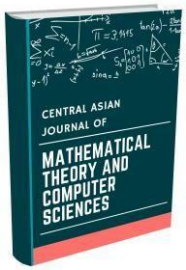

CENTRAL ASIAN JOURNAL OF MATHEMATICAL THEORY AND COMPUTER SCIENCES
https://cajmtcs.centralasianstudies.org/index.php/CAJMTCS
Volume: 06 Issue: 01 | January 2025 ISSN: 2660-5309

Article

Predictive Modeling Based on Machine Learning Techniques to Predict Tooth Loss Among Adults in Nasiriyah City, Iraq

Ghosoon k. Munahy*[1], Nabra F. Salih[2], Manar Hamza Bashah[3]

1. Department of Information Technology, College of Computer Science and Information Technology, University of Kerbala, Kerbala, Iraq
2. College of Dentistry, University of Thi Qar, Thi Qar, Iraq
3. Department of Information Technology, College of Computer Science and Information Technology, University of Kerbala, Kerbala, Iraq

* Correspondence: ghosoon.k@uokerbala.edu.iq

Abstract: Tooth loss negatively affects the overall physical and social well-being of adults. Controversy surrounds the socioeconomic variables that influence tooth loss, an issue that can negatively impact a person's quality of life. Better diagnostic tools and medical data analysis can be achieved in the healthcare industry through the application of machine learning. Machine learning is focused on mimicking human learning with data and algorithms and gradually increasing the model's accuracy. In order to predict adult tooth loss both completely and incrementally, this study will develop a machine-learning approach and compare the predictive performance of various models. We first collected the most data from the Dental clinics of the College of Dentistry at the University of Thi Qar. Then, we analyzed the data using statistical tools and developed a system using machine learning techniques for predicting tooth loss among adults. We found that tooth loss and socioeconomic factors are often linked, with individuals from lower socioeconomic backgrounds experiencing higher rates of tooth loss. Several factors contribute to this relationship, including access to dental care. Our predictive approach achieves high performance when we use a random forest algorithm, with results of 94.0%, 97.3, 97.1%, 97.4%, 97.3%, and 81.4% for AUC, CA, F1, Precision, Recall, and MCC, respectively.

Keywords: Machine Learning, Predictive Analytics, Tooth Loss, Random Forest

Citation: Munahy, G. K., Salih, N. F., Bashah, M. H. Predictive Modeling Based on Machine Learning Techniques to Predicate Tooth Loss Among Adults in Nasiriyah City, Iraq. Central Asian Journal of Mathematical Theory and Computer Sciences 2025, 6(1), 68-75.

Received: 10th Dec 2024
Revised: 23th Dec 2024
Accepted: 20h Jan 2025
Published: 31th Jan 2025

1. Introduction

Predictive analytics is a game-changing tool that can improve patient outcomes, encourage a proactive approach to oral health management, and revolutionize the delivery of dental healthcare. Through the utilization of data and analytics, dental professionals can improve preventive measures, optimize treatment plans, and advance oral health equity for all people and communities. Dental healthcare will be data-driven in the future, and predictive analytics will remain essential in determining how oral health and well-being are shaped [1].The final stage of dental disease is the loss of teeth. Loss of teeth can notably impact people's overall health. An early diagnosis of tooth disease can prevent tooth loss. Many low-income and low-academic-level adults have poor oral health due to poor routines [2]. Scientists or health care providers can use machine learning techniques to integrate large-scale heterogeneous data to understand the basis of disease better and to predict or identify the disease in its early stages [3]. Disease prediction and healthcare are some of the important applications of machine learning, where its techniques are utilized

to analyze big datasets and predict the desired output.[4] Predicting tooth loss can help healthcare officials and dentists take early measures to protect against tooth loss[5].

The objective of predictive analytics is to build machine learning models to predict patterns of behaviors and use predictions to make decisions. Multidimensional data is organized as a table of records, where a set of features represents each record; one of these features is the target to be predicated [6]. Four predictive Models were used in this paper: Naïve Bayes, support vector machine, decision tree, and random forest. Then, we used a 10-fold cross-validation for evaluating predictive models.

2. Related Works

Numerous investigations have endeavored to forecast the eventuality of tooth loss and its associated risk factors through the application of machine learning techniques. To illustrate, [7] used a two-step machine learning technique in which a decision tree was used to create rules about tooth loss and then the rules formulated were related to the outcome variable. Similarly, [8] introduced a system for predicting diabetes that worked based on three machine learning algorithms namely; Gradient Boosting, Decision Tree and Logistic Regression. The algorithm with the best performance was the Decision Tree with a score 91%. According to [9] the development of a system to predict the loss of teeth was done based on machine learning algorithms and particularly Gradient-Boosting Trees achieved the best results on their dataset. More so, [10] observed that the main reasons for the loss of teeth included dental caries (85.4%), disease (13.7%) and trauma (7.5%). Their work revealed that increased tooth loss was related to age, smoking, education levels, and brushing teeth. In another research [11], the object was to gauge the frequency of dental decay among young adults based on the Iowa Fluoride Study longitudinally determined predictors such as diet, social and economic status and an individual's past dental history. Four machine learning techniques were employed in the analysis, out of which the LASSO regression model reached the highest level of accuracy. On those two significant predictors, previous caries and sugar-sweetened beverage consumption, ages 13 to 17. Nonetheless, further confirmation must be done in diverse populations, this model appears useful as a strategy for health interventions .A study aimed[12] at estimating DMFT (Decayed, Missing, and Filled Teeth) scores based on clinical examination data was administered to 2000 patients at Gaziantep University. For the model, various machine learning algorithms were put to use: Naive Bayes, Logistic Regression, and Multilayer Perceptron. The model of Multilayer Perceptron produced the most favorable classification results, with the accuracy of 95.8%, which was indicative of the success of caries risk assessment model, as an ideal situation of implementing machine learning to the prediction of complicated oral health condition.

3. Materials and Methods

This research takes into consideration the stages of tooth loss prediction in a machine learning context and implements quite a number of them. There are five broad stages to the process as shown in Figure 1. This is where the process starts as data collection is the first step in the process where all possible data is collected from different avenues like medical records, and data from the clinics, and health-related questions amongst others in order to have all the data that involves tooth loss [13]. Data preprocessing comes next, which involves cleaning up of the data and manipulating it in such a way that errors are fixed and missing values corresponded to, this is important in increasing and improving the reliability of the model. And then, decision making about which characteristics contribute mostly to tooth loss occurs also in order to prevent unnecessary complexity of the model and to make predictions as efficient as possible. After the features are determined, the machine learning approach turns into action and the model is trained on the already prepared data in order to learn the correlations between parameters leading to proper prediction. At last, the model application is completed by a system evaluation, the

purpose of which is to evaluate the quality of the evaluation model, for example, by other measures: accuracy, recall, and precision, as well as to find out what areas are in need for improvement, ensuring sufficiency of the model.

3.1 Data collection

The data for this prospective case-control study was gathered over a time frame of six months from October 2022 until April 2023 at four primary health care centers, namely, Al-Shamiya Specialized Dental Center, Sumer Dental Center, educational clinics in the Faculty of Dentistry of University of Thi Qar, and Al-Hussein Teaching Hospital. These centers cater for a wide population mix enabling the study results to depict various income levels which improves the applicability of the study results in the real world. Five-hundred and sixty patients' charts of records were analyzed in total including a host of demographic, clinical and behavioral factors that bear influence in tooth loss outcomes [14]. Characteristics of data, including those presented by age, sex, level of education, oral health practices and the history of disease are detailed in the table below. A sample of the dataset is also shown in Table 1, indicating the collected data and its format. The large dataset will be instrumental in building models to predict and spot risk factors for tooth loss.

Table.1 Data Features Description.

Features	Description
Age	Years
Gender	Male / Female
Academic Achievement	Illiterate/intermediate education/graduate/postgraduates
Income	Limited/middle/good/ very good
Eating Sweet Times	Number of times
Eating Vegetables Times	Number of times
Brushing Teeth Times	Number of times
Smoking Habit	Smoker/nonsmoker
No. Of Decayed Teeth	Number of decayed teeth

3.2 Data Preprocessing

Data preprocessing is the process of cleaning and preparing raw data to make it suitable for use in analytical models or machine learning algorithms and includes steps such as:

1. Converting categorical data to numeric codes:

Categorical data such as "gender" and "education" must be converted to numeric codes. For example, the "male" gender can be set to 0 and "female" to 1.

2. Handling missing data

Check for missing data and replace it. The mean or most frequent value can be used to fill in the blanks.

3. Splitting the data into a training and test set

To ensure that the model is tested properly, the data must be split into a training set to train the model and a test set to evaluate its performance.

3.3 Feature selection

Feature selection is the process of selecting a subset of the most relevant features from an original dataset to use in building the model. It aims to improve model performance and reduce complexity by removing unnecessary or redundant features, which helps improve model accuracy and minimize training time. There are many methods for feature selection, one of which we used in this study is chi-square. We used the chi-square test because it fits well with the nature of the study data, as it helps determine the relationship

between each feature and the outcome of tooth loss. The chi-square test provides an easy and quick way to identify the most important features by analysing the statistical significance of each feature. Features that show a strong association with the outcome, i.e. those with low p-values, are selected for inclusion in the predictive model. This helps improve the accuracy of the model and reduces its complexity by eliminating ineffective features.

Table 2. Data set Sample,

Age	Gender	Academic Achievement	Income	Eating Sweets Times	Eating Vegetables Times	Brushing Teeth Times	Smoking Habit	No. Of Decayed Teeth	Missing Teeth
18-35	female	intermediate education	limited	3	2	1	Smoker	3	1
18-35	female	graduate	middle	4	1	0	Non -smoker	4	1
36-50	male	graduate	middle	1	3	1	Non -smoker	2	0
18-35	female	intermediate education	limited	4	1	2	Non -smoker	4	1
18-35	female	intermediate education	limited	3	1	0	Non -smoker	5	1
36-50	male	postgraduates	middle	1	3	3	Non -smoker	1	0
18-35	male	intermediate education	limited	3	3	3	Non -smoker	2	1

3.4 Machine Learning Models:

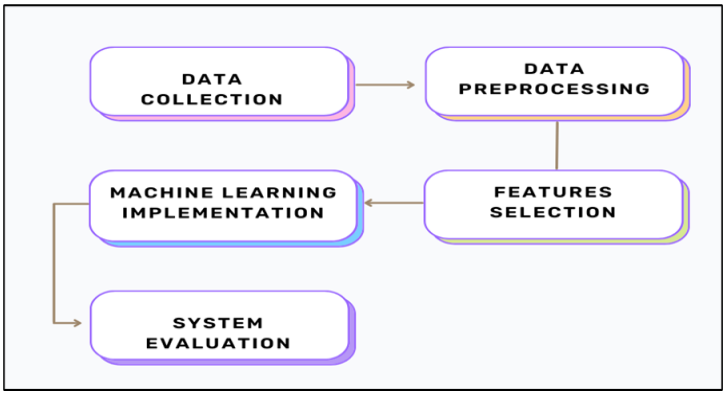

Figure 1. Proposed approach phases

In this study, we used four predictive models:

1. Naïve Bayes: Byes'theorm (Bayes' law) is a probability theory that relates the conditional and marginal probabilities of two random events. A naive Bayes classifier is a type of classification algorithm that is based on Bayes' theorem with a simple assumption called "conditional independence", which states that the presence of a certain feature in a class is not related to the presence of any other feature.

For example, a fruit may be classified as orange if it has features round, is about 4" in diameter and has an orange colour. Although each feature in this fruit depends on presence of the other features, a naive Bayes classifier supposes all of these features contribute independently to the probability that it is an orange fruit [15].

2. Sport Vector Machine:

It is a supervised machine-learning algorithm used for classification. The SVM algorithm may also be applied in statistical learning [16]. This algorithm works by finding the best "hyperplane" that separates different classes in a feature space.

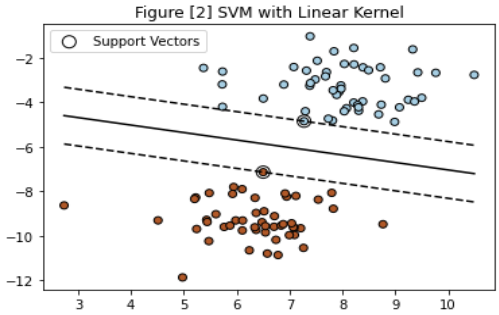

Figure 2. SVM Classifier.

"Figure 2" refers to an example of an SVM classifier. Although SVM is a delicate method, it may be computationally complex since it sometimes uses mathematical functions that need sophisticated [17]

3. Decision Tree: is a supervised machine-learning model used for classification problems. The tree consists of a number of nodes like tree structures. Each node represented a decision based on features' values. "Figure 3" refers to an example of the decision tree structure.

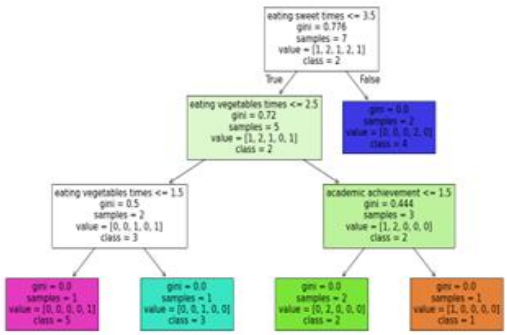

Figure 3. Part Of Tree.

4. Random Forest: is a supervised predication machine-learning model introduced by Breiman[18]. Many decision trees are constructed for model predictions in the random forest using a randomly selected training dataset. The results of each tree are combined to give a prediction for each observance [19]. This model makes a final prediction based on a majority voting of the decision trees. "Figure 4" refers to an example of a random forest structure.

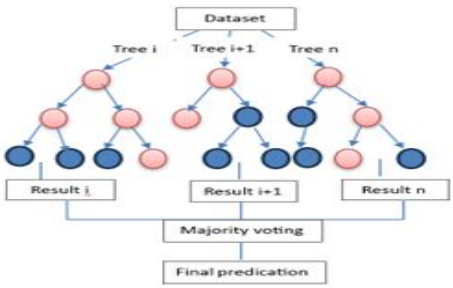

Figure 4. Example of a Random Forest Structure.

3.5 Evaluation System: In this study, we used the k-fold cross-validation method to evaluate the performance of our proposed system based on the value of AUC, F1, recall, precision and MCC. k-fold cross-validation can be built, is adaptable, and often does not

depend on model structure [20]. In this paper, we used 10 folds; we divided the data into 10 equal parts, using each part once as test data and the other nine parts as training data. This gives us an idea of how the model performs on various data splits.

4. Discussion

Table 3. Feature's Chi-Square Value.

Model	AUC	CA	F1	Prec	Recall	MCC
Random Forest	0.94	0.97	0.97	0.97	0.97	0.81
Tree	0.72	0.96	0.96	0.96	0.96	0.76
SVM	0.94	0.94	0.93	0.94	0.94	0.58
Naïve Bayes	0.93	0.92	0.92	0.92	0.92	0.53

After applying data preprocessing techniques to our collected data, we reduced the features from nine to five based on their Chi-square value "Table 3". The other four features were removed because they have a p-value of more than 0.05. "Table 4"compares the performance of the machine learning algorithms used in this paper based on six performance evaluation measures. AUC (area under the ROC curve) measures the model's ability to distinguish between classes. ROC curves for each model. CA or classification accuracy measures the percentage of correct predictions, precision(prec) measures the ratio of correct positive predictions to total positive predications, Recall calculates the ratio of true positive to the total actual positive, F1 score is the harmonic average between precision and recall, which provides a balance between them, and Matthews Correlation Coefficient (MCC) gives a single value between -1 and +1 based on confusion matrix values, is considered balanced metric. The table shows that the random forest algorithm performs better than the rest. It has high performance in all indicators. The high MCC value refers to the random forest model being a very balanced model. Although the decision tree's AUC is low, it has high classification accuracy, F1, Prec, and Recall. In general, it has less balance than the random forest model.

Table 4. Performance of The ML Algorithms.

Features	Chi Squair
No. Of Decayed Teeth	53.851
Income	27.691
Academic Achievement	21.838
Smoking Habit	10.545
Age	7.786
Brushing Teeth Times	5.824
Gender	2.508
Eating Sweet Times	1.037
Eating Vegetables Times	0.9248

The SVM and Naïve Bayes models show good AUC. Still, they are lower in other indicators, especially the MCC, which indicates that they may not be equally reliable in terms of the balance of expectations

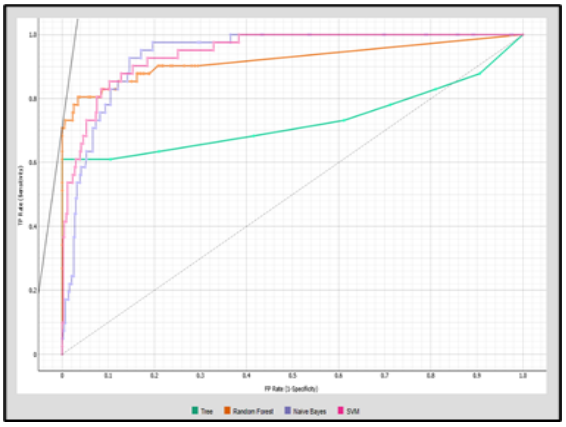

Figure 5. ROC Curve.

5. Conclusion

In this paper, we applied four machine-learning models to predict tooth loss among adults. The dataset was collected from some Specialized Dental care Centers in Nasiriyah City. Before implementing the ML algorithms on data, we preprocess this data, and then we reduce its features based on the chi-square method. our experiment performance of the random forest algorithm with 94.0%, 97.3, 97.1%, 97.4%, 97.3%, and 81.4% for AUC, CA, F1, Precision, Recall and MCC, respectively

REFERENCES

[1] M. Bari Antor, et al., "A comparative analysis of machine learning algorithms to predict Alzheimer's disease," Journal of Healthcare Engineering, vol. 2021, no. 1, p. 9917919, 2021.

[2] H. W. Elani, et al., "Predictors of tooth loss: A machine learning approach," PLoS One, vol. 16, no. 6, p. e0252873, 2021.

[3] N. M. Salem, et al., "Machine and deep learning identified metabolites and clinical features associated with gallstone disease," Computer Methods and Programs in Biomedicine Update, vol. 3, p. 100106, 2023.

[4] R. Venkatesh, C. Balasubramanian, and M. Kaliappan, "Development of big data predictive analytics model for disease prediction using machine learning technique," Journal of Medical Systems, vol. 43, no. 8, p. 272, 2019.

[5] K. Ramudu, et al., "Machine learning and artificial intelligence in disease prediction: Applications, challenges, limitations, case studies, and future directions," in Contemporary Applications of Data Fusion for Advanced Healthcare Informatics, IGI Global, 2023, pp. 297–318.

[6] Y. Jiang, et al., "FXAM: A unified and fast interpretable model for predictive analytics," Expert Systems with Applications, vol. 252, p. 123890, 2024.

[7] U. Cooray, et al., "Importance of socioeconomic factors in predicting tooth loss among older adults in Japan: Evidence from a machine learning analysis," Social Science & Medicine, vol. 291, p. 114486, 2021.

[8] C.-T. Lee, et al., "Identifying predictors of the tooth loss phenotype in a large periodontitis patient cohort using a machine learning approach," Journal of Dentistry, vol. 144, p. 104921, 2024.

[9] S. S. Bhat, et al., "A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms," Healthcare Analytics, p. 100273, 2023.

[10] R. Ahmed, M. Bibi, and S. Syed, "Improving heart disease prediction accuracy using a hybrid machine learning approach: A comparative study of SVM and KNN algorithms," International Journal of Computations, Information and Manufacturing (IJCIM), vol. 3, no. 1, pp. 49–54, 2023.

[11] I. Steinwart and A. Christmann, Support Vector Machines, Springer Science & Business Media, 2008.

[12] A. Hasuike, et al., "Machine learning in predicting tooth loss: A systematic review and risk of bias assessment," Journal of Personalized Medicine, vol. 12, no. 10, p. 1682, 2022.

[13] A. Alibrahim, et al., "Patterns and predictors of tooth loss among partially dentate individuals in Jordan: A cross-sectional study," The Saudi Dental Journal, vol. 36, no. 3, pp. 486–491, 2024.

[14] D. Stojanov, et al., "Predicting the outcome of heart failure against chronic-ischemic heart disease in elderly population–Machine learning approach based on logistic regression, case to Villa Scassi hospital Genoa, Italy," Journal of King Saud University-Science, vol. 35, no. 3, p. 102573, 2023.

[15] S. A. Pattekari and A. Parveen, "Prediction system for heart disease using Naïve Bayes," International Journal of Advanced Computer and Mathematical Sciences, vol. 3, no. 3, pp. 290–294, 2012.

[16] W. Thomson, et al., "Long-term dental visiting patterns and adult oral health," Journal of Dental Research, vol. 89, no. 3, pp. 307–311, 2010.

[17] R. Maimaitituerxun, et al., "Predictive model for identifying mild cognitive impairment in patients with type 2 diabetes mellitus: A CHAID decision tree analysis," Brain and Behavior, vol. 14, no. 3, p. e3456, 2024.

[18] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001.

[19] J. L. Speiser, et al., "A comparison of random forest variable selection methods for classification prediction modeling," Expert Systems with Applications, vol. 134, pp. 93–101, 2019.

[20] X. Zhang and C.-A. Liu, "Model averaging prediction by K-fold cross-validation," Journal of Econometrics, vol. 235, no. 1, pp. 280–301, 2023.