Analysis of Depth of Entropy and GINI Index Based Decision Trees for Predicting Diabetes

Arjun Singh Saud [1], *Subarna Shakya* [2], *and Bindu Neupane* [3]

Abstract

Diabetes is a disease caused due to malfunctioning of pancreas. In this disease, pancreas either no longer produces insulin or produces insufficient insulin. If this disease is diagnosed early, several health complications can be avoided by taking precautions timely. Otherwise, it may create serious health problems. Nowadays, machine learning models are widely researched for diabetes prediction. This research work uses decision tree classifiers for diabetes prediction and analyzed the impact of decision tree depth in diabetes prediction. Besides this, the research work compared the performances of ID3 and CART decision trees in reference to diabetes prediction. From the empirical observation, we concluded that the CART algorithm has slightly better performance than ID3 and the best prediction performance can be achieved with the decision trees of depth 4.

Keywords : **CART, decision tree classification, depth analysis, diabetes prediction, ID3**

I. INTRODUCTION

Classification is a supervised machine learning approach in which a prediction model is created from the training data and then this model is used to classify test data. Training data is the data that is given as input to the algorithm during training process and the algorithm learns by looking at the patterns available in the data. Test data is the data in observations that are not seen by the algorithm previously and for which output label needs to be predicted. Binary classification and multi-class classification are two major categories of classification algorithms. Examples of binary class classification problem are identifying males and females, determining spam and non-spam emails, predicting the presence/absence of particular disease etc. Classification problems like biometric identification, document classification, image classification, speech recognition, handwriting recognition etc. are examples of multi-class classification problems. Naïve Bayes, Support Vector Machine, Decision Tree, Random Forest, K-Nearest Neighbor etc. are examples of widely used classification algorithms [1].

Decision tree is a machine learning approach that can be used to solve classification and regression problems. It creates prediction model in the form of decision tree. The algorithm divides a dataset into smaller and smaller subsets iteratively. The final result is a decision tree in which internal nodes are decision nodes and leaf nodes represent some class. A decision node has two or more branches. Iterative Dichotomiser 3 (Id3), Classification and Regression Tree (CART), and C 4.5 are examples of widely used decision tree algorithms. Decision Tree is a preferred approach of classification because it is easy to interpret and has very few hyperparameters [1].

Nowadays, machine learning techniques are widely

Manuscript Received : October 3, 2021; Revised : November 6, 2021 ; Accepted : November 12, 2021. Date of Publication : December 5, 2021.

A. S. Saud[1] is *Assistant Professor,* with Central Department of Computer Science and IT, TribhuvanUniversity, Kathmandu, Bagmati, Nepal - 44613. Email : arjunsaud@cdcsit.edu.np ; ORCID iD : https://orcid.org/0000-0002-5235-7322
S. Shakya[2] is *Professor* with Department of Electronics and Computer Engineering, IOE, Tribhuvan University, Lalitpur, Bagmati, Nepal - 44700. Email : drss@ioe.edu.np ; ORCID iD : https://orcid.org/0000-0003-2268-0097
B. Neupane[3] is *Lecturer* with Nagarjuna College of IT, Tribhuvan University, Lalitpur, Bagmati, Nepal - 44700.
Email : bneupanesaud@gmail.com ; ORCID iD : https://orcid.org/0000-0003-1855-6253

DOI : https://doi.org/10.17010/ijcs/2021/v6/i6/167641

used to predict diseases. Basically, classification algorithms are used for solving this type of problem [2]. Diabetes is one of the chronic diseases that occurs when level of blood sugar is high. If diabetes is not identified and treated in time, many health complications may occur. Visiting a diagnostic center and consulting with the doctor is a traditional approach of disease identification. Machine learning approaches help to deal with this problem easily. Classification approaches are used to predict the likelihood of diabetes in patients [3].

This research work compares the performance of ID3 and CART Decision Trees in diabetes prediction. Furthermore, the paper evaluates the impact of the hyperparameter decision tree depth and suggests its optimal value for predicting diabetes. Using smaller depth may lead to the underfitted model and using larger value of depth may lead to overfitted model. Both underfitting and overfitting result in poor prediction accuracy. Therefore, determining the suitable value of decision tree depth is necessary.

This research paper is organized as follows: Section II sheds light on theories and models used in this research work. Section III provides review of the literature relevant to this research work. Section IV includes discussion on research methodology adopted to carry out the research work. Section V provides analysis and interpretation of the experimental results. Finally, Section VI presents research findings and recommendations for future research.

II. RELATED MODELS AND THEORIES

This section discusses ID3 Algorithm, CART algorithm, and concepts of parameter and hyperparameter.

A. ID3 Algorithm

Id3 stands for Iterative Dichotomiser 3. It uses top-down greedy approach to build decision tree model and attempts to create the smallest possible decision tree. This algorithm computes information gain for each attribute and then selects the attribute with the highest information gain. Information gain measures reduction in entropy after data transformation. It is calculated by comparing entropy of the dataset before and after transformation. Entropy is the measure of homogeneity of the sample. Entropy or expected information of dataset D is given by eq.(1) [1].

$$E(D) = -\sum_{i=1}^{m} p_i \log_2 p_i \quad (1)$$

Where p_i is the probability of a tuple in D belonging to class C_i and is estimated using eq. (2).

$$p_i = \frac{|C_{i,D}|}{|D|} \quad (2)$$

Where $|C_{i,D}|$ is the number of tuples in D belonging to class C_i and $|D|$ is the number of tuples in D.

Suppose we have to partition the tuples in D on some attribute A having v distinct values. The attribute A can be used to split D into v partitions $\{D_1, D_2, \ldots\ldots, D_v\}$. Now, the total entropy of data partitions while partitioning D around attribute A is calculated using eq. (3).

$$E_A(D) = -\sum_{j=1}^{v} \frac{|D_J|}{|D|} \text{ x } E(D_j) \quad (3)$$

Finally, the information gain achieved after partitioning D on attribute A is calculated using eq. (4).

$$IG(A) = E(D) - E_A(D) \quad (4)$$

B. CART Algorithm

CART is acronym for classification and regression tree. This algorithm constructs binary tree where each internal node has exactly two children. CART uses Gini Index that measures purity of leaf nodes. A leaf node is considered more impure if mixed training data is assigned to the node. Gini Index of D is calculated using eq. (5) [1]. The probability of incorrectly classifying a randomly chosen element in the dataset is called Gini Index.

$$\text{Gini}(D) = 1 - \sum_{i=1}^{m} p_i^2 \quad (5)$$

Where p_i is the probability that a tuple in D belongs to Class C_i

For each possible binary split, the weighted sum of the Gini Index of each partition is calculated. For example, if a binary split of D on attribute A creates partitions D_1 and D_2, the Gini index of D after the above partitioning is calculated using eq. (6).

$$\text{Gini}_{A}(D) = \frac{|D_1|}{|D|}\text{Gini}(D_1) + \frac{|D_2|}{|D|}\text{Gini}(D_2) \quad (6)$$

Finally, the reduction in impurity of the dataset after partitioning D on attribute A is calculated using eq. (7) given below.

$$\Delta\text{Gini}(A) = \text{Gini}(D) - \text{Gini}_{A}(D) \quad (7)$$

The attribute that maximizes the reduction in impurity or, equivalently, has the minimum Gini Index is selected as the splitting attribute.

C. Parameter vs Hyperparameter

Parameters are the properties of a machine learning model that can be learned during training. Weights and biases of neural network, split points in a decision tree etc. are examples of parameters. On the other hand, hyperparameters are the properties of a machine learning model that cannot be learned during the training process. These properties govern the entire training process and control model parameters. Hyperparameters have significant impact on the performance of the model being trained. Thus, hyperparameter values need to be selected carefully. Learning rate, numbers of epochs, choice of activation functions, depth of decision trees etc. are examples of hyperparameters. Hyperparameter tuning is always an empirical process and hence, its optimal value needs to be identified through experiments [4].

D. Overfitting and Underfitting

Overfitting or underfitting data is the main cause of poor performance in prediction models. Supervised machine learning is a kind of inductive learning that learns the target function from training data. Therefore, generalization is the major goal of a good machine learning model so that better predictions can be made on data that has never been seen by the model during the training process. Overfitting is the result of using an excessively complicated model. It happens when a model learns random fluctuations in the training data as concepts. This prevents the model from being generalized because these concepts do not apply to new data. On the other hand, using an excessively simple model or using very few training samples results in underfitting. In such situations, underlying trend of the data cannot be captured by a machine learning algorithm. Underfitted models can neither model the training data nor generalize on the new data [4].

III. LITERATURE REVIEW

Yuvarani and Selvarani performed a comparison of the three decision tree classification models and concluded that J-48 decision tree model is efficient and accurate [5]. Sisodia & Sisodia used Decision Tree, Support Vector Machine (SVM) and Naive Bayes (NB) algorithms to detect diabetes and found that that the NB algorithm outperformed other algorithms with 76.3% accuracy [3]. Han et al. evaluated the Naive decision trees and ID3 algorithm in predicting diabetes. From the experiments, authors observed that Naïve decision tree has 72% accuracy and ID3 decision tree has 80% accuracy [6]. Al Jarullah used decision tree model for the diagnosis of type-2 diabetes. Data pre-processing was done to improve the quality of data. The accuracy of the resulting model was 78.18% [7]. Chen, Chen, Zhang, and Wu proposed a hybrid approach that combined K-means and Decision tree classifier for predicting type-2 diabetes. From the experiments, authors observed 90.04% accuracy from the proposed model [8].

Kandhasamy and Balamurali compared Performance Decision Tree, KNN, Random Forest, and Support Vector Machine to classify diabetes mellitus patients. The result showed that decision tree classifier is best among the four [9]. Meng, Huang, Rao, Zhang, and Liu used three predictive models, namely, logistic regression, artificial neural networks, and decision tree. The study suggested that the decision tree algorithm had the best classification accuracy of 77.87% [10]. Sabariah, Hanifa, and Sa'adah combined CART and Random Forest (RF) to build the classification model to detect type-2 diabetes. The model achieved the average accuracy of 83.8%, which was higher than the single classifier CART [11]. Zou, Qu, Luo, Yin, Ju, and Tang used decision tree, random forest, and neural network to predict diabetes and observed that that the random forest model can achieve highest accuracy [12].

Yu, Liu, Valdez, Gwinn, and Khoury used support vector machine (SVM) to classify persons with and without diabetes. The result indicated that the SVM modelling is a promising classification approach for detecting common diseases like diabetes [13]. Vijayan and Ravikumar compared EM, KNN, K-means,

Amalgam KNN, and AFIS algorithm to predict diabetes. The experiment showed that amalgam KNN and ANFIS has 80% accuracy [14]. Huang, Wang, and Chan evaluated the Naive Bayes and J-48 and ensemble of five classifiers to predict disease and observed a little improvement of the ensemble approach over pure Naive Bayes and J-48 [15]. Sneha and Gangil proposed machine learning model with modified feature selection procedure for diabetes prediction and compared its performance with the Naïve Bays, Support Vector Machine, Random Forest, and KNN classifiers. The results revealed the fact that the proposed approach is able to give better performance than other approaches [16]. Polat, Güneş, and A. Arslan proposed GDS-LS-SVM learning system for diabetes classification and concluded that the proposed system is more promising than other approaches [17]. Çalişir and Doğantekin proposed LDA-MWSVM system for automatic diabetes diagnosis and found that the proposed approach is better than other approaches [18].

IV. RESEARCH METHODOLOGY

This research is based on quantitative research methodology and used deductive reasoning process. Conceptual framework of diabetes prediction system was employed to carry out this research work is given in Figure 1.

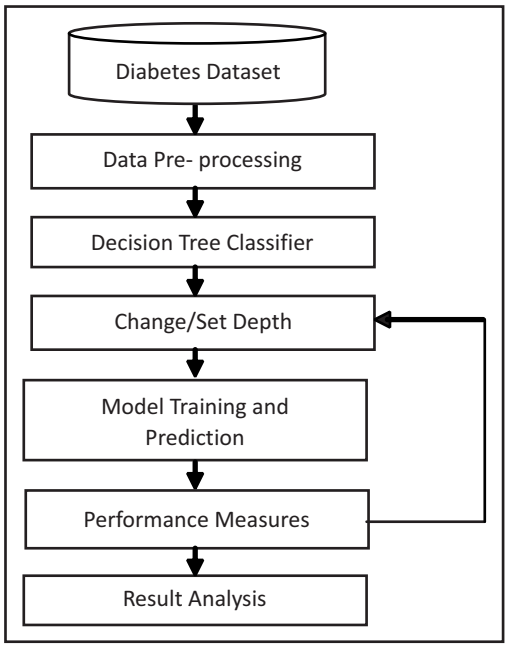

Fig. 1. Conceptual Framework of Diabetes Prediction System

A. Dataset Description

In this study Pima Indians Diabetes Dataset (PIDD) was used, which was downloaded from Kaggle. The dataset had 9 attributes 768 instances. The names of the two target classes were tested positive and tested negative. The number of instances with positive target was 268 and the number of instances with negative target was 500. The 9 attributes of the diabetes dataset were: pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age, and outcome.

B. Result Analysis

This PIDD dataset was given to the decision tree classifiers as input. The depth of decision trees varied between 1 and 20 and performance measures were captured. Finally, the captured result was interpreted and conclusion was drawn.

C. Evaluation Metrics

This research work evaluated decision tree models in terms of four measures: Precision, Recall, F1-score, and accuracy. Formulae for calculating these measures are given in eq.8 - 11 respectively.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (8)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (9)$$

$$F1 = \frac{2 \text{ x Precision x Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total Instances}} \quad (11)$$

V. EXPERIMENTAL RESULTS

The default value of the maximum depth used by Sklearn is 30. In this study, both ID3 and CART decision tree classifiers were executed 10 times and the results were

captured for each execution. The captured results for both decision trees are tabulated below.

A. Analysis of CART and ID3 Decision Trees

Table I and II show ID3 and Cart measures respectively. If we look at the graphs from Fig. 2 to 5, we can clearly see that the performance of CART decision tree is slightly better than the ID3 decision tree. CART is able to outperform ID3 in all measures. The average value of accuracy score for ID3 and CART is 0.703 and 0.738. Similarly, the average F1-score for ID3 and CART is 0.580 and 0.624 respectively. F1-score is a combined measure of recall-score and precision-score. Therefore, the average value of these two measures is not calculated separately. Even though CART performed better than ID3, both decision trees are performing poorly. This is due to overfitting of training data, which can be seen from Tables I and II. If we look at these two tables we can see that values of training accuracy score, recall score,

TABLE I.

ID3 MEASURES

Training Measures				Test Measures			
Accuracy	**Recall**	**Precision**	***F*1-Score**	**Accuracy**	**Recall**	**Precision**	***F*1-Score**
1	1	1	1	0.688	0.582	0.541	0.560
1	1	1	1	0.714	0.594	0.580	0.587
1	1	1	1	0.683	0.594	0.534	0.562
1	1	1	1	0.683	0.582	0.534	0.557
1	1	1	1	0.718	0.594	0.587	0.591
1	1	1	1	0.727	0.607	0.600	0.603
1	1	1	1	0.705	0.582	0.567	0.575
1	1	1	1	0.696	0.607	0.551	0.578
1	1	1	1	0.696	0.607	0.551	0.578
1	1	1	1	0.722	0.632	0.588	0.609

TABLE II.

CART MEASURES

Training Measures				Test Measures			
Accuracy	**Recall**	**Precision**	***F*1-Score**	**Accuracy**	**Recall**	**Precision**	***F*1-Score**
1	1	1	1	0.740	0.620	0.620	0.620
1	1	1	1	0.740	0.645	0.614	0.629
1	1	1	1	0.731	0.632	0.602	0.617
1	1	1	1	0.753	0.658	0.634	0.645
1	1	1	1	0.735	0.645	0.607	0.625
1	1	1	1	0.744	0.632	0.625	0.628
1	1	1	1	0.735	0.620	0.612	0.616
1	1	1	1	0.735	0.632	0.609	0.621
1	1	1	1	0.740	0.632	0.617	0.625
1	1	1	1	0.735	0.620	0.612	0.616

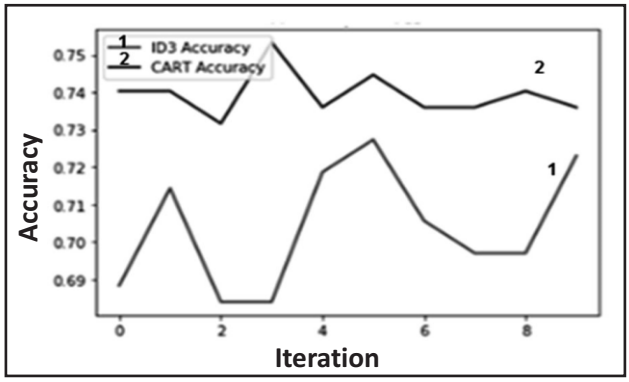

Fig. 2. Accuracy Curve

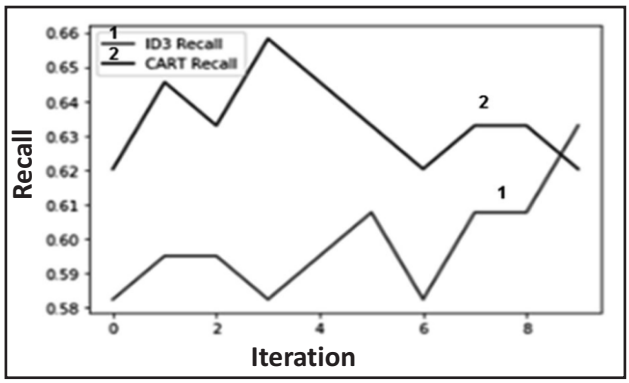

Fig. 3. Recall Curve

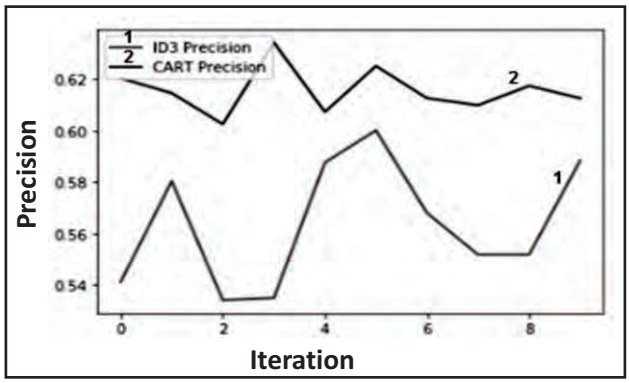

Fig. 4. Precision Curve

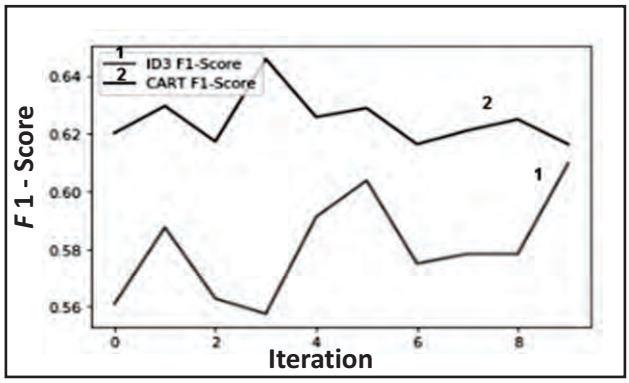

Fig. 5. F1 Score Curve

precision score and F1-score is 1 for all iterations for both trees. This is due to overfitting of training data.

B. Analysis of Depth

Depth of decision trees is one of the important hyperparameters. Its value cannot be learned by decision tree algorithms. In this research work, depth of ID3 and CART decision trees was varied from 1 to 20 and the results were captured for each experiment. The captured results are tabulated next.

If we look at the charts from Fig. 6 to Fig. 13, we can clearly see that values of test accuracy score, recall score, precision score, and F1-score are optimum at the decision tree depth four. However, the values of these scores are decreased sharply for depth value more than four. On the other hand, values of training accuracy score, recall score, precision score, and F1-score are continuously increasing upto depth 16 and 17 for both decision trees and finally, these scores reached 1. This clearly points to the fact that the decision trees were underfitted upto depth

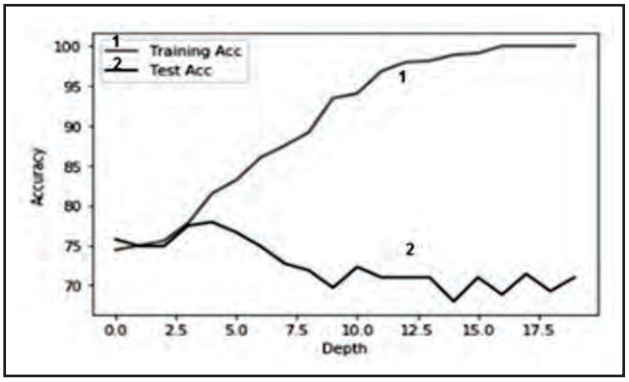

Fig. 6. Accuracy Scores of ID3

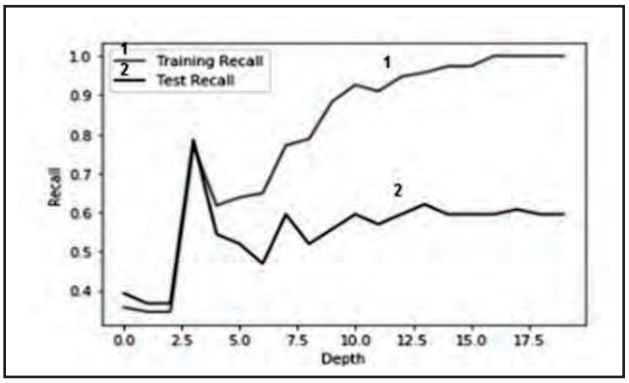

Fig. 7. Recall Scores of ID3

TABLE III.

ID3 MEASURES FOR VARYING DEPTH

	Training Measures				Test Measures			
Depth	**Accuracy**	**Recall**	**Precision**	***F*1**	**Accuracy**	**Recall**	**Precision**	***F*1**
1	0.744	0.356	0.807	0.494	0.757	0.392	0.794	0.525
2	0.750	0.345	0.855	0.492	0.748	0.367	0.783	0.500
3	0.755	0.345	0.890	0.498	0.748	0.367	0.783	0.500
4	0.777	0.771	0.656	0.709	0.774	0.784	0.639	0.704
5	0.815	0.617	0.811	0.700	0.779	0.544	0.741	0.627
6	0.832	0.638	0.845	0.727	0.766	0.518	0.719	0.602
7	0.860	0.648	0.931	0.764	0.748	0.468	0.698	0.560
8	0.875	0.771	0.857	0.812	0.727	0.594	0.602	0.598
9	0.891	0.787	0.891	0.836	0.718	0.518	0.602	0.557
10	0.934	0.882	0.927	0.904	0.696	0.556	0.556	0.556
11	0.940	0.925	0.906	0.915	0.722	0.594	0.594	0.594
12	0.968	0.909	1.0	0.952	0.709	0.569	0.576	0.573
13	0.979	0.946	0.994	0.970	0.709	0.594	0.573	0.583
14	0.981	0.957	0.989	0.972	0.709	0.620	0.569	0.593
15	0.988	0.973	0.994	0.983	0.679	0.594	0.528	0.559
16	0.990	0.973	1.0	0.986	0.709	0.594	0.573	0.583
17	1.0	1.0	1.0	1.0	0.688	0.594	0.540	0.566
18	1.0	1.0	1.0	1.0	0.714	0.607	0.578	0.592
19	1.0	1.0	1.0	1.0	0.692	0.594	0.546	0.569
20	1.0	1.0	1.0	1.0	0.709	0.594	0.573	0.583

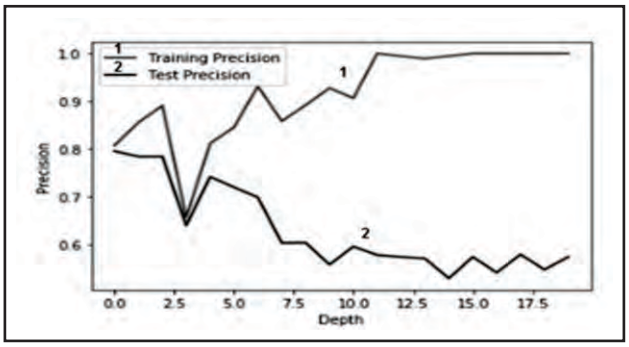

Fig. 8. Precision Scores of ID3

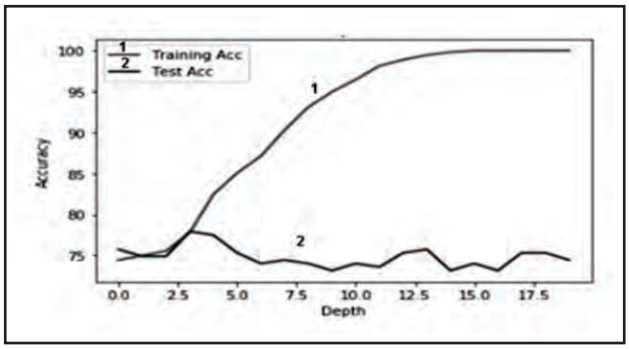

Fig. 10. Accuracy Scores of CART

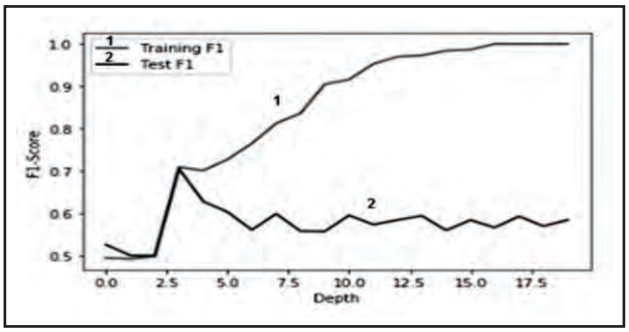

Fig. 9. F1- Scores of ID3

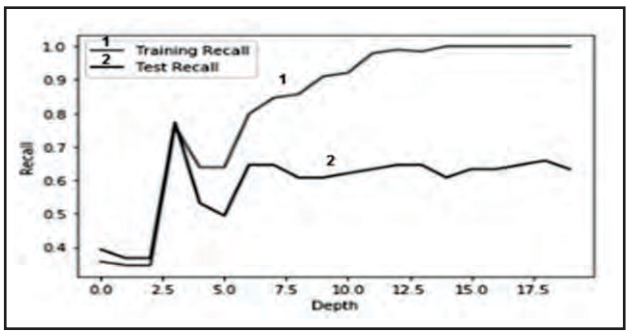

Fig. 11. Recall Scores of CART

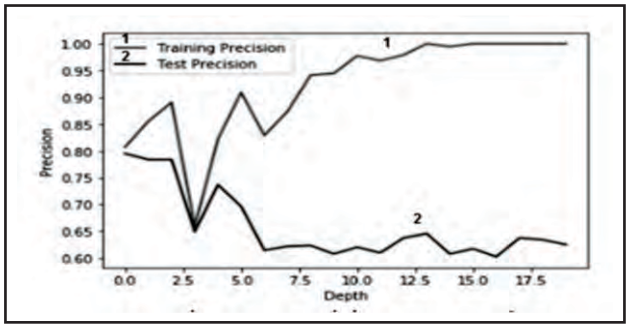

Fig. 12. Precision Scores of CART

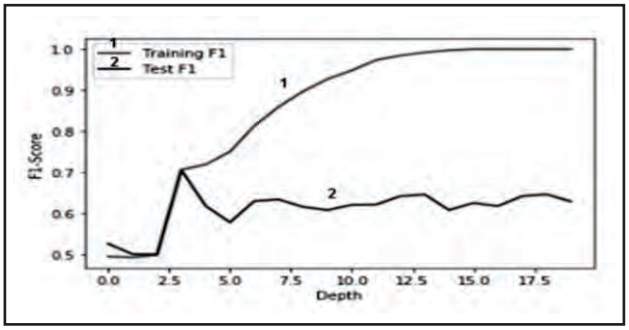

Fig. 13. F1-Scores of CART

three and the decision trees were overfitted for depth more than four.

VI. CONCLUSION

Disease prediction has long been regarded as a critical topic. Machine learning techniques are widely used to solve this type of medical care problem. Basically, classification algorithms are used for solving this type of problem. This research work compared the performance of ID3 and CART Decision Trees in predicting diabetes. Furthermore, the research work analyzed the hyperparameter "depth" of decision trees and suggested the optimal value of the hyperparameter for both decision trees in reference to diabetes prediction. From the experiments, we concluded two facts. First, CART is slightly better than the ID3 decision tree in diabetes prediction. Second, depth of decision trees is a very sensitive hyperparameter and its value may lead to

TABLE IV.

CART MEASURES FOR VARYING DEPTH

	Training Measures				Test Measures			
Depth	**Accuracy**	**Recall**	**Precision**	***F1***	**Accuracy**	**Recall**	**Precision**	***F1***
1	0.744	0.356	0.807	0.494	0.757	0.392	0.794	0.525
2	0.750	0.345	0.855	0.492	0.748	0.367	0.783	0.500
3	0.755	0.345	0.890	0.498	0.748	0.367	0.783	0.500
4	0.777	0.760	0.658	0.706	0.779	0.772	0.648	0.705
5	0.824	0.638	0.821	0.718	0.774	0.531	0.736	0.617
6	0.850	0.638	0.909	0.750	0.753	0.493	0.696	0.577
7	0.871	0.797	0.828	0.813	0.740	0.645	0.614	0.629
8	0.902	0.845	0.873	0.859	0.744	0.645	0.621	0.633
9	0.930	0.856	0.941	0.896	0.740	0.607	0.623	0.615
10	0.949	0.909	0.944	0.926	0.731	0.607	0.607	0.607
11	0.964	0.920	0.977	0.947	0.740	0.620	0.620	0.620
12	0.981	0.978	0.968	0.973	0.735	0.632	0.609	0.621
13	0.988	0.989	0.978	0.984	0.753	0.645	0.637	0.641
14	0.994	1.0	0.994	0.997	0.731	0.607	0.607	0.607
15	0.998	1.0	0.994	0.997	0.731	0.607	0.607	0.607
16	1.0	1.0	1.0	1.0	0.740	0.632	0.617	0.625
17	1.0	1.0	1.0	1.0	0.731	0.632	0.602	0.617
18	1.0	1.0	1.0	1.0	0.753	0.645	0.637	0.641
19	1.0	1.0	1.0	1.0	0.753	0.658	0.634	0.645
20	1.0	1.0	1.0	1.0	0.744	0.632	0.625	0.628

underfitted or overfitted decision tree models. Optimal value of depth for predicting diabetes for both decision trees is 4.

There are many ways to extend this research work. One way to increase reliability of the result of this research is to replicate the research work with other diabetes datasets. Another way is to compare the results of this research work with other classification approaches. Third way of extending this research work is to perform similar type of evaluations for other variations of decision trees like C4.5, C5.0 etc. Another interesting research work that can be carried out is to analyze the attributes that are most relevant for predicting diabetes.

AUTHORS' CONTRIBUTION

All the authors had been actively involved in the presented work. Arjun Singh Saud is the key person behind designing conceptual framework of the study. Subarna Shakya is advisor of the research work and provided feedback at every step of the research work. Bindu Neupane worked on implementation of the research work. Saud and Neupane participated equally in creating the research document.

CONFLICT OF INTEREST

The authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in the manuscript.

FUNDING ACKNOWLEDGEMENT

The authors received no financial support for the research, authorship, and/or for the publication of the article.

REFERENCES

[1]J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techn.,* 3rd ed. Burlington,USA: Morgan Kaufmann, 2011.

[2] C.- H. Weng, T. C.-K. Huang, and R.-P. Han, "Disease prediction with different types of neural network classifiers," *Telemat. Informatics*, vol. 33, no. 2, pp. 277–292, 2016.

[3] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia Comput. Sci.,* vol. 132, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning (Adaptive Computation and Mach. Learning Series)*. Massachusetts: The MIT Press, 2016.

[5] S. Yuvarani and R. Selvarani, "An analysis of decision tree models for diabetes," *Int. Res. J. Eng. Technol.,* vol. 3, no. 11, pp. 680–684, 2016. [Online]. Available: https://www.irjet.net/archives/V3/i11/IRJET-V3I11118.pdf

[6] J. Han, J. C. Rodriguez, and M. Beheshti, "Diabetes data analysis and prediction model discovery using rapidminer," in *2nd Int. Conf. Future Generation Communication and Networking,* Hainan, China, Dec. 13-15, 2008, pp. 96–99, doi: 10.1109/FGCN.2008.226.

[7] A. A. Al Jarullah, "Decision tree discovery for the diagnosis of type II diabetes," in *Int. Conf. Innovations Inform. Technol.,* Apr. 25-27, 2011, pp. 303–307, doi: 10.1109/INNOVATIONS.2011.5893838

[8] W. Chen, S. Chen, H. Zhang, and T. Wu, "A hybrid prediction model for type 2 diabetes using K-means and decision tree," in *2017 8th IEEE Int. Conf. Software Eng. and Service Sci.,* 2017, pp. 386–390, doi: 0.1109/ICSESS.2017.8342938

[9] J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict Diabetes Mellitus," *Procedia Comput. Sci.,* vol. 47, pp. 45–51, 2015, doi: 10.1016/j.procs.2015.03.182

[10] X.-H. Meng, Y.-X. Huang, D.-P. Rao, Q. Zhang, and Q. Liu, "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors," *The Kaohsiung J. Med. Sci.,* vol. 29, pp. 93–99, Feb. 2013, doi: 10.1016/j.kjms.2012.08.016

[11] M. T. M. K. Sabariah, S. T. A. Hanifa, and M. T. S. Sa'adah, "Early detection of type II Diabetes Mellitus with random forest and classification and regression tree (CART)," in *Int. Conf. Advanced Informatics: Concept, Theory and Application (ICAICTA),* Bandung, Indonesia, 2014, pp. 238–242, doi: 10.1109/ICAICTA.2014.7005947

[12] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Front. Genet.,* vol. 9, Nov. 2018, doi: 10.3389/fgene.2018.00515

[13] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes," *BMC Med. Inform. Decis. Mak.,* vol. 10, no. 1, Art. no. 16, 2010, Art no 16, doi: 10.1186/1472-6947-10-16

[14] V. Vijayan V. and A. Ravikumar, "Study of data mining algorithms for prediction and diagnosis of Diabetes Mellitus," *Int. J. Comput. Appl.,* vol. 95, no. 17, pp. 12–16, Jun. 2014. [Online]. Available: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.670.9608&rep=rep1&type=pdf

[15] F. Huang, S. Wang, and C.-C. Chan, "Predicting disease by using data mining based on healthcare information system," in *2012 IEEE Int. Conf. Granular Computing,* 2012, pp. 191–194, doi: 10.1109/GrC.2012.6468691

[16] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *J. Big Data,* vol. 6, no. 1, 2019, Art. no. 13, doi: 10.1186/s40537-019-0175-6

[17] K. Polat, S. Güneş and A. Arslan, "A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine," *Expert Syst. Appl.,* vol. 34, no. 1, pp. 482–487, Jan. 2008, doi: 10.1016/j.eswa.2006.09.012

[18] D. Çalişir and E. Doğantekin, "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier," *Expert Syst. Appl.,* vol. 38, no. 7, pp. 8311–8315, Jul. 2011, doi: 10.1016/j.eswa.2011.01.017

About the Authors

Arjun Singh Saud received M.Sc. degree in Computer Science and Information Technology from Tribhuvan University, Kathmandu, Nepal in 2006. He is a Ph.D. fellow and Assistant Professor with the same department. His research interests include Artificial Intelligence, Machine Learning, Deep Learning, and Data Science.

Subarna Shakya received M.Sc. and Ph.D. degrees in Computer Engineering from the Lviv Polytechnic National University, Ukraine, in 1996 and 2000 respectively. Currently, he is a Professor at the Department of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering, Tribhuvan University. His research interests include E-Government system, Distributed & Cloud computing, Software Engineering & Information System, Deep Learning, and Data Science.

Bindu Neupane received M.C.A. degree (Computer Application) from Purbanchal University, Biratnagar, Nepal in 2020. Currently, she is a faculty member of Nagarjuna College of IT, Tribhuvan University. Her research interests include Data Mining and Bioinformatics.