2 p_i, \quad (1)$$

with p_i represents the probability associated with each class.

In Table 3, the complete data is divided into two classes $\boldsymbol{C_1}$ and $\boldsymbol{C_2}$ where 13, 'g' is in $\boldsymbol{C_1}$ and 22, 'ng' is in $\boldsymbol{C_2}$. Therefore

$$I(C_1, C_2) = \left(-\frac{13}{35}\right) * \log\left(\frac{13}{35}\right) - \left(\frac{22}{35}\right) * \log\left(\frac{22}{35}\right) = 0.951762676$$

Further, for the number of values, m_j in the range R_j of attribute A_k then entropy of A_k, E(A_k) is given as Equation 2:

$$E(A_k) = \sum_{j=1}^{r} \frac{m_j}{N} \sum_{i=1}^{c} p_i \log_2 p_i \quad (2)$$

Here, association with the four different attributes are given as follows:

$$E(A_1) = E(PHC) = 0.309678301$$

$$E(A_2) = E(SBC) = 0.702953112$$

$$E(A_3) = E(CHC) = 0.685134682$$

$$E(A_4) = E(RPHC) = 0.378114904$$

Net gained information $Gain\_info(A_k)$ for different attributes are given by Equation 3

$$Gain\_info(A_k) = I(C_1, C_2, C_3, \cdots, C_k) - E(A_k) \quad (3)$$

Hence, we have

$$Gain\_info(A_1) = 0.642084374 \ bits$$

$$Gain\_info(A_2) = 0.248809563 \ bits$$

$$Gain\_info(A_3) = 0.266627993 \ bits$$

$$Gain\_info(A_4) = 0.573647772 \ bits$$

Further, information of split, $Split\_info(A_k)$ has been measured for each attribute using the following Equation 4

$$Split\_info(A_k) = \sum_{j=1}^{r} \frac{m_j}{N} \log_2 \left\{ \frac{m_j}{N} \right\} \quad (4)$$

Therefore, we get

$$Split\_info(A_1) = 1.993608561.$$

$$Split\_info(A_2) = 1.587522864.$$

$$Split\_info(A_3) = 1.926630222.$$

$$Split\_info(A_4) = 0.863120569.$$

Net gained ratio information i.e. $Gain\_ratio(A_k)$ for each attribute 'A_k' is calculated using the Equation 5

$$Gain\_ratio(A_k) = \frac{Gain\_info(A_k)}{Split\_info(A_k)} \quad (5)$$

Thus, we have

$$Gain\_ratio(A_1) = 0.32207143 \ \ bits$$

$$Gain\_ratio(A_2) = 0.15672817 \ \ bits$$

$$Gain\_ratio(A_3) = 0.13839085 \ \ bits$$

$$Gain\_ratio(A_4) = 0.66462067 \ \ bits$$

We note that the ratio of the fourth attribute is obtained with discrete values and without intervals. For example, RPHC is the maximum. Therefore this attribute is designated as the root of the decision tree and will divide the data initially into two parts at the threshold. Hence, only two branches will emerge for root. Further, the above process of gained ratio is repeated for branches or subtrees until we get the terminal nodes as 'g' or 'ng'. The resultant decision tree for the data in Table 3, using the C4.5 Algorithm, is given in Figure 1.

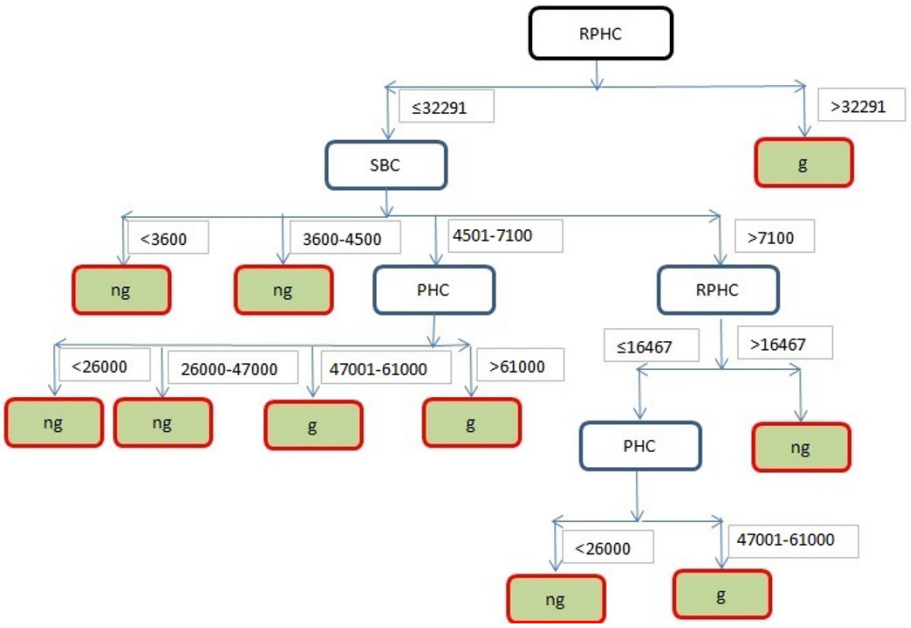

Figure 1. C4.5 Algorithm based Decision Tree

The decision tree obtained by Kumar et al. (2016) based on Varma (1966) is given in Figure 2, and for rules induced, we can refer to Kumar et al. (2016). Information required, entropy values, split information and gained information ratios are calculated for each attribute of the training data set. The root of the decision tree has the attribute with the highest gained ratio, and the other remaining attributes are arranged as in the nodes of its branches. The same procedure is repeated for the nodes of the branches. The rules formed while moving from the root towards the leaf or terminal node of the final decision tree are helpful in setting classification criteria. These classification rules form the strong base for the systematic analysis and the development of the system.

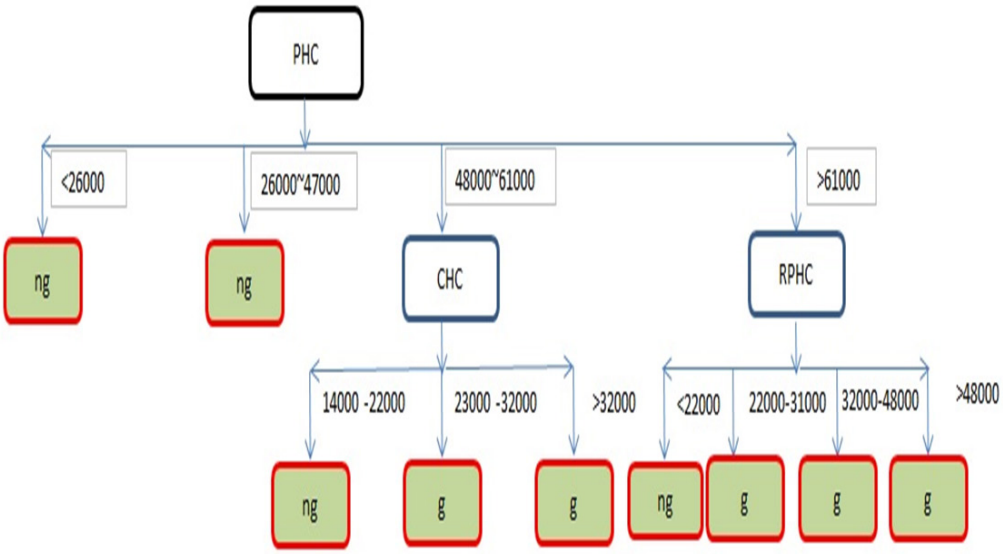

Figure 2. Varma Entropy-based Decision Tree

RESULTS AND DISCUSSION

Rules Induced

Some rules have been induced using 'if-then', based on the resultant C4.5 algorithm-based decision tree, given in Figure 1. Thus, further improving the decision making criteria during the evaluation or implementation of the health-related projects for the development of the health care system. These rules can be summarised as follows:

1. If RPHC is less than or equal to 32291 and SBC is less than 4500, then for any value of PHC and CHC, health care services are 'not good' in that region.

2. If RPHC is less than or equal to 32291 and SBC is within 4501 to 7100 and PHC is less than 47000, then for any value of CHC, health services are 'not good' in that region.
3. If RPHC is less than or equal to 32291and SBC is within 4501 to 7100 and PHC is greater than 47000, then for any value of CHC, health services are 'good' in that region.
4. If RPHC is less than or equal to 32291 and SBC is greater than 7100, again RPHC is less than or equal to 16467 and PHC is less than 26000, then for any value of CHC, health services are 'not good' in that region.
5. If RPHC is less than or equal to 32291 and SBC is greater than 7100, again RPHC is less than or equal to 16467 and PHC is within 47001 to 61000, then for any value of CHC, health services are 'good' in that region.
6. If RPHC is less than or equal to 32291 and SBC is greater than 7100 and RPHC is more than 16467, then for any value of CHC, health services are 'not good' in that region.
7. If RPHC is greater than 32291 then for any value of SBC, PHC and CHC, health services are 'good' in that region.

The rules have been framed while moving from the root towards the terminating nodes in the resultant decision tree. Also, rules 4 and 5 can be further simplified by combining the conditions for RPHC as less than or equal to 16467. The regions with good health care models satisfy rules 3, 5 and 7, while rules 1, 2, 4 and 6 help identify the regions where health care models need to be updated.

Comparison between the Decision Trees

Two decision trees obtained by the modified ID3 algorithm based on Varma Entropy and C4.5 Algorithm have been compared and summarised in Table 4. Some conditions or rules have been formed with the help of decision trees so that the correct information can be collected. Thus, proper decisions can be made before implementing required policies or projects for the improvement of health care services. Furthermore, the rules obtained from the differently trained data (by considering the values of one attribute without intervals) and then applying the C4.5 Algorithm have resulted in more precise rules than those obtained by Kumar et al. (2016).

Table 4

Comparison of the Decision Trees

Sr. no.	Factors	Varma Entropy based Decision Tree	C4.5 Algorithm based Decision Tree
1	Attribute at root	PHC	RPHC
2	Minimum branch length	1	1
3	Maximum branch length	2	4
4	Smallest Rule	If 'PHC' less than 26000 then 'ng'	If 'RPHC' is greater than 32291 then 'g'
5	Largest Rule	If 'PHC' is greater than 61000 and 'RPHC' is greater than 48000 then 'g'	If 'RPHC' is less than or equal to 32291, 'SBC' is greater than 7100, 'RPHC' is greater than equal to 16467 and 'PHC' is less than 26000 then 'ng'.
6	Rules at depth 1	2	1
7	Rules at depth 2	7	2
8	Rules at depth 3	NA	5
9	Rules at depth 4	NA	2
10	Total number of Rules	5	8

CONCLUSION

The proper decision-making process is always helpful in the successful completion of time-bound projects. The selection of the appropriate classification method is based upon the nature of the training data set. The rules were based on the decision tree obtained by categorised data using the C4.5 algorithm is compared with that of the improved ID3 algorithm, based upon Varma entropy for a particular value of parameters. It was observed that, on using discrete values without intervals, better results are obtained. Thus, in comparison, if the C4.5 algorithm is used for classification, after categorisation of the collected data into different classes, more refined rules have been developed. Such rules help analyse the systematic and progressive development of any introduced system. It has been observed from the rules induced that available processes and measures need improvement in its current models to capture key elements of health facilities. Moreover, the unique features of regions where the health care models are in good condition can be implemented in those regions where improvement is required.

ACKNOWLEDGEMENT

Authors appreciate the incomparable support and inspiration by Honourble Chancellor Dr Ashok K. Chitkara and Honourble Pro-Chancellor Dr Madhu Chitkara, Chitkara University, Himachal Pradesh, India, to write this manuscript.

REFERENCES

Afulani, P. A., Phillips, B., Aborigo, R. A., & Moyer, C. A. (2019). Person-centred maternity care in low-income and middle-income countries: Analysis of data from Kenya, Ghana, and India. *The Lancet Global Health*, *7*(1), e96-e109. https://doi.org/10.1016/S2214-109X(18)30403-0

Alkema, L., Chou, D., Hogan, D., Zhang, S., Moller, A. B., Gemmill, A., Fat, D. M., Boerma, T., Temmerman, M., Mathers, C., & Say, L. (2016). Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenario-based projections to 2030: A systematic analysis by the UN Maternal Mortality Estimation Inter-Agency Group. *The Lancet*, *387*(10017), 462-474. https:// 10.1016/S0140-6736(15)00838-7

Antonella, P., & Mariangela, S. (2017). Weighted distance-based trees for ranking data. *Advances in Data Analysis and Classification, 13*(2), 427-444. https://doi.org/10.1007/s11634-017-0306-x.

Assembly, U. G. (2000, September 6-8). United Nations millennium declaration. In *Millenium Summit of the United Nations*. New York.

Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, *8*(6), 866-883. https:// 10.1109/69.553155

Gondek, D., & Hofmann, T. (2007). Non-redundant data clustering. *Knowledge and Information Systems*, *12*(1), 1-24. https://doi.org/10.1007/s10115-006-0009-7

Jamaludin, M. H., Wah, Y. B., Nawawi, H. M., Yung-An, C., Rosli, M. M., & Annamalai, M. (2020). Classification of familial hypercholesterolaemia using ordinal logistic regression. *Pertanika Journal of Science & Technology*, *28*(4), 1163-1177. https://doi.org/10.47836/pjst.28.4.03

Jonsson, Å., Orwelius, L., Dahlstrom, U., & Kristenson, M. (2020). Evaluation of the usefulness of EQ-5D as a patient-reported outcome measure using the Paretian classification of health change among patients with chronic heart failure. *Journal of Patient-Reported Outcomes*, *4*(1), 1-11. https://doi.org/10.1186/s41687-020-00216-7

Karim, A., & Frank, P. F. (2017). Local generalized quadratic distance metrics:Application to the k-nearest neighbors. *Advances in Data Analysis and Classification, 12*(2), 341-363. https://10.1007/s11634-017-0286-x.

Kruk, M. E., Nigenda, G., & Knaul, F. M. (2015). Redesigning primary care to tackle the global epidemic of noncommunicable disease. *American Journal of Public Health*, *105*(3), 431-437. https://10.2105/AJPH.2014.302392

Kruk, M. E., Porignon, D., Rockers, P. C., & Van Lerberghe, W. (2010). The contribution of primary care to health and health systems in low-and middle-income countries: A critical review of major primary care initiatives. *Social Science & Medicine*, *70*(6), 904-911. https://10.1016/j.socscimed.2009.11.025

Kumar, A., Taneja, H. C., & Chitkara A. K̈. (2016, January 18-19). Analysis of health conditions using generalized information measure based ID3 algorithm. In *4th Annual International Conference on Operations Research and Statistics (ORS-2016)* (pp. 33-37). Singapore. https://10.5176/2251-1938_ORS16.11

Macarayan, E. K., Gage, A. D., Doubova, S. V., Guanais, F., Lemango, E. T., Ndiaye, Y., Waiswa, P., & Kruk, M. E. (2018). Assessment of quality of primary care with facility surveys: A descriptive analysis in ten low-income and middle-income countries. *The Lancet Global Health*, *6*(11), e1176-e1185. https://doi.org/10.1016/S2214-109X(18)30440-6

Mackintosh, M., Channon, A., Karan, A., Selvaraj, S., Cavagnero, E., & Zhao, H. (2016). What is the private sector? Understanding private provision in the health systems of low-income and middle-income countries. *The Lancet*, *388*(10044), 596-605. https://doi.org/10.1016/S0140-6736(16)00342-1

Maria, T. G., & Gunter, R. (2016). Probabilistic clustering via Pareto solutions and significance tests. *Advance Data Analysis and Classification, 12*(2), 179-202. https://10.1007/s11634-016-0278-2.

OGD. (2015). *Open government data (OGD) platform India.* Retrieved June 6, 2015, from https://data.gov.in/.

Okada, M., Tanaka, T., Oseto, M., Takeda, N., & Shinozaki, K. (2006). Genetic analysis of noroviruses associated with fatalities in healthcare facilities. *Archives of Virology*, *151*(8), 1635-1641. https://doi.org/10.1007/s00705-006-0739-6

Panagiotis, T., & Christos, T. (2016). T3C: Improving a decision tree classification algorithm's interval splits on continuous attributes. *Advances in Data Analysis and Classification, 11*(2), 353-370. https://doi.org/10.1007/s11634-016-0246-x.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1) 81-106. https://doi.org/10.1007/BF00116251

Rokach, L., & Maimon, O. (2014). *Data mining with decision trees: Theory and applications*. World Scientific. https://doi.org/10.1142/9097

Salzberg, S. L. (1994). C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning, 16*, 235-240. https://doi.org/10.1007/BF00993309

Sarka, B., Maia, Z., Peter, F., Thomas, O., & Christian, B. (2018). Clustering of imbalanced high-dimensional media data. *Advances in Data Analysis and Classification, 12*(2), 261-284. https://doi.org/10.1007/s11634-017-0292-z.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379-423. https://10.1002/j.1538-7305.1948.tb01338.x

Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research*, *4*(4) 2094-2097.

Shi, L. (2012). The impact of primary care: A focused review. *Scientifica, 2012*, Article 432892. https://10.6064/2012/432892

Tzirakis, P., & Tjortjis, C. (2017). T3C: Improving a decision tree classification algorithm's interval splits on continuous attributes. *Advances in Data Analysis and Classification*, *11*(2), 353-370. https://doi.org/10.1007/s11634-016-0246-x

Varma, R. S. (1966). Generalizations of Renyi's entropy of order α. *Journal of Mathematical Sciences, 1*(7), 34-48.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., & Zhou, Z. H. (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, *14*(1), 1-37. https://doi.org/10.1007/s10115-007-0114-2

Zeng, J., Shi, L., Zou, X., Chen, W., & Ling, L. (2015). Rural-to-urban migrants' experiences with primary care under different types of medical institutions in Guangzhou, China. *PloS One*, *10*(10), Article e0140922. https://doi.org/10.1371/journal.pone.0140922

Zhang, J., Kang, D. K., Silvescu, A., & Honavar, V. (2006). Learning accurate and concise naïve Bayes classifiers from attribute value taxonomies and data. *Knowledge and Information Systems*, *9*(2), 157-179. https://doi.org/10.1007/s10115-005-0211-z

Zhu, P., & Wen, Q. (2010). Some improved results on communication between information systems. *Information Sciences*, *180*(18), 3521-3531. https://doi.org/10.1016/j.ins.2010.05.028