Pertanika J. Sci. & Technol. 32 (4): 1753 - 1764 (2024)

PERTANIKA JOURNALS

SCIENCE & TECHNOLOGY

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

Optimizing Bagged Trees in an Ensemble Classifier for Improved Prediction of Diabetes Prevalence in Women

Jose Candia Jr.*, Airish Mae Adonis and Jesica Perlas

Department of Computer Science, North Eastern Mindanao State University – Tagbina Campus, Poblacion, Tagbina, 8308 Surigao del Sur, Philippines

ABSTRACT

This study aims to optimize the performance of the bagged tree in an ensemble classifier for predicting diabetes prevalence in women. The study used a dataset of 1,888 women with six features: age, BMI, glucose level, insulin level, blood pressure, and pregnancy status. The dataset was divided into training and testing sets with a 70:30 ratio. The bagged tree ensemble classifier was used for the analysis, and five-fold cross-validation was applied. The study found that using all features during training resulted in a 92.3% training accuracy and a 99.5% testing accuracy. However, applying optimization techniques such as feature selection, parameter tuning, and a maximum number of splits improved model performance. Feature selection optimized the accuracy performance by 0.2%, while parameter tuning improved the test accuracy by 0.2%. Moreover, decreasing the maximum number of splits from 1322 to 800 or 600 resulted in an optimized model with 0.1% higher validation accuracy. Finally, the optimized bagged tree models were evaluated using various performance metrics, including accuracy, precision, recall, and F1 score. The study found that Model 1, which used 800 maximum number of splits and 50 learners, outperformed Model 2 in terms of recall and F1 score, while Model 2, which used 600 maximum number of splits and 50 learners, had a higher precision score. The study concludes that optimization techniques can significantly improve the performance of the bagged tree in predicting diabetes prevalence in women.

Keywords: Bagged trees, diabetes prevalence, ensemble classifier, feature selection, model optimization, parameter tuning

ARTICLE INFO

Article history:
Received: 07 March 2023
Accepted: 12 September 2023
Published: 22 July 2024

DOI: https://doi.org/10.47836/pjst.32.3.16

E-mail addresses:
jqcandia@nemsu.edu.ph (Jose Candia Jr.)
airishadonis09@gmail.com (Airish Mae E. Adonis)
perlasjesica09@gmail.com (Jesica Perlas)
*Corresponding author

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

INTRODUCTION

Diabetes is a major public health problem worldwide, affecting millions of people and imposing a significant economic burden on healthcare systems. In the Philippines, diabetes is a growing concern, with a prevalence rate of 7.5% among women aged 20–79 years. The high prevalence of diabetes in the Philippines can be attributed to various factors, including sedentary lifestyles, unhealthy diets, and genetic predisposition. Therefore, an urgent need is to develop accurate and reliable methods for predicting diabetes prevalence in women to prevent complications and improve health outcomes.

Machine learning algorithms have shown promise in predicting diabetes prevalence in women (Nishat et al., 2021). Decision trees and ensemble classifiers are popular techniques for classification and prediction tasks because of their simplicity, interpretability, and ability to handle both categorical and numerical data. Ensemble classifiers, in particular, are known to improve the accuracy and robustness of the model by combining the predictions of multiple decision trees. Bagged trees are an ensemble classifier that has gained popularity in recent years. Bagging is a resampling method that involves randomly selecting subsets of the training data and training decision trees on each subset. The predictions of the individual trees are then combined through a voting or averaging scheme to produce a final prediction. Bagged trees have improved the accuracy and stability of decision tree models, particularly for complex and noisy datasets.

Several studies have used machine learning algorithms to predict diabetes prevalence in the Philippines. For instance, Tan et al. (2019) used decision trees and logistic regression to predict the risk of type 2 diabetes among Filipino adults. The study found that the most important predictors of diabetes were age, body mass index (BMI), waist circumference, and family history of diabetes. Similarly, a study by Abayadeera et al. (2019) used decision trees and logistic regression to predict the risk of type 2 diabetes among urban Filipinos. The study found that the most important predictors of diabetes were age, BMI, physical activity, and education level. However, few studies have used bagged trees to predict diabetes prevalence in the Philippines. One study by Pang et al. (2017) used bagged trees to predict the risk of diabetes among Chinese adults, achieving an accuracy of 78.4%. Another study by Jia et al. (2018) used bagged trees to predict the risk of diabetes among urban Chinese residents, achieving an accuracy of 84.2%. These studies demonstrate the potential of bagged trees in predicting diabetes prevalence and risk factors.

This study aims to optimize the bagged tree ensemble classifier for predicting diabetes prevalence in women in the Philippines. We will use data from the 2019 Philippine National Nutrition Survey, a nationally representative survey that collects information on Filipinos' health and nutrition status. We will focus on six key predictors of diabetes prevalence in women: pregnancies, glucose levels, blood pressure, BMI, age, and diabetes pedigree function. By optimizing the parameters of the bagged tree algorithm, we aim to improve

the accuracy and robustness of the model. The results of this study could have important implications for the early detection and prevention of diabetes in women, as well as for the development of more effective machine-learning algorithms for predicting chronic diseases in resource-limited settings.

Conceptual Framework

The conceptual framework for this study on optimizing bagged trees in an ensemble classifier for improved prediction of diabetes prevalence in women is presented in the diagram below:

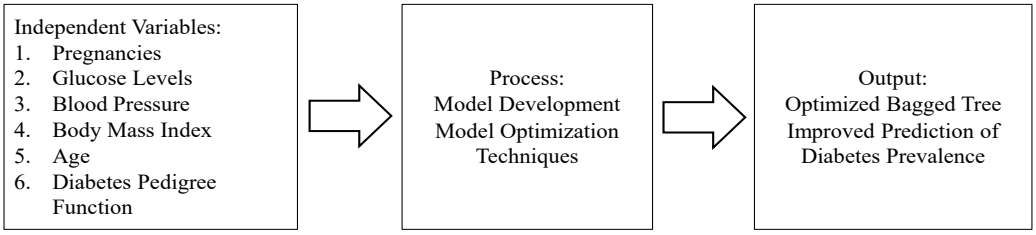

Figure 1. Conceptual framework of optimizing bagged trees in an ensemble classifier for improved prediction of diabetes prevalence in women

The independent variables in this study include the predictive factors of diabetes prevalence in women: pregnancies, glucose, blood pressure, BMI, age, and diabetes pedigree function. These variables will be used as input features for the bagged trees in the ensemble classifier. The dependent variable is the prediction of diabetes prevalence in women.

The bagged trees algorithm will be used as the main classifier for predicting diabetes prevalence in women. The bagged trees algorithm is an ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce the risk of overfitting. The bagged trees algorithm will be optimized using various parameters such as the number of trees, maximum depth, and minimum number of samples required to split a node.

The study will also employ feature selection techniques to determine the most important predictive factors for diabetes prevalence in women. The feature selection process will be performed using wrapper methods, specifically the recursive feature elimination algorithm. The recursive feature elimination algorithm is a backward selection method that recursively removes features from the model and evaluates their impact on model performance. The output of the bagged trees algorithm will be evaluated using performance metrics such as accuracy, precision, recall, and F1 score.

MATERIALS AND METHODS

This study used ensemble classifier model development and optimization techniques to predict diabetes prevalence in women. The study was conducted in the following phases:

Phase 1: Data Collection

The dataset used for this study was obtained from the Philippine Statistics Authority's National Nutrition Survey 2019 (Philippine Statistics Authority 2020), which includes information on women diagnosed with diabetes. The dataset will be preprocessed to remove missing values and normalize the data.

Table 1
Factors affecting maternal risk during pregnancy

Factors	Descriptions
Pregnancies	Number of pregnancies a woman has had
Glucose Level	Blood glucose levels in terms of molar concentration, mmo/L
Blood Pressure	Lower value of blood pressure in mmHg
Body Mass Index	Measure of body fat based on a person's height and weight
Diabetes Pedigree Function	Measure of diabetes family history
Age	Age in years

The collected data was divided into two sets: a training set comprising 70% of the data and a testing set comprising 30%. The training set was utilized to construct and train a machine learning model, while the testing set was used to assess the model's performance based on previously unseen data.

Table 2 presents the data distribution between the training and testing datasets for predicting diabetes prevalence in women. The training dataset comprises 1,323 data points, while the testing dataset contains 565.

Table 2
Training and testing datasets of diabetes prevalence in women

Diabetes Prevalence	Training 70%	Testing 30%	Total
Yes	457	182	639
No	866	383	1,249
Total	1,323	565	1,888

Phase 2: Feature Selection

In this phase, a wrapper method using the Recursive Feature Elimination algorithm will be utilized to select the most important features to predict diabetes prevalence in women. Recursive Feature Elimination works by removing one feature at a time and evaluating the impact of the feature on the model's performance until the best subset of features is determined. This method aims to

increase the model's accuracy and reduce overfitting by removing irrelevant features. The selected features will train the model in the next phase.

Phase 3: Parameter Tuning

The Bagged Trees algorithm will be used as the primary classifier for predicting diabetes prevalence in women. In this phase, the Bagged Trees algorithm will be optimized by varying its parameters, such as the maximum number of splits and learners. This phase aims to improve the model's performance by determining the parameters' optimal values. Parameter tuning is crucial as it can significantly impact the model's accuracy and generalization ability.

Phase 4: Performance Evaluation

In this phase, the performance of the Bagged Trees algorithm will be evaluated using performance metrics such as accuracy, precision, recall, and F1 score. The evaluation will involve splitting the dataset into training and testing sets using a 5-fold cross-validation approach. Since the dataset had skewed data with many more observations of one class, a stratified 5-fold cross-validation was applied. The metrics will be used to evaluate the model's ability to predict diabetes prevalence in women accurately. Accuracy is the proportion of true predictions out of all predictions made, while precision is the proportion of true positive predictions out of all positive predictions made. Recall is the proportion of true positive predictions out of all actual positive cases, and the F1 score is the harmonic mean of precision and recall. The higher the values of these metrics, the better the model's performance.

The study will use MATLAB software, specifically the Statistics and Machine Learning Toolbox. MATLAB is a powerful programming language and software environment widely used for scientific computing and data analysis. The bagged tree ensemble classifier will be optimized and trained on the diabetes dataset using this toolbox. The resulting model will then be used to predict diabetes prevalence in women in the Philippines, which could have important implications for early detection and prevention of diabetes in this population.

RESULTS

The results and discussion present the study's findings on optimizing bagged trees in an ensemble classifier for improved prediction of diabetes prevalence in women. It highlights the performance of the optimized model in predicting diabetes prevalence based on the selected features and tuned parameters. It also discusses the implications of the study’s findings for the early detection and prevention of diabetes in women in the Philippines and for developing more effective machine-learning algorithms for predicting chronic

diseases in resource-limited settings. Overall, the results and discussion aim to provide a comprehensive understanding of the effectiveness of the optimized bagged trees model in predicting diabetes prevalence in women and its potential impact on public health.

Table 3
Ensemble classifier model summary

Model Type	Ensemble Method	Learner Type	Validation Method	Training Accuracy	Testing Accuracy
Bagged Tree	Bag	Decision Tree	5-fold Cross Validation	92.3%	99.5%
Bagged Tree	Bag	Decision Tree	Stratified 5-fold Cross Validation	81.4%	82.5%

The primary goal of this research is to construct a robust machine-learning model capable of accurately predicting the prevalence of diabetes among women in the Philippines. To fulfill this objective, we employed an ensemble classifier model trained on the entire feature set and evaluated using standard and stratified 5-fold cross-validation techniques. An overview of the model's performance can be found in Table 3.

The outcomes reveal that our bagged tree ensemble classifier model demonstrated a remarkable performance, achieving a notable accuracy of 92.3% during training and an impressive 99.5% accuracy during testing. When addressing the imbalanced data issue through stratified cross-validation, the accuracy rates for training and testing slightly decreased to 81.4% and 82.5%, respectively.

Table 4
Bagged tree model optimization using feature selection

Bagged Tree Model	Feature Selection	Validation Accuracy	Test Accuracy	Remarks
1	All features were used in the model (6/6)	81.4%	82.5%	Initial Model
2	All features used except: Pregnancy (5/6)	81.5%	81.9%	Unoptimized
3	All features used except: Blood Pressure (5/6)	81.0%	82.5%	Unoptimized
4	All features used except: Pregnancy and Blood Pressure (4/6)	81.4%	81.1%	Unoptimized
5	All features used except: Blood Pressure and Age (4/6)	79.5%	80.9%	Unoptimized

The findings from the experiments provide insights into the relationship between feature selection and the performance of the bagged tree model in predicting the likelihood of diabetes prevalence. Here is a summary of the observations:

Model 1 (Initial Model): The model trained with all features achieved a validation accuracy of 81.4% and a test accuracy of 82.5%. It serves as the baseline for comparison. Removing the "Pregnancy" feature (Model 2) while retaining the rest resulted in a slightly decreased validation accuracy of 81.5% and a test accuracy of 81.9%. It indicates a minor reduction in model performance, suggesting that "Pregnancy" might play a role in accurate predictions. Excluding the "Blood Pressure" feature (Model 3) while keeping others led to a validation accuracy of 81.0% and a test accuracy of 82.5%, consistent with the initial model's performance. Removing both "Pregnancy" and "Blood Pressure" features (Model 4) caused a validation accuracy of 81.4% and a test accuracy of 81.1%. The performance drop suggests that both features somewhat contribute to model accuracy. Omitting "Blood Pressure" and "Age" features (Model 5) resulted in a validation accuracy of 79.5% and a test accuracy of 80.9%, representing the least optimized configuration among the tested scenarios.

Table 5
Optimization of bagged tree model through varying learners

Bagged Tree Model	Number of Learners	Validation Accuracy	Test Accuracy	Remarks
1	30	81.4%	82.5%	Initial Model
2	40	81.5%	82.5%	Unoptimized
3	50	82.5%	81.6%	Unoptimized
4	60	82.2%	82.3%	Unoptimized
5	70	82.2%	82.8%	Optimized

The optimization of the developed model was also carried out using parameter tuning. Table 5 summarizes the results of the bagged tree model optimization by varying the number of learner parameters, which determines the number of decision trees in the ensemble. The optimization technique increased training and test accuracy by 0.8% and 0.3%, respectively, when the number of learners was increased from 30 to 70. The improvement in accuracy suggests that the increase in the number of learners resulted in a more robust and accurate model.

The study also investigated the impact of varying the maximum number of splits in optimizing the bagged tree model's performance. Table 6 summarizes the results of the experiment. It shows that increasing the maximum number of splits from 20 to greater than 100 improved the validation accuracy by at least 9.3% and test accuracy by 15.4%.

This result suggests that setting a higher value for maximum splits can prevent the model from overfitting and improve its generalization performance.

Table 6
Optimization of bagged tree model through varying splits

Bagged Tree Model	Maximum Number of Splits	Validation Accuracy	Test Accuracy	Remarks
1	20	82.2%	82.8%	Initial Model
2	50	88.2%	91.0%	Optimized
3	100	91.5%	98.2%	Optimized
4	200	91.9%	99.5%	Optimized
5	300	91.8%	99.8%	Most Optimized
6	500	92.3%	99.6%	Most Optimized

Table 7
Performance metrics of the optimized bagged tree model

Metrics	Model 1 Splits: 300 Learner: 70		Model 2 Splits: 500 Learner: 70	
	Validation	Test	Validation	Test
Accuracy	91.8%	99.8%	92.3%	99.6%
Precision	92.7%	99.7%	93.4%	99.5%
Sensitivity (Recall)	94.7%	100%	94.7%	100%
Specificity	86.7%	99.5%	87.8%	98.9%
F1 Score	93.7%	99.9%	94.1%	99.7%

The performance metrics of the optimized Bagged Tree model, evaluated on both the validation and test datasets, are presented in Table 7. The model was tested using Model 1 with 300 splits and Model 2 with 500 splits while maintaining a constant learner count of 70.

Model 1 achieved an accuracy of 91.8% on the validation set and an impressive 99.8% on the test set. Model 2 exhibited similar performance, with an accuracy of 92.3% on the validation set and a slightly lower accuracy of 99.6% on the test set.

The precision values for both models were consistently high as well. For Model 1, precision was recorded at 92.7% on the validation set and an exceptional 99.7% on the test set. Model 2 displayed slightly improved precision, with values of 93.4% on the validation set and 99.5% on the test set.

Sensitivity (recall), which measures the model's ability to correctly identify positive instances, was remarkably high for both models. Model 1 exhibited a sensitivity of 94.7%

on both the validation and test sets, indicating its strong capability to identify positive cases. Similarly, Model 2 achieved a perfect sensitivity of 100% on both datasets.

Regarding specificity, which reflects the model's ability to correctly identify negative instances, both models displayed respectable values. Model 1 recorded a specificity of 86.7% on the validation set and a noteworthy 99.5% on the test set. Model 2 showed a slight increase in specificity, with values of 87.8% on the validation set and 98.9% on the test set.

The F1 score, which balances precision and recall, showcased consistent and high values for both models. Model 1 attained an F1 score of 93.7% on the validation set and an impressive 99.9% on the test set. Model 2's F1 score was slightly higher, with values of 94.1% on the validation set and 99.7% on the test set.

DISCUSSION

The study aimed to optimize the bagged tree ensemble classifier model to improve the prediction of diabetes prevalence in women. Through the analysis of the four findings, several important insights were uncovered.

First, the high accuracy results collectively indicate a performance exhibited by the model, showcasing its ability to provide accurate predictions regarding the likelihood of diabetes prevalence among women. It lends substantial support to the practical applicability of machine learning algorithms, particularly emphasizing the efficacy of the bagged tree ensemble classifier in forecasting diabetes prevalence among women in resource-constrained settings such as the Philippines, which is consistent with previous studies that used machine learning algorithms for diabetes prediction. For instance, a study by Mujumdar and Vaidehi (2019) used a machine learning approach based on decision trees and random forest algorithms to predict diabetes and achieved an accuracy of 91.78%. Another study by Zhao et al. (2019) applied a support vector machine (SVM) algorithm to predict diabetes and reported an accuracy of 95.03%. These studies demonstrate the potential of machine learning algorithms in accurately predicting diabetes prevalence.

Second, while prior studies have consistently highlighted the enhanced performance of machine learning models through the application of feature selection techniques (Nguyen et al., 2020; Zhang et al., 2018), the current investigation presents a unique scenario where each predictor included in the model appears to bear significant relevance to the predictive outcomes.

Third, the finding is consistent with the literature on ensemble methods, which suggests that increasing the number of trees in an ensemble generally improves the model's accuracy up to a certain point (Biau, 2012). Additionally, other studies on parameter tuning in bagged tree models have also shown that increasing the number of trees can lead to improved performance (Chen et al., 2004; Wang et al., 2018). However, it is important to note that increasing the number of learners beyond a certain point can also lead to overfitting,

where the model becomes too complex and starts to memorize the training data rather than generalize it to new data. Therefore, finding the optimal number of learners that maximizes the model's accuracy without overfitting is important.

Fourth, the finding is consistent with the related literature, which showed that increasing the maximum number of splits can reduce overfitting and improve the performance of decision tree-based algorithms (Breiman et al., 1984; Quinlan, 1993). In a study by Han and Kamber (2001), the authors demonstrated that maximizing the depth of decision trees in a bagged ensemble model can improve its generalization performance. The finding highlights the importance of fine-tuning the hyperparameters of machine learning models to optimize their performance.

Overall, the findings of this study provide valuable insights into optimizing the bagged tree ensemble classifier model for predicting diabetes prevalence in women. The study highlights the importance of feature selection, parameter tuning, and a maximum number of split optimizations in improving the model's performance. Future studies may further explore these techniques and other potential methods for optimizing the bagged tree ensemble classifier model for predicting diabetes prevalence in women.

CONCLUSION

In conclusion, this study aimed to optimize the bagged tree ensemble classifier for improved prediction of diabetes prevalence in women. The study found that applying feature selection and parameter tuning techniques increased model accuracy, precision, recall, specificity, and F1 score. Specifically, the optimized bagged tree model using 300 splits and 70 learners showed the best performance of the model based on various evaluation metrics in a test dataset, while the optimized model using 500 maximum number of splits and 70 learners showed the best performance in the validation dataset. These findings suggest that using bagged tree ensemble classifiers and optimization techniques can provide improved predictions for diabetes prevalence in women. The insights gained from this study could significantly improve healthcare outcomes for women at risk of diabetes. Further research can be conducted to test the applicability of this approach to other datasets and populations.

ACKNOWLEDGEMENTS

The authors express heartfelt gratitude to the following individuals for their invaluable support and guidance throughout this research: Dr. Ariston O. Ronquillo, Campus Director, of North Eastern Mindanao State University -Tagbina Campus, Philippines, for his unwavering support and encouragement and for providing us the necessary resources and creating a conducive environment for research; Dr. Born Christian Isip, Dean of the College of Information Technology Education, North Eastern Mindanao State University -Main Campus, Philippines, for her insightful feedback and expertise in enhancing the

quality of this work; Dr. Myelinda Baldelovar, Department Chair of the Computer Science Department of North Eastern Mindanao State University -Tagbina Campus, Philippines, for her crucial guidance and support within the department; and Dr. Rolly Salvaleon, the Vice-President for Research and Extension of North Eastern Mindanao State University, Philippines, for approving the financial support for the publication of this research. Thank you all for your significant contributions to this research.

REFERENCES

Abayadeera, N., Jayawardena, R., & Byrne, N. M. (2019). Machine learning-based models for diabetes risk prediction in urban Filipinos. *Journal of Diabetes Research, 2019*, 1-8. https://doi.org/10.1155/2019/3709346

Biau, G. (2012). Analysis of a random forests model. *Journal of Machine Learning Research, 13*(1), 1063-1095.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. CRC press.

Chen, T., Li, Z. H., Yuan, C. X., & Wong, K. C. (2004, July 4-8). *Improving bagging algorithms: Anti-overfitting by bagging adaptive boosting*. [Paper presentation]. Proceedings of the Twenty-first International Conference on Machine Learning (ICML), Alberta, Canada.

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. Morgan Kaufmann Publishers.

International Diabetes Federation. (2019). *IDF diabetes atlas* (9th ed.). International Diabetes Federation. https://www.diabetesatlas.org/upload/resources/material/20200302_133351_IDFATLAS9e-final-web.pdf

Jia, Q., Chen, F., Wang, Y., Huang, B., & Chen, Y. (2018). Application of bagged decision trees for predicting diabetes mellitus in urban Chinese residents. *Journal of Healthcare Engineering, 2018*, 1-10.

Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Procedia Computer Science, 165*, 292-299. https://doi.org/10.1016/j.procs.2020.01.047

Nguyen, T. T., Tran, T. H., & Nguyen, H. H. (2020, November 12-14). *Feature selection techniques for diabetes prediction*. [Paper presentation]. Proceedings of the International Conference on Advanced Data Mining and Applications, Foshan, China.

Nishat, M. M., Faisal, F., Mahbub, M. A., Mahbub, M. H., Islam, S., & Hoque, M. A. (2021). Performance assessment of different machine learning algorithms in predicting diabetes mellitus. *Bioscience Biotechnology Research Communications, 14*(1), 74-82. https://doi.org/10.21786/bbrc/14.1/10

Pang, B., Wang, C., Lu, Y., Cao, J., Zhang, Y., & Jing, L. (2017). Predicting the risk of diabetes mellitus using machine learning techniques. *Frontiers in Genetics, 8*, 1-8.

Philippine Statistics Authority. (2020). *2019 National nutrition survey final results*. Philippine Statistics Authority. https://psa.gov.ph/nutrition-statistics/2019NNSTables

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers.

Tan, C., Ona, E. T., Yu, W., Garabiles, M. R., & Sy, R. G. (2019). Predictive modeling of type 2 diabetes risk among Filipinos using decision trees and logistic regression. *Diabetes Research and Clinical Practice, 153*, 177-185.

Wang, Y., Cao, Y., & Zhang, Y. (2018). An adaptive bagging algorithm for imbalanced data classification. *Applied Soft Computing, 71*, 1018-1030.

Zhang, Y., Liu, L., & Li, Q. (2018, July 16-20). *A feature selection method based on PSO-SVM for diabetes prediction*. [Paper presentation]. IEEE International Conference on Software Quality, Reliability and Security Companion, Lisbon, Portugal.

Zhao, Y., Feng, X., Li, L., Liu, Y., & Zhang, X. (2019). Prediction of diabetes using support vector machine algorithm based on medical examination data. *BMC Medical Informatics and Decision Making, 19*(2), 1-9.