Student IDs 23208466
24222286

ASSIGNMENT TWO

PAPER NAME: Data Mining and Machine

Learning PAPER CODE: COMP809

TOTAL MARKS: 100

Students' Names: Vedant Marwadi, Xeniya Obolonkova

Students' IDs: 23208466, 24222286

- Due date: 09 Jun 2024 midnight NZ time.
- Late penalty: maximum late submission time is 24 hours after the due date. In this case, a **5% late penalty** will be applied.
- Submit the actual code (no screenshot) separately with appropriate comments for each task.

Note: This assignment should be complemented by a group of two students and both students MUST contribute in each part.
Submission: a soft copy needs to be submitted through the canvas assessment link.

INSTRUCTIONS:

1. **The following actions may be deemed to constitute a breach of the General Academic Regulations Part 7: Academic Discipline,**
    - Communicating with or collaborating with another person regarding the Assignment
    - Copying from any other student work for your Assignment
    - Copying from any third-party websites unless it is an open book Assignment
    - Uses any other unfair means
2. **Please email DCT.EXAM@AUT.AC.NZ if you have any technical issues with your submission on Canvas immediately**
3. **Attach your code for all the datasets in.**

Assignment 2
PART A

Vedant Marwadi
23208466
Auckland University of Technology,
nxj4679@autuni.ac.nz

Xeniya Obolonkova
24222286
Auckland University of Technology,
9236@aut.ac.nz

***Abstract*—Predicting the maintenance of teeth is essential for making well-informed decisions and effectively planning treatments within the dental field. As periodontitis Is reported to be the one of the most prevalent diseases and a leading cause of tooth loss, it is important to predict the tooth loss as well as periodontitis development and progression [1].**

Machine learning (ML) is increasingly being utilized in various sectors, including periodontology, due to its remarkable predictive capabilities. Model performance measures like accuracy, root-mean-squared error (RMSE) and AUC-ROC enable the evaluation of the model and help in identifying the need for adjustments. Factors such as model complexity, sample size, class imbalance, and prediction timeline must be carefully addressed during the development and validation of a prediction model. This paper aims to explore the diverse methodologies employed in predicting outcomes of periodontal disease, investigate problem-solving methods, algorithms, strengths and weaknesses, and address potential issues related to predictions using ML techniques.

I. INTRODUCTION

The present paper includes methodology and data explanation from six studies:

- Identifying Factors Associated with Periodontal Disease Using Machine Learning [2];
- Predicting dental caries outcomes in young adults using machine learning approach [3];
- Using machine learning algorithms to investigate factors associated with complete edentulism among older adults in the United States [4];
- Systemic Periodontal Risk Score Using an Innovative Machine Learning Strategy [5];
- Systemic Periodontal Risk Score Using an Innovative Machine Learning Strategy: An Observational Study [5];
- Association between Body Mass Index and Severity of Periodontal Disease among Adult South Indian Population [6].

All studies are concentrated to a periodonititis and related health conditions risk prediction, such as tooth loss. Using various mechanisms like Least Absolute Shrinkage and Selection Operator (LASSO) regression, Negative Binomial regression, Generalized Boosting Machines (GBM), Extreme Gradient Boosting (XGBOOST), and Adaptive Boosting (AdaBoost) techniques for developing ML models leads to excessive performance in achieving the task. Random Forest, C-Tree, and k-fold validation feature engineering techniques provide advanced feature selection capabilities. However, range of various ML approaches provide wide opportunity to improve the model parameters by adjusting them toward to reaching best possible performance [7] or use alternative techniques the produce better results. The evaluation of methodologies employed in research is grounded in the examination of diverse techniques utilized, exploring metrics, and uncovering the methodology of advancing studies through the utilization of alternative machine learning solutions.

II. PREDICTING CARIES OUTCOME USING LASSO, GBM, NEGGLM, AND XGBOOST

The study aims to use machine learning techniques to analyze a set of longitudinally-obtained predictor variables from the Iowa Fluoride Study Data to predict outcomes of dental caries in young adults and identify the most significant predictors. Data contained 51 features that were independent with 4 factors related to sociodemographic state and 47 other features. Sum of decayed, missing and filler tooth surface for age 23 was defined as a target variable D2+MFS.

Data preprocessing involved imputation of missing values using K-nearest neighbors and scaling and normalization of the data. The K-Nearest Neighbors (KNN) algorithm is ideal for categorical variables, excelling in classifying and clustering such data. It efficiently handles missing data by imputing values based on nearby data points. However, its performance can be affected by high dimensionality, causing potential inaccuracies in predictions as distances between data points become less meaningful with increased features.

For determining the correlation between target variable 23 (D2+MFS). and family income, mother's level of education, and composite SES bivariate analyses base on Mann-Whitney U tests was used [3].To explore relationship between outcome and continuous independent variables(home water fluoride concentration, total fluoride intake, and beverage intake variables Spearman (Rho) correlation tests were performedd [3]. Utilizing statistical analysis for data exploration represents a sophisticated approach to examining the interplay among variables, offering insights into the significance of each variable and enhancing the efficacy of model performance [3].

The methodology is based on utilizing the LASSO regression, generalized boosting machines (GBM) Negative binomial regression (NegGLM), Extreme gradient boosting (XGBOOST) algorithms. These models were chosen for their ability to handle high-dimensional data, perform variable selection, and work with different data types and distributions with few assumptions [3].The models were trained and tested using the nested resampling technique with 5-fold cross-validation to ensure unbiased performance estimation. Data preprocessing techniques such as k-nearest neighbor imputation, scaling, and normalization were applied before fitting the models. The TunedModel function in the MachineShop package was used to fine-tune the models by selecting optimal parameters through cross-validation.

The model evaluation in the study included assessing the performance using root mean square error (RMSE), mean absolute error (MAE), and the R-squared value. Lower RMSE and MAE values are indicators of better model performance, while a higher R-squared value suggests better performanced [3]. The best-performing model was selected based on RMSE and R2, with MAE helping to understand overall model performance. The LASSO regression model demonstrated superior performance, achieving an RMSE of 0.70, R2 of 0.44, and MAE of 0.48. The GBM and the Negative binomial GLM also displayed respectable performance, with RMSE values of 0.74 and 0.76, respectively. In contrast, the XGBOOST model exhibited the poorest performance, recording an RMSE score of 0.79 [3]. While LASSO regression demonstrates a high level of performance with its feature selection mechanism, Random Forest feature selection technique presents a more sophisticated approach to feature selection. Employing Random Forest prior to model fitting can reduce the model development time and potentially enhance performance [2].

The results of the study involved 258 participants meeting the inclusion criteria, with details on prevalence, mean values of predictor variables, and associations with caries counts at different ages. The associations between the D2+MFS23 count and various variables, such as family income, composite SES, brushing frequency, and caries experience at different ages, were statistically significant. As XGBoost showed the lowest performance from all the models, it is recommended to implement Catboost model instead as is specifically designed to handle categorical variables efficiently and shows higher performance result [8]. Unlike other gradient boosting algorithms that require one-hot encoding or label encoding for categorical variables, CatBoost automatically handles categorical variables internally. Overall, the model evaluation emphasized the performance metrics and their implications for understanding the predictive capabilities of the models used in this study The incorporation of Mann-Whitney U tests and KNN algorithms had a beneficial impact on the model's performance. However, the adoption of more sophisticated machine learning methodologies such as CatBoost may enhance the handling of categorical variables in a more optimal manner.

III. Feature Selection based on CTree and Random Forest algorithms.

A study with 4699 participants investigated the relationship between chronic conditions in PD and non-PD groups. Variable selection for PD prediction was done using CTree and Random Forest regression models. Approximately 30% of the data lacked health insurance and other values, with a missing category created for values missing over 1%. The severity of PD was based on clinical attachment loss (CAL), categorizing PD into mild, moderate, and severe based on CAL measurements. The research examined the common variations in chronic conditions using a conditional inference regression model (CTree). This model was structured with a limit of 100 items for parent and child nodes, and a p-value of 0.001. Using the CTree model, the prediction rate for PD was found to be better based on sociodemographic and behavioral variables than chronic conditions. To predict moderate/severe PD, age and level of education were identified as important factors. Bias was discovered for elderly people as the data from 2013 to 2014 did not include individuals with edentulous conditions. This may explain the lower level of PD among older individuals. Information related to education revealed that 45% of individuals had a low level of education, which may influence their access to healthcare.

In order to answer the main research query, CTree analysis was limited to chronic conditions only for the presence and severity of PD where individuals with PD predominantly lacked hypertension, arthritis, and diabetes, resulting in over 80% of participants reporting a combination of these conditions with

PD. Among individuals with diabetes, the highest percentage of moderate/severe PD (over 80%) was noted. On the other hand, those without diabetes but with hypertension and asthma showed the lowest percentage of moderate/severe PD (around 60%). The prevalence of moderate/severe PD varied from sixty to just above eighty percent among participants with varying combinations of chronic conditions. The prevalence of moderate/severe PD varied from sixty to just above eighty percent among participants with varying combinations of chronic conditions [2].

Random Forest was also employed as a secondary model for three most critical variables for PD are age, alcohol use, and health insurance. When focusing on moderate/severe PD, the top six significant variables are age, education level, type of health insurance, the ratio of family income to poverty, gender, and race. These variables are determined by considering chronic conditions, sociodemographic factors, and behavioral variables. The consistent presence of these variables in the CTree models confirmed the validity of the models and underscores the significance of sociodemographic and behavioral factors [2].

For presenting PD and moderate/severe PD, logistic regression analysis was applied. According to the Random Forest analysis, the top three most vital variables for PD and moderate/severe PD, based solely on chronic conditions, are hypertension, arthritis, and diabetes [2]. Specifically for moderate/severe PD, the most crucial variables are diabetes, hypertension, and asthma.

Chronic conditions and sociodemographic/behavioral factors play crucial roles in Periodontal Disease (PD). Sociodemographic factors have a broader impact than serious chronic conditions. The study suggests that focusing on sociodemographic factors can enhance PD prediction and prevent its progression in middle-aged individuals. In the study, CTree can capture complex relationships and create a decisive model to identify key factor combinations linked to PD. In contrast, Random Forest assesses a data subset through bootstrapping to rank the most important variable for the desired outcome. While Random Forest may not pinpoint the most frequent combination of variables associated with PD, it can ascertain whether the top predictors identified align with the critical variables in CTree models. Thus, both machine learning techniques can automatically detect the interaction and nonlinear correlation of the variables. The approach of feature selection based on the CTree and Random Forest methods was found to be suitable for PD risk prediction; however, as the dataset included a small number of records with reported chronic diseases, the data for regression modeling may be imbalanced, affecting the model's accuracy.

IV. Using AdaBoost to investigate factors associated with complete edentulism among older adults

The study utilized data from the Behavioral Risk Factor Surveillance System (BRFSS) netrwork. The data contained 401,958 records of health-related information for adult residents with a focus on respondents aged 65 and above. The study included 30 explanatory variables covering various domains, including oral health variables like time since last dental visit and number of permanent teeth lost [4].

Data was cleaned in R, removing missing observations. Data was standardized with MinMaxScaler from sklearn. Feature selection reduced dimensionality. Machine learning algorithms trained models, and hyperparameters were optimized with five-fold cross-validation. Demographic characteristics were studied among edentulous and dentate groups to explore edentulism prevalence across age groups and factors like health status and healthcare access.

The study utilized a range of machine learning algorithms, including Naive Bayes, K-Nearest Neighbors, Random Forest, AdaBoost, Logistic Regression, Ensemble, and Gradient Boosting. These algorithms were employed to identify factors associated with complete edentulism among older adults. The research incorporated 30 explanatory variables covering sociodemographic factors, health care access, health behavior, health status, chronic health conditions, and disability variables. Specific oral health variables included "time since the last dental visit" and "number of permanent teeth lost." The outcome variable was deduced from responses to the query regarding the state of dentition, with the reference category being "completely edentulous/loss of all permanent teeth" [4]. The features were prioritized according to their correlation with the outcome variable (edentulism). The list of machine learning algorithms employed in this research across various feature ranking techniques included Naive Bayes, K-Nearest Neighbors, Random Forest, AdaBoost, Logistic Regression, Ensemble, and Gradient Boosting.

The Adaptive Boosting Machine Learning Algorithm achieved the highest accuracy with an AUC of 84.9%, whereas k-NN exhibited the lowest accuracy at 73.8%. Among the variables assessing healthcare access in the study, the last dental visit emerged as the most significant factor linked to complete edentulism in our top-performing model. This trend remained consistent across all other models as well [4]. AdaBoosting's iterative approach of adjusting weights for misclassified instances can lead to higher accuracy in predicting factors associated with complete edentulism in older adults. Moreover, this algorithm demonstrates relevant robustness to overfittings [9], conducts intricate feature selection to accurately identify the most influential factors contributing to complete tooth loss in elderly individuals and show versatility to various types of data. Nevertheless, AdaBoosting may be sensitive to noisy data, which could impact the model's performance in predicting factors associated with complete edentulism if the dataset contains outliers or inaccuracies [10].

For further model effectiveness assessing it is recommended to extend the list of model evaluation techniques. Some of these methods including hyperparameter that can enhance the model's performance by finding the best combination of parameters that improve accuracy and reduce overfitting through techniques like grid search or random search. Analyzing the confusion matrix and calculating metrics such as accuracy, precision, recall, and F1 score can give a more detailed understanding of the model's performance in terms of true positives, true negatives, false positives, and false negatives.These metrics can help evaluate the model's ability to correctly predict outcomes.

Random Forest, Gradient Boosting, or XGBoost can enhance the model's predictive power by combining multiple models and leveraging their strengths to improve overall performance. Random Forest algorithm is recommended for feature selection in this study due to its robust nature as an ensemble learning technique. It can offer valuable insights into the importance of variables for predicting the outcome of interest [2].

V. Implementing RuleFit algorithms to identify potential significant features associated with tooth loss

Data extraction was performed using a Structured Query Language (SQL) query to identify patients meeting the selection criteria and having documented periodontitis diagnosis from electronic health records (EHRs). The primary outcome considered in the study was the number of tooth loss occurrences following the initial visit, excluding third molars from analysis. Eighteen initial variables were scrutinized during the initial comprehensive periodontal examination, including age, sex, medical history, lifestyle habits, and specific periodontal indicators. These variables, identified as potentially relevant from existing literature, were reliable and comprehensive in the EHRs [1]. The study employed a two-stage predictive modeling approach to forecast tooth loss. In the first stage, the Rule-Fit algorithm facilitated feature selection and rule generation to determine pivotal predictors of tooth loss. This algorithm enabled the extraction of high-order feature combinations to enhance machine learning model performance by identifying complex variable interactions. Subsequently, count regression models were utilized in the second stage to predict tooth loss, considering demographic and clinical variables along with the follow-up time component to account for temporal effects on tooth loss [1].

To refine the model and enhance interpretability, the LASSO method was applied to the rule-set generated by Rule-Fit. LASSO allowed for variable selection, regularization, and model sparsity, minimizing overfitting and improving model accuracy and generalization. Post-LASSO selection, importance scores of rules were calculated to identify and retain the most significant rule per feature, mitigating redundancy and enhancing model interpretability without compromising prediction accuracy [1].

The hybrid approach of Rule-Fit for feature selection and LASSO for regularization and interpretation improved the model's accuracy and robustness while maintaining transparency in the model's predictive capabilities. This methodology ensured a comprehensive understanding of the factors contributing to tooth loss in the studied cohort and provided an insightful approach to predictive modeling in dental research [1]. Further research may explore fine-tuning the rule-selection process to account for threshold variations and enhance model stability, thereby advancing predictive accuracy and interpretability. To ensure robust model evaluation, a separate test set was utilized to validate the performance of the predictive model. By evaluating the model on unseen data, the study aimed to gauge its effectiveness in generalizing to new patient cases and estimating tooth loss counts accurately. The test set allowed for unbiased evaluation of the model's performance, providing insights into its real-world applicability and predictive power beyond the training data [1]. Furthermore, the identified significant rules generated by the Rule-Fit algorithm – whether individual variables or complex feature interactions – were integrated into the count regression model for prediction purposes. The contribution of these rules to the model's predictive accuracy and interpretability was assessed through p-values and 95% confidence intervals of high-order features. This analysis reaffirmed the significance of the extracted rules in predicting tooth loss outcomes and provided a statistical framework for validating the Rule-Fit model's results. Overall, the model evaluation process encompassed comprehensive validation steps, including RMSE calculation on the test set and statistical analysis of significant rules, to ensure the reliability, accuracy, and interpretability of the predictive model in forecasting tooth loss counts for patients with periodontitis. For model evaluation additional metric such as MAE and R2 could be implemented to ensure all aspect of the model performance estimation [3].

VI. Comprehensive Approach with Feature Selection and Explainability Analysis based on MLP model

The study aimed to predict periodontal health using various analysis strategies. For data preprocessing encoding binary and ordinal variables was implemented. Various techniques that were applied sequentially in the machine learning process included encoding variables, feature selection, data scaling, and utilizing a multilayer perceptron model. BorutaPy was employed for feature selection to identify significant variables for the prediction model.

The subjects' profiles were visualized using UMAP and clustered using the DBSCAN algorithm based on selected variables. A multilayer perceptron algorithm was optimized for performance using scikit-optimize. The research performed on population of 532 subjects showed variations in periodontal health scores by age groups. Three distinct clusters were identified based on sociodemographic characteristics and risk factors. BorutaPy highlighted relevant variables for the final model, such as age, BMI, systemic pathologies, and lifestyle habits. Random Forest BorutPy, a feature selection algorithm that combines Random Forest with the Boruta method, is a recommended approach for identifying important features in a dataset by evaluating attribute importance scores.

In the machine learning pipeline, a multilayer perceptron model was employed after min-max scaling. The model was optimized using training data with weighted F1-scores of 0.60 ± 0.03 for training and 0.57 ± 0.08 for validation datasets [5]. While the model performed well in predicting healthy periodontium and periodontitis, it struggled with accurate prediction of gingival inflammation. The evaluation metrics and ROC curve illustrated the model's specificity and sensitivity for each prediction group. The "kernelSHAP method" was utilized for interpreting predictions, highlighting important attributes for periodontal health prediction. Age, systemic pathologies, hormonal status, dietary habits, and sugary drink consumption were identified as significant variables. The model indicated that age, diet, smoking, and pathologies increased the risk of periodontitis, while variables like education did not show a strong impact. The analysis of SHAP values highlighted correlations between variables like age, BMI, diet, and hormonal status in predicting periodontitis. Individual-level risk predictions demonstrated how specific factors influenced the likelihood of periodontitis diagnosis for different age groups and genders. The methods used offer benefits like interpreting model predictions and identifying key predictors for periodontal health. However, the model faces challenges in accurately predicting gingival inflammation and interpreting variable interactions. To enhance the model's performance in predicting gingival inflammation, reevaluating feature selection or exploring specialized models may be beneficial. Additionally, it is considered to collect more data or balancing the dataset to improve the accuracy and generalizability of the model's predictions. Exploring ensemble methods such Random Forest or hybrid models that combine different ML techniques to leverage the strengths of each approach and enhance overall prediction performance [2]. Moreover, to demonstrate the innovativeness of Multilayer Perceptron (MLP), it is beneficial to compare its results with other advanced techniques such as LASSO, XGBoost, and Decision Trees.

VII. Implementation of Multivariate Logistic Regression for investigating association between Body Mass Index and Severity of Periodontal Disease

The study aims to investigate the relationship between BMI and periodontitis, exploring the impact of poor diet, nutrient deficiency, and high sugar/fat intake on periodontal risk. It also examines the inflammatory role of adipose tissue in obesity and its link to chronic periodontal disease. By analyzing BMI's association with pocket probing depth, the study aims to understand the complex connection between obesity and periodontitis in adults, while considering covariates and confounders.

Bivariate analyses and Multivariate logistic regression (MLR) models have been applied. Bivariate analyses is conducted to determine whether a statistical association exists between any of two variables. MLR model is an example of a broad class of models known as generalized linear models (GLM). Covariates included sociodemographic factors (age, gender, community status, income, marital status, education, occupation) and lifestyle characteristics (smoking, alcohol consumption, fluoridated products use, dental visits frequency, salt intake, physical inactivity). Periodontal disease was classified as mild, moderate, or severe. Previous research recommended a minimum sample size of 26 in each group for 95% confidence and 80% power. To accommodate additional factors and account for attrition, the sample size was increased to 180. The associated factors populations were carried forward to a multivariate model. Multivariate logistic regression models have been used to estimate odd ratios (OD) and associated 95% confidence intervals(CIs). A P-value of `<0.05` was set as statistically significant. The stepwise method was used for regression. The study is credible as it aligns with previous research on the prevalence of periodontal disease. However, it did not find a significant association between obesity and periodontal disease, indicating that smoking may be the primary factor in this population. This contrasts with other studies showing a strong link between obesity and periodontitis, highlighting the complexity of factors like smoking, diabetes, aging, and alcohol consumption. Previously, it was demonstrated that periodontitis and obesity exist simultaneously, and their association is bidirectional. Therefore, the study utilized a cross-sectional approach, encompassing all potential covariates at risk for both obesity and periodontal disease. This strengthens the study findings. The final assessment did not investigate the temporal causation of periodontal disease with obesity but focused on identifying various risk factors associated with these chronic diseases in the adult population. This approach strengthens the study's conclusions by emphasizing the diverse risk factors linked to these chronic diseases in adults instead of establishing a causal relationship between periodontal disease and obesity over time.

Population's selection bias, information bias and Berksonian bias were proved to be eliminated. The study observed that among smokers, the prevalence of periodontal disease was 50.0% (n = 53). After adjusting for confounders, multivariate analysis revealed that smokers had a 3.24 times higher risk compared to nonsmokers. Within the limitations of the study, obesity is not statistically associated with periodontal disease parameters in the urban and tribal regions, but other covariates, such as smoking, alcohol consumption, age, and history of diabetic mellitus, are strongly associated with periodontal disease. The MLR model limitation are multicollinearity issues, which were not checked in the observer research and could alter the conclusions. Additionally, incorporating excessive independent variables in a regression model can result in overfitting, causing the model to capture the noise in the data rather than the true underlying relationship [11]. Multivariable regression assumes linear relationships between the independent variables and the dependent variable, otherwise it may not accurately capture the underlying patterns in the data.

VIII. Conclusion

The studies conducted on periodontal health and its association with various factors utilized advanced analysis techniques to predict and understand the complexities of periodontal diseases in adult populations. While different studies focused on factors such as obesity, smoking, alcohol consumption, age, and diabetes, they underscored the importance of considering multiple variables in assessing periodontal health. The use of multivariate logistic regression and machine learning models provided valuable insights into the associations between these factors and periodontal diseases. Recommendations for future research include improving model performance through reevaluating feature selection and exploring specialized models to enhance predictive accuracy and understanding the diverse risk factors associated with chronic diseases in adult populations. There is no perfect algorithm that provides a fully adjusted solution to achieve absolute results from a probabilistic perspective [7]. Therefore, there is always room for improving model performance by adjusting parameters to achieve the best possible outcomes. Models like ADABoosting and LASSO Regression have shown impressive results in predicting periodontitis, signaling an opportunity to enhance performance and develop models capable of making predictions across various data scopes. Adjusting parameters is essential to reaching optimal performance levels.

References

[1] C.-T. Lee, K. Zhang, W. Li, K. Tang, Y. Ling, M. Walji, and X. Jiang, "Identifying predictors of tooth loss using a rule-based machine learning approach: A retrospective study at university-setting clinics," *Journal of periodontology*, vol. 94, 04 2023.

[2] H. M. Alqahtani, S. M. Koroukian, K. Stange, N. K. Schiltz, and N. F. Bissada, "Identifying factors associated with periodontal disease using machine learning," *Journal of International Society of Preventive and Community Dentistry*, vol. 12, no. 6, pp. 612–622, Nov–Dec 2022, https://doi.org/10.4103/jispcd.JISPCD_188_22.

[3] C. Ogwo, G. Brown, J. Warren, D. Caplan, and S. Levy, "Predicting dental caries outcomes in young adults using machine learning approach," *BMC Oral Health*, vol. 24, no. 1, p. 529, 2024. [Online]. Available: https://doi.org/10.1186/s12903-024-04294-7

[4] A. M. Oladayo, H. W. Miyuraj Harishchandra, E. Zeng, D. J. Caplan, A. Butali, and L. Marchini, "Using machine learning algorithms to investigate factors associated with complete edentulism among older adults in the united states," *Special Care in Dentistry*, vol. 44, no. 1, pp. 148–156, 2024, https://doi.org/10.1111/scd.12832.

[5] P. Monsarrat, D. Bernard, M. Marty, C. Cecchin-Albertoni, E. Doumard, L. Gez, J. Aligon, J.-N. Vergnes, L. Casteilla, and P. Kemoun, "Systemic periodontal risk score using an innovative machine learning strategy: An observational study," *Journal of Personalized Medicine*, vol. 12, no. 2, 2022. [Online]. Available: https://www.mdpi.com/2075-4426/12/2/217

[6] M. Venkat and C. Janakiram, "Association between body mass index and severity of periodontal disease among adult south indian population: A cross-sectional study," *Indian Journal of Community Medicine*, vol. 48, no. 6, pp. 902–908, Nov–Dec 2023, https://doi.org/10.4103/ijcm.ijcm_148_22.

[7] N. Z. Bashir, Z. Rahman, and S. L.-S. Chen, "Systematic comparison of machine learning algorithms to develop and validate predictive models for periodontitis," *Journal of Clinical Periodontology*, vol. 49, no. 10, pp. 958–969, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jcpe.13692

[8] J. M. H. Pinheiro and M. Becker, "Breast cancer classification using gradient boosting algorithms focusing on reducing the false negative and shap for explainability," Mar 2024, https://jhttps://arxiv.org/pdf/2403.09548.

[9] B. Schlkopf, Z. Luo, and V. Vovk, "Empirical inference: Festschrift in honor of vladimir n. vapnik," 2014, url=https://doi.org/10.1007/978-3-642-41136-6.

[10] A. Kalai and R. A. Servedio, "Boosting in the presence of noise," in *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing*, ser. STOC '03. New York, NY, USA: Association for Computing Machinery, 2003, p. 195–205. [Online]. Available: https://doi.org/10.1145/780542.780573

[11] A. MAXWELL, "Limitations of the use of the multiple linear regression model," *British Journal of Mathematical and Statistical Psychology*, vol. 28, pp. 51 – 62, 08 2011, https://doi.org/10.1111/j.2044-8317.1975.tb00547.x.