ISSN: 2633-4828 Vol. 5, No.3, September 2023

*International Journal of Applied Engineering and Technology*

# Evaluation of Supervised Machine Learning Algorithms in Diabetes Prediction

Donald Jim L. Auza[1] and John Vianne B. Murcia[2]

[1,2]*University of Southeastern Philippines – Mintal Campus, Davao City, Philippines*

[2]*University of Mindanao, Davao City, Philippines*

[1]djiauza00617@usep.edu.ph, [2]jv_murcia@umindanao.edu.ph

***Date of Submission: 11th August 2023 Revised: 29th September 2023 Accepted: 10th October 2023***

*How to Cite:* **Auza, D. J. and Murcia, J. B. (2023). Evaluation of Supervised Machine Learning Algorithms in Diabetes Prediction. International Journal of Applied Engineering Research 5(1), pp.88-92.**

***Abstract* - This study is conducted to assess various classifiers in diabetes prediction. The diabetes dataset consists of 70,692 instances and is evenly divided between respondents with and without diabetes. Based on the classification procedures, RandomForest algorithm has the highest classification accuracy among other classifiers. On the other hand, Logistic algorithm has the highest classification precision based on a five-fold cross-validation. For lBK, k-NN 7 performed the best over its other variants, whereas J.48 achieved the most significant classification accuracy with a confidence factor of 25%. Compared to other classifiers, RandomForest has the most optimal F-measure, correctly classified instances, and kappa statistic. Future work should evaluate the implemented classifiers on larger datasets, as recommended. By utilizing larger datasets, it is possible to confirm the accuracy and generalizability of the classifiers. This validation process will guarantee the accuracy of the classifiers and provide more trustworthy insights when applied to real-world scenarios.**

***Index Terms* - Classifiers, Data Mining, Diabetes, Prediction, Supervised Machine Learning.**

## INTRODUCTION

Diabetes is a pathological illness that is associated with various life-threatening complications, including myocardial infarctions, neuropathy, renal insufficiency, and cerebrovascular accidents. Diabetes can be attributed to various factors, such as suboptimal dietary patterns, sedentary behaviors, and occupational stress. The increasing prevalence of diabetes among patients has led public health institutions to compile medical records, resulting in the development of a comprehensive database containing valuable data for analytical endeavors. One strategy that can be employed is the utilization of machine learning algorithms to discern patterns within the given dataset.

The healthcare industry extensively utilizes large-scale datasets. The utilization of data analytics in the examination of extensive datasets enables the extraction of valuable insights and the generation of dependable forecasts through the identification of previously unobserved patterns and information.

The healthcare sector encompasses various applications of big data analytics. Currently, hospitals offer a range of diagnostic techniques for diabetic patients, allowing for personalized therapy based on individual patient outcomes. Despite these, it was revealed that the current technique exhibits inadequate categorization and prediction accuracy[1]. This is where machine learning (ML) techniques come into the scene, as these are known to demonstrate high efficacy in the prediction and early diagnosis of diabetes[2].

Early detection of diabetes is essential for preventing future complications and slowing the disease's course. Text, photos, data, and electronic medical records (EMRs) are only two examples of the vast amounts of data generated by the healthcare industry. However, making sense of and acting on this data is still a significant difficulty. New machine-learning techniques may be used to unearth previously unseen patterns that may assist in diagnosing diabetes at an early stage. Neural Networks, Decision Trees (DTs), Deep Learning (DLs), and NaiveBayes (NBs) are all very accurate functional classifiers, with an accuracy ranging from 90% to 98%[3]. This research aims to find the ultimate diabetes classifier using a variety of practical unsupervised learning techniques. Data mining methods, their definitions, the research methodology, the datasets employed, the findings, and the conclusions are all discussed in this part.

Technological advancements greatly influence the modern medical sector. The aged population has a disproportionately high rate of diabetes. There is presently no cure for diabetes. Problems as serious as blindness may develop if diabetes is not addressed. However, if diagnosed early, diabetes may be controlled. Machine learning algorithms have been shown to be successful in illness identification, making them a good fit for the diagnosis of diabetes. K-Nearest Neighbors, Logistic Regression, Naive Bayes, Random Forests, and Decision Trees were only few of the well-established machine learning methods we used in our investigation[4].

K-means and clustering by agglomeration were two unsupervised learning methods for classification that we sought to deploy alongside traditional supervised learning algorithms. The accuracy, recall, sensitivity, and sensitivity of each model have been measured. By comparing the outcomes, researchers were able to determine which model performed better in detecting diabetes, our primary goal. The effectiveness of our models was also compared to that of previously published work, and we discovered that our models outperformed the earlier work[4].

Several studies, such as Vijayan and Anjali[5], presented automated databases for the prediction of diabetes and their use of different classifiers was discussed. The accuracy and efficacy of the system may be dramatically enhanced by carefully choosing reliable classifiers. In fact, the study introduced algorithms known as AdaBoost and Decision Stump used as the foundation of a suggested classification decision assistance system. AdaBoost's accuracy as a primary classifier is checked using the following techniques: NaiveBayes, function logistics, and Decision Tree.

The study by Chang et al.[6] presented electronic diagnostic systems for classification models. These models use three different machine learning methods, then tested to see whether they correctly identified whether or not an individual had diabetes mellitus based on eight criteria. The classifier RandomForest exceeded the J48 and NaiveBayes on the diabetes dataset of Pima Indian, with scores of 79.57% accuracy, specificity of 75.00%, and F-score of 85.17%, while the J48 decision tree had the greatest sensitivity (88.43%) of the three. The disparity between class 0 and class 1 is the reason for the disproportion between accuracy and specificity[6].

Hassan et al. [7] tried to apply unsupervised learning techniques for classification using *K*-means and agglomerative clustering in addition to supervised learning algorithms. The accuracy, recall, sensitivity, and sensitivity of each model have been measured. By comparing the outcomes, we succeeded to determine which model performed better in detecting diabetes, our primary goal. The effectiveness of our models was also compared to that of previously published work, and we discovered that our models outperformed the earlier work.

## METHODS

### *Dataset*

The dataset was acquired the Behavioral Risk Factor Surveillance System (BRFSS) of the Centers for Disease Control (CDC), containing 70,692 instances of clean diabetes data. It has a 50/50 split between respondents with and without diabetes. Diabetes has two classes, 0 for the absence of diabetes and 1 for the presence of diabetes. This balanced dataset contains 21 feature variables and 1 independent variable.

### *Data Preparation*

The datasets were visually inspected, and it was found that it has no deficiencies and appears to be balanced. The numeric dataset requires the unsupervised attribute filters numeric to nominal to be transformed into a nominal dataset. The unsupervised filter's AttributeIndices transforms numeric attribute numbers Diabetes, HighBP, CholCheck, HighChol, Smoker, Stroke, HeartDieaseorAttack, PhysActivity, Fruits, Veggies, HvyAlcoholConsump, AnyHealthCare, NoDocbcCost, DiffWalk, and Sex into nominal attributes. The dataset contains a total of 15 nominal attributes and seven numeric attributes. After conversion, the class attribute's colors were changed from black and white to nominal data type, to red (class 2 with a label of 1) and blue (class 1 with a label of 0). There are total of 15 nominal attributes, and 7 numerical attributes are present in the dataset.

**Table 1.**
**Dataset Description**

| Attribute | Type | Description |
|---|---|---|
| Diabetes_binary | Numeric | Have Diabetes, 0 – No, 1 - Yes |
| HighBP | Numeric | Have High Blood Pressure, 0 means No, 1 means Yes |
| HighChol | Numeric | Have High Cholesterol level, 0 means No, 1 means Yes |
| CholCheck | Numeric | No Check of cholesterol in 5 years, 0 – No, 1 - Yes |
| BMI | Numeric | Body Mass Index of the respondents , actual number |
| Smoker | Numeric | Smoking 100 cigarettes or more in your entire life, 0 means No, 1 means Yes |
| Stroke | Numeric | Experienced stroke, 0 means No, 1 means Yes |
| HeartDieaseorAttack | Numeric | Have Heart Disease, 0 means No, 1 means Yes |
| PhysActivity | Numeric | Physical activity , 0 means No, 1 means Yes |
| Fruits | Numeric | Fruit consumption at least once per day, 0 means No, 1 means Yes |
| Veggies | Numeric | Vegetable consumption at least once per day, 0 means No, 1 means Yes |
| HvyAlcoholConsump | Numeric | More than 14 drinks per week for adult male and more than 7 drinks per week for adult female, 0 means No, 1 means Yes |
| AnyHealthCare | Numeric | Have any kind of health insurance, 0 means No, 1 mean Yes |
| NoDocbcCost | Numeric | Needed to see a doctor but could not afford, 0 means No, 1 means Yes |
| GenHlth | Numeric | Would you say that in general, your health is: scale of 1-5 |
| MentHlth | Numeric | Feel depressed scale 1-30 days |
| PhysHlth | Numeric | Feel sick in past 30 days scale 1-30 |
| DiffWalk | Numeric | Walking problems, 0 means No, 1 means Yes |
| Sex | Numeric | Sex, 0 means female, 1 means male |
| Age | Numeric | Respondent's Age, 1-13 categorized |
| Education | Numeric | Educational level/attainment scale 1-6 |
| Income | Numeric | Monthly income in dollars – scale 1-8 |

### *Selection of Attributes*

Before the classification process, the three most well-known feature selection algorithms in Weka were used to select attributes, ensuring that only relevant ones were included[8]. The suggested approach chooses techniques for feature selection based on correlation (CorrelationAttributeEval), information gain (InfoGainAttributeEval), and learning (WrapperSubsetEval).

The attributes GenHlth (r=0.408), HighBP (r=0.382), BMI (r=0.293), HighChol (r=0.289), and Age (r=0.279) were found to be the most highly correlated, and the five attributes with minimum correlation were Vegetables (r=0.079), Fruits (r=0.054), Sex (r=0.044), NoDocbcCost (r=0.041), and AnyHealthcare (r=0.023).The top five attributes according to InformationGain (entropy)-based feature selection are GenHlth (r=0.134), HighBP (r=0.108), BMI (r=0.082), Age (r=0.68), and HighChol (r=0.061); the bottom five are Veggies (r=0.004), Fruits (r=0.002), Sex (r=0.001), NoDocbcCost (r=0.001), and AnyHealthcare (r=0.000). Learner-based feature selection (WrapperSubsetEval) was utilized to determine the number of folds for the estimation of accuracy. For accurate estimation of the attributes, five folds cross validation is suggested based on the results.

### *Data Classification and Cross-validation*

In order to classify the dataset and predict the class label in the test set, four classifiers were chosen. We used the most popular classifiers, including J48 and RandomForest decision trees, *k*-NN, and NaiveBayes. Due to the binary nature of the class labels to be predicted, additional classifier, Logistic, was also included.

Typically, there are tuning parameters for each classifier. In this study, we tweaked the *k*-NN parameter value and the decision tree's confidence factor (J48). The error rate in k-NN reduces with increasing *k*[9], while the parameter that is altered to examine post-pruning performance is the confidence factor in J48[10]. Classifications were carried out in the k-NN3, k-NN5, and k-NN7 datasets. Because test examples frequently receive labels that are similar to the neighboring example in the training set, we made the decision to omit the default *k*-NN 1 (also known as simple *k*-NN)[11]. Starting with *k*=3, we look at the three most similar groups and choose the one with the highest frequency to label[12]. In the J48 classifier had its confidence factor of 0.25, 0.5, and 0.75.

Cross-validation is used to evaluate the classifier, and what is entered in the folds text field in terms of folds[13]. As stated above, 5 folds cross validation is suggested for accurate estimation of the attributes based on the learner-based feature selection (WrapperSubsetEval). A 10 folds cross validation was also performed to differentiate the results of the classifiers.

## RESULTS AND DISCUSSION

Using the training set test option, the classifier's ability to predict the category of the instances it was trained on is evaluated. There are nine classification tests performed under five classifiers: one for NaiveBayes, three for lazy.lBK, three for trees.J48, one for the trees.RandomForest, and one for the functions.Logistics. Based on the simulations, the trees.RandomForest achieved the greatest classification accuracy (99.40%). Among the lBK variants, the best overall performance was found with k-NN 3 (82.73%), while the highest classification accuracy was attained under trees.J48 with confidence factor of 75% (91.12%). Comparatively, the accuracy of Logistic classification is 74.81% while that of NaiveBayes classification is 71.63%.

**Table 2.**
**Classification accuracy of classifiers on training dataset**

| Classifiers | Variant | Correctly Classified Instances (%) |
|---|---|---|
| Naïve.Bayes | - | 50637 (71.63%) |
| Lazy.lBK (k-NN) | 3 | 58482 (82.73%) |
| Lazy.lBK (k-NN) | 5 | 56040 (79.27%) |
| Lazy.lBK (k-NN) | 7 | 54981 (77.78%) |
| Trees.J48 | 0.25 | 59638 (84.36%) |
| Trees.J48 | 0.50 | 63337 (89.60%) |
| Trees.J48 | 0.75 | 64418 (91.12%) |
| Trees.RandomForest | - | 70269 (99.40%) |
| Functions.Logistics | - | 52882 (74.81%) |

Set We run the nine cross-validations at 10- and 5-folds. Function.Logistic has the highest (74.75% and 74.78%) classification accuracy among other classifiers both in 10 and 5 folds. Among its variants, k-NN 7 had the highest classification accuracy (71.03%) in both the 10-fold and five-fold tests, while J.48 with a confidence factor of 25% achieved the highest accuracy (72.21% and 72.32%) in both tests. The classification accuracy of NaïveBayes is 71.62% at 10 folds and 71.60% at five folds, whereas that of RandomForest is 73.68% at 10 folds and 73.70 percent at five folds. Hence, we can see that the classifiers performed better when training the dataset as it is rather than performing either five- or ten-fold cross-validation.

**Table 3.**
**Performance of cross-validation of dataset (10 folds vs. 5 folds)**

| Classifier | Variant | Cross-Validation | |
|---|---|---|---|
| | | Classification Accuracy (10 folds) | Classification Accuracy (5 folds) |
| Naïve.Bayes | - | 71.62 | 71.60 |
| Lazy.lBK (k-NN) | 3 | 69.03 | 68.96 |
| Lazy.lBK (k-NN) | 5 | 70.30 | 70.25 |
| Lazy.lBK (k-NN) | 7 | 71.03 | 71.03 |
| Trees.J48 | 0.25 | 72.21 | 72.32 |
| Trees.J48 | 0.50 | 68.98 | 69.39 |
| Trees.J48 | 0.75 | 67.55 | 67.90 |
| Trees.RandomForest | - | 73.68 | 73.70 |
| Functions.Logistics | - | 74.76 | 74.78 |

Classification accuracy on the training data set is shown in Table 4 for all used classifiers and their modifications. The F-measure for Trees. Random Forest (0.994) is the highest of the nine classifiers, has correctly classified instances of 70,269 and kappa statistic value of 0.988. Kappa statistics is a measure of true agreement that is not purely derived by chance[14]. The trees. Random Forest algorithm is a collective classifier that creates many decision trees using a bagging method, allowing for the selection of multiple instances of the same sample while not selecting any instances of other samples at all[14]. In the lBK classifier variants, *k*-NN 3 has the highest F-measure (0.827), with 58,482 correctly-classified instances and kappa statistics of 0.6546. In J48, the confidence factor 75% has the highest F-measure (0.911), with 64,418 correctly classified instances, and kappa statistics of 0.8225.

## Conclusion

Using unsupervised data mining methods, the study was able to identify the best classifier for the diabetes dataset. The findings show that when testing on the training set, trees.RandomForest outperforms other classifiers in terms of classification accuracy. Functions.Logistics has the greatest categorization accuracy based on five-fold cross-validation. However, J.48 attained the best accuracy in classification among its variations at a confidence factor of 25%, whereas k-NN 7 was the highest for lBK. As a whole, trees.RandomForest performed the best in terms of the F-measure, percentage of properly categorized cases, and the kappa statistics compared with other classifiers.

## Recommendations

External validation of developed classifiers using independent datasets is crucial. In order to ascertain the generalizability of the models, it is imperative to gather supplementary data from a variety of sources or healthcare settings. This procedure will enhance the level of confidence in the dependability and effectiveness of the classifiers when utilized on novel, unobserved data.

Future research should aim to evaluate the implemented classifiers on larger datasets. By employing larger datasets, one can validate the precision and applicability of the classifiers. The implementation of this validation process will ensure the precision of the classifiers and yield more reliable insights when utilized in real-world contexts. Furthermore, it is imperative to incorporate supplementary independent variables or attributes, such as a familial predisposition to diabetes and other contributing factors, into the dataset in order to enhance the precision of classification.

Additionally, given that trees.RandomForest demonstrated the highest classification accuracy among the classifiers tested on the training set, trees.RandomForest should be regarded the primary classifier for diabetes prediction. Due to its ability to manage high-dimensional datasets and capture intricate feature relationships, it is a strong candidate for accurate classification. Meanwhile, although functions. Logistic did not achieve the greatest classification accuracy on the training set, it achieved the highest classification accuracy during five folds of cross-validation. As a result, it is beneficial to investigate functions. Logistic as a substitute diabetes prediction classifier. Its interpretability and simplicity make it a useful instrument for understanding the impact of individual predictor characteristics.

## References

[1] Mujumdar, A. and Vaidehi, V. Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, pp. 292-299, 2019.

[2] Chauhan, T., Rawat, S., Malik, S., and Singh, P.. Supervised and unsupervised machine learning based review on diabetes care. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 581-585), 2021. IEEE.

[3] Naz, H. and Ahuja, S. Deep learning approach for diabetes prediction using PIMA Indian dataset. Journal of Diabetes & Metabolic Disorders, 19, pp.391-403, 2020.

[4] Singh, S. and Vazirani, V. Classification vs clustering: Ways for diabetes detection. In 2022 IEEE 7th International conference for Convergence in Technology (I2CT), 2022.

[5] Vijayan, V. V., and Anjali, C. Prediction and diagnosis of diabetes mellitus—A machine learning approach. In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS) (pp. 122-127), 2015. IEEE.

[6] Chang, V., Bailey, J., Xu, Q. A., and Sun, Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Computing and Applications, pp.1-17, 2022.

[7] Hassan, M. M., Mollick, S., and Yasmin, F. An unsupervised cluster-based feature grouping model for early diabetes detection. Healthcare Analytics, 2, pp.100-112, 2022.

[8] Gnanambal, S., Thangaraj, M., Meenatchi, V., and Gayathri, V. Classification algorithms with attribute selection: an evaluation study using WEKA. International Journal of Advanced Networking and Applications, 9(6), pp. 3640-3644, 2018.

[9] Dudani, S. A. The distance-weighted k-nearest-neighbor rule. IEEE Transactions on Systems, Man, and Cybernetics(4), pp.325-327, 1976.

[10] Rajesh, P., and Karthikeyan, M. A comparative study of data mining algorithms for decision tree approaches using Weka tool. Advances in Natural and Applied Sciences, 11(9), pp.230-243, 2017.

[11] Enas, G. G., and Choi, S. C. Choice of the smoothing parameter and efficiency of k-nearest neighbor classification. In Statistical methods of discrimination and classification (pp. 235-244), 1986. Elsevier.

[12] Alejandrino, J. C., Bolacoy Jr, J., and Murcia, J. V. B. Supervised and unsupervised data mining approaches in loan default prediction. International Journal of Electrical & Computer Engineering (2088-8708), 13(2), pp.1837-1847, 2023.

[13] Kirkby, R. and Frank, E. WEKA Explorer User Guide for Version 3-4. 2007.

[14] Jankovic, R. Classifying cultural heritage images by using decision tree classifiers in WEKA. Proceedings of the 1st International Workshop on Visual Pattern Extraction and Recognition for Cultural Heritage Understanding Co-Located with 15th Italian Research Conference on Digital Libraries (IRCDL 2019), 2019. Pisa, Italy,

## AUTHOR INFORMATION

**John Vianne Murcia,** College Professor of Business Analytics, University of Mindanao, and concurrent Professorial Lecturer, PhD Program, University of Southeastern Philippines – Mintal Campus, Davao City.

**Donald Jim Auza,** PhD student, University of Southeastern Philippines – Mintal Campus, Davao City.