Original Research Article

Machine Learning Regression Tree Approach for Age Prediction from Eruption Status of Permanent Teeth in Sri Lankan Children

Abstract

Prediction of age is a predominant facet in forensic and clinical fields. Forensic odontology is used to predict an age by using permanent teeth resistant to high temperatures and any mass disaster than other parts of the body. Although many studies have been carried out in other countries, this is the first study of age prediction using the eruption status of permanent teeth for Sri Lankans. Age-related memory loss, memory loss associated with dementia, and the absence of official documents to verify their age are the main reasons people have no knowledge about their age. Therefore, age prediction is used in various situations, such as identification, admission purposes, employment, criminal issues and judicial punishments. The main objective of this study is to predict the age of a child using the eruption status of permanent teeth. This cross-sectional study was conducted on 3321 individuals (1681 males and 1640 females) from 7 provinces and 20 schools in Sri Lanka. Regression tree algorithms in Machine learning were used for age prediction. Classification and regression trees (CART), gradient boosting (GB) classifier and extreme gradient boost (XGBoost) classifier were used to make predictions for the age of a child. Results were validated using cross-validation techniques, and root mean squared error (RMSE) and R-squared values were used as accuracy measures to select the best model. The best model for age prediction was the XGBoost model, which gives the highest accuracy (88%). This is the first study of age prediction using eruption status of permanent teeth for Sri Lankan children. The study results provide an XGBoost machine learning classifier as the most suitable method for age prediction with higher precision.

Keywords: Decision Trees, Machine Learning, Permanent Dentition, Regression Analysis, Tooth Eruption.

I Introduction

All humans have an identity in life; age is one of the essential criteria for identification both in living and dead bodies. Age prediction plays an important aspect in various fields such as the clinical practice of Pediatric medicine, Pediatric dentistry, Endocrinology, identification of crimes and accidents etc. [1]. Forensic odontology is used to predict an age by using permanent teeth which are highly durable and resist putrefaction, fire, chemicals etc. Age-related memory loss, memory loss associated with dementia, and the absence of official documents to verify their age are the main reasons people do not know their age. Therefore, age prediction is used in various situations, such as identification, admission purposes, employment, criminal issues and judicial punishments [2].

The tooth consists of two major parts, namely the root (hidden in the gums) and crown (visible part of the tooth). The process in which the teeth enter the mouth and become visible is known as tooth eruption. The first human teeth to appear, or the eruption of primary teeth, occurs roughly between 6 months to 2 years of age [3]. Permanent teeth are the next set of teeth to erupt into the oral cavity after exfoliating primary teeth. There are 32 permanent teeth with both maxilla and mandible are consisted of 16 each. As shown in Figure 1, an arch consists of two central incisors, two lateral incisors, two canines, four premolars, and six molars. Moreover, the teeth that fail to erupt through the gums are impacted or unerupted. 3rd molars (wisdom teeth) are the most common teeth to face with this problem due to the fact they are the last teeth to emerge [4-6].

Although age assessment is an archaic exercise that forensic anthropologists have practised in recent decades, a newly emerging branch known as "Forensic Odontology" is used to predict an age by using tooth/teeth [7] because teeth can be preserved for an extended period compared to the other parts of the body which are resistant to high temperature and any mass disaster [7, 8].

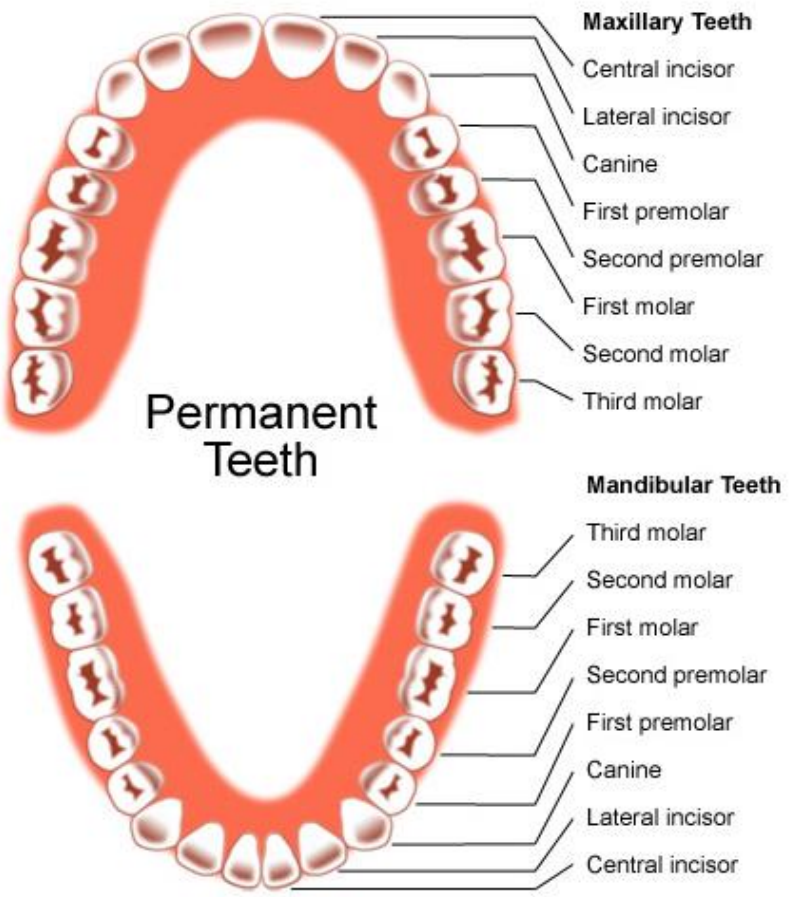

Figure 1: Structure of permanent teeth with names.

One of the fastest, low-cost and noninvasive methods to identify dental age in subadults is dental eruption acquired through visual examination of the oral cavity. A study was conducted to estimate the age of subadults with ages ranging from 1 to 17 years based on dental eruption status [2]. The Least Squares regression analysis technique was used to compare the teeth present at different ages separately by gender and primary and permanent teeth.

Human identification is one of the most challenging subjects humans have been confronted with. Although many studies have been carried out to make this process more efficient and effective [3, 4], this is the pioneer study of age prediction using the eruption status of permanent teeth for the Sri Lankan population. This article is organized as follows. Section 2 describes the nature of the data set utilized for analysis, regression techniques used to predict the eruption status of permanent teeth with its model validation techniques. In contrast, Section 3 expands on the statistical analysis conducted with R statistical software Version 4.1.2 and the key results derived. In discussion, Section 4 sums up the main findings.

II Methods

This section outlines the research methods that were followed in this study. It provides information about the nature of the dataset utilized for analysis and different regression techniques used to analyze the data with its model validation techniques.

1. Data

In this cross-sectional study, 3321 school children (1681 males and 1640 females) have been used, and the eruption status of permanent teeth was noted in the age group from 4 to 18 years. Age is continuous, obtained from the date of birth of a child from the school records, and the other variables are categorical, which are eruption status of 28 permanent teeth. The study was carried out with the permission of parents of the school children and the management of the schools.

In here, a tooth was considered erupted if more than 1/3 of the crown is visible, unerupted if not present in the oral cavity. Trained investigators at the school premises were recorded the eruption status of permanent teeth. The eruption time of the third molar is 17-21 years. Most of the individuals' age was lower than 17 years. Therefore, the eruption status of permanent teeth with the exception of the third molar was noted in this study.

2. Regression Techniques

The Classification and Regression Tree (CART) is a decision tree algorithm with a tree-like structure [9]. There are two types of machine learning algorithms called supervised and unsupervised. CART is a supervised machine learning algorithm that lies at the foundation of machine learning decision tree algorithms. The most commonly used supervised machine learning methods are tree-based algorithms which construct predictive models with high precision, stability and ease of interpretation [10]. Since tree-based algorithms are non-parametric and map non-linear relationships quite well, an essential advantage over linear models, they are among the leading supervised learning methods in machine learning.

A decision tree is a binary tree that is constructed by recursive partitioning. The partitioning procedure starts from the first parent node, called the root node, and each node is split into

two parts: the left child node and the right child node [11]. Then each child node becomes a parent node and is split into two nodes as left and right child nodes. This recursive partitioning procedure will repeat until each child node is pure. Starting from the root node, the feature that results in the most significant Information Gain (IG) is selected as the parent node at each partitioning step.

In practice, this may result in a tree with an abysmal depth and many nodes. Hence, to minimize the overfitting, it is necessary to prune the tree by setting a limit for the maximal depth of the tree [11]. The accuracy of a tree is heavily affected by the use of strategic splits. There are two types of decision trees: classification trees and regression trees used for classification problems and regression problems. Decision criteria for classification and regression trees are different [12]. For regression trees, the decision criteria are the weighted mean squared error (MSE) of the children nodes, which is minimized to choose split points.

$$\boldsymbol{MSE(t)} = \frac{\mathbf{1}}{\boldsymbol{N_t}} \sum_{\boldsymbol{i \epsilon D_t}} \left(\boldsymbol{y^{(i)}} - \boldsymbol{\hat{y}_t}\right)^{\mathbf{2}} \tag{1}$$

Here, N_t is the number of training samples at node t, Dt is the training subset at node t, $y^{(i)}$ is the true target value, and $\hat{y}_t$ is the predicted target value [12].

$$\boldsymbol{\hat{y}_t} = \frac{\mathbf{1}}{\boldsymbol{N_t}} \sum_{\boldsymbol{i \epsilon D_t}} \boldsymbol{y^{(i)}} \tag{2}$$

After comparing MSE values among variables, the variable or point with the lowest MSE is selected as the split point, a recursive process.

Boosting is a machine learning ensemble technique in which the predictor variables are not made independently, but sequentially which converts weak learners into strong learners [13]. Gradient Boosting is a technique that constructs a collection of models and trains them in a sequential, gradual and additive manner. This technique is based on the logic that subsequent predictors learn from the mistakes of the previous predictors [14]. Consequently, the observations have an unequal probability of appearing in subsequent models, and the ones

with the highest error appear most. Hence, the observations are selected based on the error. Gradient Boosting reduces the bias and variance.

Gradient boosting technique mainly focuses on optimizing the cost function, using weak learners to make predictions and addition of vulnerable learners to minimize the loss function using an additive model. In the gradient boosting technique, the loss function is a measure used to identify the model's coefficient at fitting the underlying data, and decision trees are used as weak learners [14]. Specifically, regression trees are used that output absolute values for splits and whose output can be added together, allowing subsequent model outputs to be added and "correct" the residuals in the predictions. Decision trees are added continuously at each step while the existing trees in the model remain unchanged.

E**x**treme **G**radient **B**oosting (XGBoost) is a fast machine learning algorithm that implements gradient boosting algorithms with several regularization techniques. Moreover, XGBoost is a supervised learning method that optimizes specific loss functions based on function approximation [15]. Decision tree-based algorithms are highly used for small-to-medium structured data in prediction problems, while artificial neural networks are used for unstructured data.

XGBoost and Gradient Boosting Machines (GBMs) are both ensemble tree methods that apply the principle of boosting weak learners (CARTs generally) using the gradient descent architecture [16]. However, XGBoost improves upon the base GBM framework through systems optimization and algorithmic enhancements. XGBoost has improved the system performances through Parallelization, tree pruning, hardware optimizations [17].

3. Model Validation

Root Mean Square Error (RMSE) and R-Squared values are used to validate the performance of the fitted models. RMSE is the standard deviation of the prediction errors (residuals) and a measure that represents the spread of data points around the regression line. To calculate the RMSE, the following equation is used [18].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2} \quad (3)$$

where *n*-number of samples, f-actual values and o-observed values.

R-squared (R^2) is a measure that can be used to check the model performance based on the proportion of variance of the dependent variable, which is explained by an independent variable or variables in a regression model. The formula for the R-Squared is defined as follows [18].

$$R^2 = 1 - \frac{Explained\ Variation}{Total\ Variation} \quad (4)$$

Cross-validation is a data resampling method to estimate models' true prediction error and true model parameters [19]. The basic form of cross-validation is a k-fold cross-validation data set that is portioned into *k* equally sized sets. *k* iterations are used to validate and train the model, and *k*-1 folds are used to learn. A 10-fold cross-validation technique whare *k* = 10 is used to validate the model. The data set is divided into training and testing groups in this method. The training set is used to learn or train the model, while the test set predicts the model.

III Results and Discussion

We used several statistical methods, algorithms and classifiers to fit a model for age prediction using eruption status of permanent teeth. In this section, the results of all these methods are discussed. The mean age with a standard deviation of total sample, females and males are 11.26±3.12, 11.20±3.13 and 11.32±3.11 years, respectively. Table 1 shows the accuracy of using the three different age prediction methods: CART algorithm, Gradient Boosting algorithm, and XGBoost classifier.

Table 1: Accuracies of age prediction using three different techniques.

Model	RMSE	R-squared	Accuracy (%)
CART	0.1396	0.7591	86.04

GDBM	0.1201	0.8124	87.99
XGBoost	0.1194	0.8276	88.06

Among the three regression methods used to predict the age of children using all 28 variables as predictors, the XGBoost model has the lowest RMSE value, the highest R-squared value and the highest accuracy. Therefore, extreme gradient boosting classifier is the best method to predict the age among these methods with the most influential variables upper left central incisor, lateral incisor, canine, first premolar, second premolar, upper right lateral incisor, canine, first premolar, second premolar, the second molar, lower left lateral incisor, second molar and lower right lateral incisor, canine, first premolar, second premolar as predictors.

Being able to predict the age using permanent teeth becomes an essential task in Forensic Odontology. Hence, the present study was conducted to predict the age of Sri Lankan children using eruption status of permanent teeth, and machine learning regression tree techniques have been used for age prediction due to the existence of multicollinearity among independent variables. All three regression tree techniques were performed better in age prediction with accuracy higher than 80%. XGBoost model has the highest accuracy of 88% and the highest R-square value 0.828 compared to the other two models, proving that this model can be used to predict the age of school children from eruption status of permanent teeth.

Across the world, many studies were conducted for age prediction using permanent teeth [1, 2, 3, 7, 8]. Different features/measurements of permanent teeth have been used in various studies. A similar study has been carried out to predict the age of children in Romania using the eruption status of permanent teeth, which is only one so far [2]. Thus, this study becomes a novel study for Sri Lankan children. Besides, it indicates that the R-squared values of the fitted models were higher than 80% in both the present study and the Romanian study. Since the usage of different sampling techniques and research methodology of the studies, model performances can differ. Hence, a direct comparison of model performances cannot be conducted for those studies. Furthermore, in future studies, the performance of these models can be improved using a larger dataset that is gathered for a more extended period and collected from different areas.

IV Conclusion

The eruption time difference was statistically significant between males and females for all permanent teeth, except for the central incisor and second premolar. No statistical difference was found between the eruption time of the left and right sides. The eruption time difference between the mandible (lower jaw) and maxilla (upper jaw) was statistically insignificant in the second premolar and first molar**.** In this study, the extreme gradient boosting (XGBoost) model, a machine learning regression model, has been proposed to predict age with 88% accuracy.

References

[1] Gupta, B., Sudha, P., Anegundi, R., & Indushekar, K. R. (2012). Importance of Dental Age and Skeletal Age in Forensic Sciences for the Assessment of Pediatric Growth and Development. *Forensic Medicine and Toxicology*, *6*(2), 20-24.

[2] Dogaroiu, C., Cosma, A., Gherghe, E. V., Morosanu, G., & Avramoiu, M. (2015). Age estimation in subadults using teeth eruption examination. *Romanian Society of Legal Medicine*, *23*, 49-56. DOI: 10.4323/rjlm.2015.49

[3] Hulland, S. A., Lucas, J. O., Wake, M. A., & Hesketh, K. D. (2000). Eruption of the primary dentition in human infants: a prospective descriptive study. *Pediatric dentistry*, *22*(5), 415-421.

[4] Jayatileke, A., Nawarathna, L. S., Ravishanker, N., Wijekoon, P., Nawarathna, R. D., Vithanaarachchi, V. S., & Wijeyaweera, R. L. (2019). New standards for eruption time and sequence of permanent dentition in Sri Lankan children. *Communications in Statistics: Case Studies, Data Analysis and Applications*, *5*(2), 92-100. https://doi.org/10.1080/23737484.2019.1579075

[5] Renton, T., & Wilson, N. H. (2016). Problems with erupting wisdom teeth: signs, symptoms, and management. *The British journal of general practice: The Journal of the Royal College of General Practitioners*, 66, no. 649 (2016): e606-e608. DOI: 10.3399/bjgp16X686509

[6] Vithanaarachchi, V. S. N., Nawarathna, L. S., & Wijeyeweera, R. L. (2018). Eruption times and patterns of permanent teeth in Sri Lankan school children in Western province. *Sri Lanka Dent J*, *48*(01), 25-31.

[7] Singh, C., & Singal, K. (2017, 12). Teeth as a Tool for Age Estimation: A Mini Review. *Forensic Sciences And Criminal Investigation*, *6*(3), 555695. DOI: 10.19080/JFSCI.2017.06.555695.

[8] Uzuner, F. D., Kaygisiz, E., & Darendeliler, N. (2017). Defining dental age for chronological age Determination. In *Post Mortem Examination and Autopsy-Current Issues From Death to Laboratory Analysis*. IntechOpen. DOI: 10.5772/intechopen.71699

[9] Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR)*, *5*(4), 2094-2097. DOI:10.21275/v5i4.nov162954

[10] Kareem, S. A., Raviraja, S., Awadh, N. A., Kamaruzaman, A., & Kajindran, A. (2010). Classification and regression tree in prediction of survival of aids patients. *Malaysian Journal of Computer Science*, *23*(3), 153-165. DOI: 10.22452/mjcs.vol23no3.2

[11] Loh, W.-Y. (2008). Classification and Regression Tree Methods. *Encyclopedia of Statistics in Quality and Reliability*, 1, 315-323. https://doi.org/10.1002/9780470061572.eqr492

[12] Patel, B. R., & Rana, K. K. (2014). A Survey on Decision Tree Algorithm For Classification. *International Journal Of Engineering Development And Research*, *2*(1). DOI:10.21275/v5i4.nov162954

[13] Patri, A., & Patnaik, Y. (2015). Random Forest And Stochastic Gradient Tree Boosting Based Approach For The Prediction Of Airfoil Self-Noise. *Procedia Computer Science*, *46*, 109-121. https://doi.org/10.1016/j.procs.2015.02.001

[14] Natekin, A., & Knoll, A. (2013). Gradient Boosting Machines, A Tutorial. *Frontiers In Neurorobotics*, **7**, 21. DOI: 10.3389/fnbot.2013.00021

[15] Zhou, Y., Li, T., Shi, J., & Qian, Z. (2019). A CEEMDAN and XGBOOST Based Approach to Forecast Crude Oil Prices. *Complexity*, 15. https://doi.org/10.1155/2019/4392785

[16] Ramraj, S., Uzir, N., Sunil, R., & Banerjee, S. (2016). Experimenting XGBoost algorithm for prediction and classification of different datasets. *International Journal of Control Theory and Applications*, *9*, 651-662. DOI: http://dx.doi.org/10.1145/2939672.2939785

[17]Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 785-794. https://doi.org/10.1145/2939672.2939785

[18] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to Statistical Learning (Vol. 112, p. 18). New York: Springer.

[19] Danniel, B. (2018). Cross-Validation. *Encyclopedia of Bioinformatics and Computational Biology*, *1*, 542-545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X