Decision Trees as Readable Models for Early Childhood Caries

Vladimir Ivančević, Nemanja Igić, Branko Terzić, Marko Knežević and Ivan Luković

Abstract Assessing risk for early childhood caries (ECC) is a relevant task in public health care and an important activity in fulfilling this task is increasing the knowledge about ECC. Discovering important information from data and sharing it in an understandable format with both experts and the general population could be beneficial for advancing and spreading the knowledge about this disease. After having experimented with association rule mining, we investigate the possibility of using decision trees as readable models in risk assessment. We build various decision trees using different algorithms and splitting criteria, favouring compact decision trees with good predictive performance. These decision trees are compared to the previous ECC models for the same analyzed population, namely a logistic regression model and an associative classifier, as well as to decision trees for caries from other studies. The results indicate flexibility and usefulness of decision trees in this context.

Keywords Early childhood caries · Risk assessment · Decision tree

V. Ivančević (✉) · N. Igić · B. Terzić · M. Knežević · I. Luković
Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
e-mail: dragoman@uns.ac.rs

N. Igić
e-mail: nemanjaigic@uns.ac.rs

B. Terzić
e-mail: branko.terzic@uns.ac.rs

M. Knežević
e-mail: marko.knezevic@uns.ac.rs

I. Luković
e-mail: ivan@uns.ac.rs

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies 2016*,
Smart Innovation, Systems and Technologies 57,
DOI 10.1007/978-3-319-39627-9_39

1 Introduction

The importance of understanding and treating early childhood caries (ECC) is being slowly acknowledged across the world. This disease represents a significant threat to the sensitive population it targets, namely children under the age of six, as well as to the whole society. Even in many developed countries, which make considerable investments in prevention and treatment, it is difficult to eradicate ECC, as there are various subgroups at high caries risk. This pattern has been observed across different continents and it is present in countries such as USA [1], Brazil [2], and Serbia [3].

Despite numerous medical advances, many open questions about ECC remain. One such issue, which has been recognized in the conclusion paper of the 2014 ECC Conference, is the assessment of caries risk [4]. As there is a large body of investigated risk factors for which there is no reliable confirmatory evidence, additional studies concerning such factors are needed in order to make progress in battling ECC. This problem is further aggravated by the diversity and interrelatedness of considered factors, which include microbiological, dietary, hygienic, and even social risks [4, 5].

In our previous studies on ECC in the South Bačka area (Autonomous Province of Vojvodina, Republic of Serbia), we employed data mining, namely association rule mining, for the following two purposes: (i) to identify potential risk factors for ECC by analyzing associations rules about ECC and candidate risk factors [6]; and (ii) to create an understandable predictive model for ECC by combining the association rules into a rule-based classifier [7]. Our primary motivation was to create readable models from real-world data that could be understood by general population or utilized to facilitate communication between data mining experts, who are looking for valuable patterns, and domain experts, who know how to interpret such information. Association rules appeared especially convenient, as they present information in a straightforward manner, allow for grouping into readable classifiers, have comprehensive character, and lack structural restrictions of decision trees.

In the present study, we are turning attention to decision trees, as they are also readable models that may be used both for explanation and prediction. According to [8], decision trees possess the following three advantages: (i) easy to represent and interpret; (ii) without conflicting knowledge; and (iii) identify and reveal important knowledge features. As they might be simpler to read and apply than typical rule-based classifiers, such as associative classifiers, we considered it worthwhile to investigate readable decision trees that could match the accuracy of previous models.

Our main goal is to construct a decision tree that could be of similar or better predictive performance when compared to the previous models for ECC in the South Bačka area: a logistic regression model [9] and an associative classifier [7]. To this end, we employed traditional tree generation by recursive partitioning and summarization of association rules into a decision tree. We evaluate the resulting trees with respect to their predictive power and structural properties, as well as compare them to other tree-based caries models from similar studies.

Besides Introduction and Conclusion, the paper is organized into four additional sections. In Sect. 2, we describe related results from previous studies on ECC modelling and discuss the usage of decision trees in ECC research. In Sect. 3, we elaborate on the employed data set and the applied methods. In Sect. 4, we present the resulting decision trees, as well as caries models from other studies. In Sect. 5, we review these decision trees and compare them to the models from other studies.

2 Related Work

In this section, we first comment on models from the related studies and then discuss the suitability of decision trees for ECC modelling.

2.1 Previous ECC Models

The initial model of ECC in South Bačka was a logistic regression model composed of five child-related variables (city, gender, birth order, birth weight and use of medical syrups), which was constructed by medical experts and based on the results of a statistical analysis of the ECC data [9]. It was followed by rule-based classifiers composed of association rules about ECC (associative classifiers) [7]. We demonstrated that 6-rule or 8-rule associative classifiers may outperform the initial logistic regression model. Moreover, these models incorporated many factors not directly related to the child, e.g., socioeconomic status of the family and factors concerning parents.

However, this improvement was coupled with a cost. Although understandable, associative classifiers generally do not provide a clear overview of all the included factors as important rules may combine numerous factors and a classifier may incorporate numerous diverse rules. This was controlled to some extent by clustering factors and favouring classifiers with dissimilar factors. Nonetheless, we decided to investigate decision trees as readable models with the hope that their structure could be more readable and their performance comparable to that of the previous models.

2.2 Decision Trees for Caries

Over the past 25 years, there have been at least six studies that utilized decision trees as caries models [10–15]. These studies mostly rely on classification and regression trees (CART) [16], the only exceptions being the use of C 5.0 [17] in [12] and CHAID [18] in [14]. An overview of those models is given in Sects. 4 and 5.

On the other hand, there are other types of decision tree formation. Decision trees that use previously mined association rules are often neglected in practice despite having some promising advantages. For instance, associative classification trees (ACTs) seem to be smaller and slightly more accurate than traditional decision trees [8], which should present sufficient motivation to evaluate them in our search for a compact and accurate model of ECC. Moreover, as we already possess a large set of association rules about ECC, we decided to form ACTs for ECC.

At present, there are various suggestions on how to build such decision trees [8, 19–22]. However, a practical problem with ACTs is that there are few implementations readily available in popular data analysis and mining tools. Given the detailed explanation of ACT creation and the positive evaluation results reported in [8], we chose to implement that particular ACT variant (see Sect. 3).

3 Materials and Methods

In this section, we describe an ECC data set and outline the creation of decision trees.

3.1 Data

The ECC data was recorded by Tušek [9] for a 10 % sample of children from the South Bačka area, which is part of the Autonomous Province of Vojvodina, Republic of Serbia. The data set features 341 records, one for each examined child, across a selection of 36 categorical variables. The output variable indicates ECC presence (30.5 %) or absence (69.5 %), while input variables correspond to potential risk factors such as socioeconomic status, dietary habits, health awareness, and behaviour of the child and the parents. More information about the data set may be found in [6].

3.2 Decision Tree Formation

The decision trees that we created could be divided into the following two groups: traditional decision trees and associative classification trees.

Formation of Traditional Decision Trees. We used RapidMiner Studio [23], a software tool for data mining, to create five traditional decision trees. Their formation is based on recursive partitioning, which, in each step, requires selection of a splitting attribute according to one of the predefined criteria. We used the RapidMiner implementation of the algorithm for decision tree generation (RDT) to create four decision trees, one for each of the four splitting criteria supported in the tool (accuracy, gain ratio, information gain, and the Gini index). We post-pruned each tree and applied the evolutionary computation approach to optimize the confidence parameter. The Gaussian mutation was used to maintain generic diversity across the population of five individuals, while the tournament selection approach with the tournament fraction set to 0.25 was used to perform selection in the population. The fifth decision tree was created using the CHAID implementation in RapidMiner, which utilizes a χ^2-based splitting criterion. In order to obtain more nodes in the tree, we set the minimum node size for split to 18 and the minimum leaf size to 9. For all five trees, the city variable was excluded, as it led to less general models. The maximum tree depth was set to seven, based on the average human capacity to process information (Miller's Law [24]).

Associative Classification Trees. The ACT formation process is described in detail in [8]. The root node contains a starting set of association rules, which is recursively split among the child nodes until one of the predefined termination criteria is fulfilled (homogenous rule set achieved or a specific threshold met). In addition to specifying the set of association rules and the splitting criterion needed to select the attribute for node splitting (confidence gain or entropy gain), ACT formation requires threshold values δ_{SUPPORT}, $\delta_{\text{CONFIDENCE}}$, and $\delta_{\text{TREE HEIGHT}}$, which determine boundary values for the node support, confidence, and height within the tree, respectively. The predictions given by ACTs generally depend on the rule sets in the leaves. We used the Java programming language to implement formation and application of ACTs.

Formation of Associative Classification Trees. The association rules for ACT formation were mined from the ECC data set according to the specification in [6], but the resulting rule set was not pruned. The complete rule set encompassed 18 rules about ECC presence and 88142 rules about ECC absence. We believe that there may be three potential reasons for the disproportionate shares of the two rule subsets: (i) in the examined population, ECC presence is less common than ECC absence (30.5 vs. 69.5 %); (ii) the size of the ECC data set (341 cases) might be too small to reliably detect less frequent patterns about ECC presence; and (iii) although there are numerous potential risk factors for ECC [25], it is possible that many of them have only a minor role in the examined population, which could manifest as a multitude of less frequent (and difficult to detect) patterns. We created three ACTs, one for each implemented splitting criterion: confidence gain, entropy gain, and lift gain. The first two criteria initially yielded poor predictive performance, so we added the lift gain criterion by modifying the confidence gain criterion to take into account the change in the lift of a group of rule sets instead of the confidence change. The threshold values for support, confidence, and height were set to 0.01, 0.8, and 7, respectively.

3.3 Decision Tree Evaluation

We aimed for decision trees of compact structure (readability) and good predictive performance. The readability was evaluated by node count (ND) and tree height (H). The predictive performance was evaluated on the whole data set by accuracy (ACC) and true skill statistic (TSS) [26], which is a measure somewhat similar to kappa [27]. Both TSS and kappa yield values between −1 and 1, with positive values denoting performance better than random guessing, but TSS is simpler to calculate and does not appear to be dependent on prevalence [26]. TSS was used so that we could directly compare the decision trees to the previous ECC models: the logistic regression model (TSS = 0.325) and the associative classifier (TSS = 0.41) [7]. For each tree group, we measured the overall predictive performance associated with each supported splitting criterion as the average accuracy in a stratified tenfold cross-validation (CV ACC). We used the dot program [28] to visualize the selected decision trees.

We also inspected decision trees for caries from other studies [10–15]. In this overview, we considered data set characteristics (population and predictor variables), tree formation procedure (TFP), variable count in the tree (V), tree height (H), and predictive performance as measured by sensitivity (SN) and specificity (SP).

4 Results

The eight resulting decision trees are presented in Table 1. An overview of various decision trees from the other studies and three representative decision trees from the present study, which are denoted by "~", is given in Table 2. With respect to both prediction and structure, we selected for discussion a single tree from each group: the ACT for splitting by lift gain (see Fig. 1) and the CHAID decision tree (see Fig. 2).

Table 1 The resulting decision trees

Tree group	Variable selection	ND	H	ACC (%)	TSS	CV ACC (%)
Traditional	Accuracy	19	5	76	0.23	69
Traditional	Gain ratio	19	6	74	0.15	68
Traditional	Information gain	55	6	85	0.59	66
Traditional	The Gini index	77	6	87	0.68	65
Traditional	CHAID	22	4	75	0.32	69
ACT	Confidence gain	6	2	70	0.00	69
ACT	Entropy gain	4	1	67	0.11	67
ACT	Lift gain	6	2	72	0.17	69

Table 2 Decision trees from caries studies

Study	Data set	TFP	V	H	SN (%)	SP (%)	Remark
[10]	N = 1024 (age 5–9) Portland, ME, USA 32 predictors	CART (Gini)	2	2	62	77	/
	N = 914 (age 5–8) Aiken, SC, USA 32 predictors		9	9	64	86	/
[11]	N = 466 (elementary school age) Rochester & Finger Lakes areas, NY, USA 5 predictors	CART (Gini)	3	3	/	55	/
[12]	N = 500 (age 5–8) Ena & Nakatsugawa areas, Gifu, Japan 12 predictors	C 5.0	5	4	73	77	/
[13]	N = 442 (age 20–64) Osaka, Japan 5 predictors	CART	2	2	34	85	Primary caries
			3	2	72	73	Secondary caries
[14]	N = 1681 (age 1–4) Dundee, Scotland, UK 56 predictors	CHAID	3	3	65	69	High caries-risk (n = 784)
[15]	N = 1322 (age 0–5) KY, USA 7 predictors	CART	3	4	43	78	/
~	N = 341 (age 1–5) South Bačka area, Vojvodina, Serbia 35 predictors	RDT (acc.)	7	5	25	98	/
		CHAID	7	4	41	90	/
		ACT (lift)	2	2	22	95	/

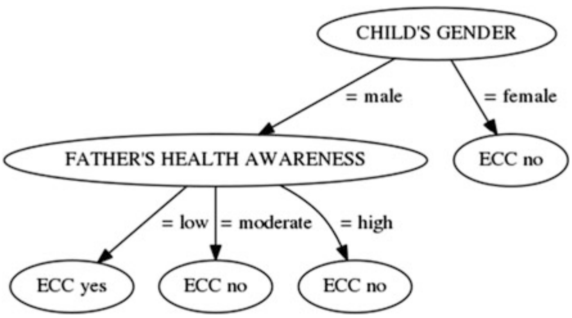

Fig. 1 An ACT obtained when splitting by lift gain

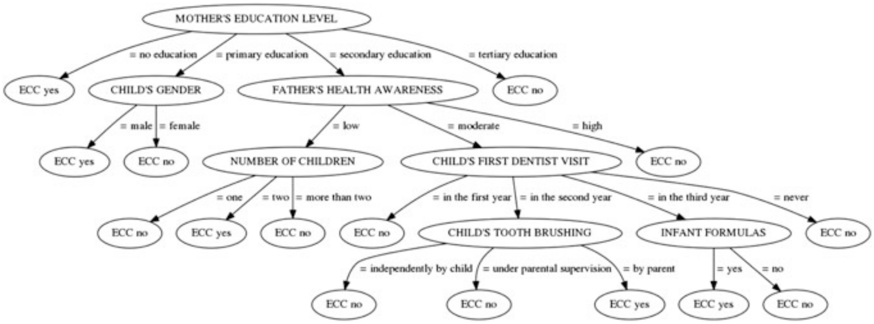

Fig. 2 A traditional decision tree obtained when applying CHAID

5 Discussion

The obtained ACTs are almost trivial and, with the exception of the ACT obtained when splitting by lift gain, have little value (see Table 1). The selected ACT (see Fig. 1) has better performance as it encapsulates an important pattern from a previous study [6]. It seems that the strong disproportion between the rules about ECC presence and absence (18 vs. 88142), prevents formation of a sound ACT.

The traditional decision trees outperform the ACTs in prediction, but are less readable (see Table 1). Splitting by the Gini index or information gain provides strong accuracy. However, the resulting trees are complex and there is most probably overfitting. Consequently, the most convenient traditional decision trees were those formed when splitting by accuracy or applying CHAID. Both decision trees are relatively compact trees of solid accuracy and both include variables about mother's education level and child's gender. Our preference for the CHAID tree was based on its higher TSS value. Some of the relevant variables that were identified in previous studies [6, 7, 9] may be observed in the CHAID tree, namely child's gender and father's health awareness. This decision tree also contains potential patterns as it includes variables concerning oral hygiene and related behaviour. Moreover, as shown in Fig. 2, if the mother has no education, or has completed primary school only and the child is male, the child is more likely to have ECC (20 of 29 instances, 69 %). When compared to the previous ECC models for the same area, the CHAID tree exhibits performance comparable to that of the logistic regression model.

Among the similar studies (see Table 2), the properties of the used data sets vary considerably. The largest data sets were used in [10, 14], and include numerous variables from various categories. Other data sets were focused mainly on narrow sets of predictors that include microbiological variables [11–13]. With respect to population's age, only the data sets used in [14, 15] seem to fully fit the early childhood category (less than 6 years of age), while other authors generally examined data about children aged 5 or more [10–12] and adults [13]. When compared to other data samples, the present data set has the fewest cases, but a relatively large set of potential predictors, and focuses completely on early childhood.

All of the reviewed trees (see Table 2) are generally readable as they contain from two to nine variables. The listed predictive performances of these models should not be compared directly because there are large differences in goals, evaluation methods and data sets across the studies. However, it appears that our selected decision trees have somewhat lower sensitivity, which could be attributed to the absence of microbiological predictors in the available data set. The key strengths of our selected decision trees are their high specificity and the possibility to apply them in practice more easily as they do not include variables that require oral examination of a child.

6 Conclusion

The decision trees appear to be very flexible models since their construction may be finely tuned to provide compact trees or strong predictive performance, depending on the actual needs. They may be well-suited to ECC research and ECC risk assessment, as they provide a readable but powerful common ground between domain experts and data analysts. Although decision trees that are summarized from association rules seem to have great potential, they did not provide satisfactory results. The particularities of the analyzed population and the data set, including data imbalance and low number of association rules for the positive class, might have been the reasons behind the poor performance. Our future research may include creating new algorithms for building associative decision trees that would be more resistant to these problems.

Acknowledgments The research presented in this paper was supported by the Ministry of Education, Science, and Technological Development of the Republic of Serbia under Grant III-44010. The authors are most grateful to Ivan Tušek and Jasmina Tušek for the provided data set and valuable support throughout the study.

References

1. Ghazal, T., Levy, S.M., Childers, N.K., Broffitt, B., Cutter, G.R., Wiener, H.W., Kempf, M.C., Warren, J., Cavanaugh, J.E.: Factors associated with early childhood caries incidence among high caries-risk children. Commun. Dent. Oral Epidemiol. **43**(4), 366–374 (2015)
2. Corrêa-Faria, P., Martins-Júnior, P.A., Vieira-Andrade, R.G., Marques, L.S., Ramos-Jorge, M. L.: Factors associated with the development of early childhood caries among Brazilian preschoolers. Braz. Oral Res. **27**(4), 356–362 (2013)

3. Tušek, I., Carević, M., Tušek, J.: Prevalence of early childhood caries among members of different ethnic groups in the South Bačka area (in Serbian). Vojnosanit. Pregl. **69**(12), 1046–1051 (2012)
4. Garcia, R., Borrelli, B., Dhar, V., Douglass, J., Ramos Gomez, F., Hieftje, K., Horowitz, A., Li, Y., Ng, M.W., Twetman, S., Tinanoff, N.: Progress in early childhood caries and opportunities in research, policy, and clinical management. Pediatr. Dent. **37**(3), 294–299 (2015)
5. Berkowitz, R.J.: Causes, treatment and prevention of early childhood caries: a microbiologic perspective. J. Can. Dent. Assoc. **69**(5), 304–307b (2013)
6. Ivančević, V., Tušek, I., Tušek, J., Knežević, M., Elheshk, S., Luković, I.: Using association rule mining to identify risk factors for early childhood caries. Comput. Methods Programs Biomed. **122**, 175–181 (2015)
7. Ivančević, V., Knežević, M., Tušek, I., Tušek, J., Luković, I.: Human friendly associative classifiers for early childhood caries. In: 7th KES International Conference on Intelligent Decision Technologies (KES-IDT 2015), pp. 243–253. Springer (2015)
8. Chen, Y.-L., Hung, L.T.-H.: Using decision trees to summarize associative classification rules. Expert Syst. Appl. **36**, 2338–2351 (2009)
9. Tušek, I.: The Influence of social environment and ethnicity on caries prevalence in the early childhood (in Serbian). Ph.D. thesis, University of Belgrade (2009)
10. Stewart, P.W., Stamm, J.W.: Classification tree prediction models for dental caries from clinical, microbiological, and interview data. J. Dent. Res. **70**(9), 1239–1251 (1991)
11. Gansky, S.A.: Dental data mining: potential pitfalls and practical issues. Adv. Dent. Res. **17**, 109–114 (2003)
12. Tamaki, Y., Nomura, Y., Katsumura, S., Okada, A., Yamada, H., Tsuge, S., Kadoma, Y., Hanada, N.: Construction of a dental caries prediction model by data mining. J. Oral Sci. **51**, 61–68 (2009)
13. Ito, A., Hayashi, M., Hamasaki, T., Ebisu, S.: Risk assessment of dental caries by using classification and regression trees. J. Dent. **39**, 457–463 (2011)
14. MacRitchie, H.M.B., Longbottom, C., Robertson, M., Nugent, Z., Chan, K., Radford, J.R., Pitts, N.B.: Development of the Dundee Caries Risk Assessment Model (DCRAM)—risk model development using a novel application of CHAID analysis. Commun. Dent. Oral Epidemiol. **40**, 37–45 (2012)
15. Li, H.F.: Data mining and pattern discovery using exploratory and visualization methods for large multidimensional datasets. Ph.D. thesis, University of Kentucky (2013)
16. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth, Belmont CA (1984)
17. Kuhn, M., Johnson, K.: Applied Predictive Modeling. Springer (2013)
18. Kass, G.V.: An exploratory technique for investigating large quantities of categorical data. Appl. Stat. **29**(2), 119–127 (1980)
19. Säuberlich, F., Gaul, W.: Decision tree construction by association rules. In: 23rd Annual Conference of the Gesellschaft für Klassifikation, pp. 245–253. Springer (2000)
20. Wang, K., Zhou, S., He, Y.: Growing decision trees on support-less association rules. In: 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'00), pp. 265–269. ACM (2000)
21. Abdekhalim, A., Traore, I., Sayed, B.: RBDT-1: a new rule-based decision tree generation technique. In: International Symposium on Rule Interchange and Applications (RuleML 2009), pp. 108–121. Springer (2009)
22. Peng, Y., Ye, Y., Yin, J.: Decision tree construction algorithm based on association rules. In: 2nd International Conference on Computer Application and System Modeling (ICCASM 2012), pp. 754–756. Atlantis Press (2012)
23. RapidMiner Studio—RapidMiner. https://rapidminer.com/products/studio/
24. Miller, G.A.: The magical number seven plus or minus two: some limits on our capacity for processing information. Psychol. Rev. **63**(2), 81–97 (1956)

25. Fontana, M.: The clinical, environmental, and behavioral factors that foster early childhood caries: evidence for caries risk assessment. Pediatr. Dent. **37**(3), 217–225 (2015)
26. Allouche, O., Tsoar, A., Kadmon, R.: Assessing the accuracy of species distribution models: prevalence, kappa and the True Skill Statistic (TSS). J. Appl. Ecol. **43**, 1223–1232 (2006)
27. Cohen, J.: A coefficient of agreement for nominal scales. Educ. Psychol. Measur. **20**(1), 37–46 (1960)
28. Graphviz—Graph Visualization Software. http://graphviz.org/