Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

A Comparative Analysis of Prediction of Student Results Using Decision Trees and Random Forest

Narayan Prasad Dahal[1], Subarna Shakya[2]

[1]Nepal Open University, Lalitpur, Nepal
[2]Pulchowk Campus, Institute of Engineering, Tribhuvan University, Lalitpur, Nepal

E-mail: yourdarpan@gmail.com, drss@ioe.edu.np

Abstract

Many types of research are based on students' past data for predicting their performance. A lot of data mining techniques for analyzing the data have been used so far. This research project predicts the higher secondary students' results based on their academic background, family details, and previous examination results using three decision tree algorithms: ID3, C4.5 (J48), and CART (Classification and Regression Tree) with other classification algorithms: Random Forest (RF), K-nearest Neighbors (KNN), Support Vector Machine (SVM) and Artificial Neural Network (ANN). The research project analyzes the performance and accuracy based on the results obtained. It also identifies some common differences based on achieved output and previous research work.

Keywords: Data mining, decision tree, random forest

1. Introduction

In recent years, Data mining is the most influenced topic of technology research. Data mining is used in different areas to generate new ideas or output based on previous datasets. It is also the emerging technique for analyzing the students' performance, academic success, achievement, the effectiveness of education, etc. So, it is necessary to research the students' result prediction to prepare early and take necessary actions before admission and final examination for more academic success.

The previous research found using different classification algorithms for predicting students' performance. However, there are many differences in accuracy and performance for large and small datasets. The research project has used small datasets containing around 1000

DOI: https://doi.org/10.36548/jtcsst.2022.3.001
Received: 18.06.2022, received in revised form: 06.07.2022, accepted: 22.07.2022, published: 01.08.2022

records of students acquired from online repositories. The research project has used two different datasets, one having 4 attributes and another having 32 attributes. The first dataset used students' past results obtained in 3 different subjects (Physics, Math, and Chemistry) and made the prediction of Pass or Fail. The second dataset uses students' academic background, family background, and past results to make the prediction of Pass or Fail in the final examination. The first dataset is suitable for prediction after the entrance examination and the second suggests the final academic performance score before the final examination.

2. Problem Statement

In the past years, many researchers tried to find the highest performance algorithms with more accurate results. Some of the algorithms are complex to understand and others are easier to know the logic used in it. Decision trees are simpler algorithms to understand and are a very effective method for supervised learning [1]. There are many decision tree algorithms and they have different performances in different researches. It may depend on the nature and volume of the data. This research primarily focuses on finding the best performance decision tree algorithm with a low and medium number of attributes for smaller datasets with around 1000 records. To know the performance of such decision trees compared to other classification algorithms, other commonly used classification algorithms like KNN, SVM, ANN, etc. are also used in the research project work.

3. Scope and Limitations

The decision tree algorithms are widely used data mining tools for prediction and ID3, C4.5 and CART are the most used decision tree algorithms [2] chosen for the proposed research project. These algorithms have satisfactory accuracy and performance to make different predictions in education, health, business, etc. However, other several classification algorithms have shown significantly high performance for different nature and sizes of datasets.

4. Literature Review

The decision tree algorithms are widely used data mining tools for prediction and ID3, C4.5 and CART are the most used decision tree algorithms [2] chosen for the purposed research project. These algorithms have satisfactory accuracy and made different predictions

in education, health, business, etc. However, several classification algorithms have shown significant performance for large and small datasets.

Also, single decision tree the algorithms like ID3, C4.5, and CART are leading the research to satisfactory results [3]. However, these algorithms have been proved better than any others being used recently [4]. These algorithms can be suitable for small data sets for prediction [5]. Out of these, C4.5 is the most preferred algorithm in machine learning and students' performance based on past results [3], [1] and-but another paper shows CART is the best algorithm for the classification of data for predicting student's performance in education [2]. The RF provided more accuracy compared to other supervised machine learning algorithms [6].

5. Methodology

The research project is based on the qualitative method of research. It used the data mining model that contains Data preparation, selection and transformation, and data mining. In Educational Data Mining, different educational information is extracted using different academic or non-academic data [3]. After the mining process, the result is analyzed and discussed. It used the Python Code and libraries for data preprocessing, construction of data mining modules, and display results.

5.1. Data Preparation

The data collection for the evaluation is used from secondary sources. The data are obtained from open online dataset repositories www.kaggle.com and UCI Machine Learning Repository (https://archive.ics.uci.edu/).

The two different data sets having 4 and 33 attributes are used. The first dataset contains 1000 records of the student. The data contains relatively four variables and a sample of the data looks like this:

Table 1. Sample collection data

Maths	Physics	Chemistry	Result
17	27	22	0
72	82	77	1
97	18	13	0
8	42	37	0

32	25	20	0
15	73	68	0
63	67	62	1

- Maths- Marks obtained in Maths (0-100)
- Physics- Marks obtained in Physics (0-100)
- Chemistry- Marks obtained in Chemistry (0-100)
- Result- The final result i.e. PASS (1) or FAIL (0)

The second dataset contains 650 records of students who studied the Portuguese language as a major subject and 396 records of students who studied Mathematics combined in a single database. The data contains relatively 33 variables.

Table 2. Variables used in dataset 2

S.No	Attribute Name	Information	Data Type
1	School	School Name: GP or MS (Gabriel Pereira or Mousinho da Silveira)	Binary
2	Sex	Male or Female (F or M)	Binary
3	Age	(Between 15 to 22)	Numeric
4	Address	Address: U or R (urban or rural)	Binary
5	Famsize	Size of Family: LE3 or GT3 (≤ 3 or >3)	Binary
6	Pstatus	Status of Parents' Cohabitation: T or A (living together or apart)	Binary
7	Medu	Mother's Education: 0 to 4 (0-none, 1- up to 4th grade, 2- between 5th to 9th grade, 3- secondary school, or 4 - higher education)	Numeric
8	Fedu	Father's Education: 0 to 4 (0-none, 1- up to 4th grade, 2- between 5th to 9th grade, 3- secondary school, or 4 - higher education)	Numeric

9	Mjob	Mother's Occupation: teacher, health care, civil services, at home or other	Nominal
10	Fjob	Father's Occupation: teacher, health care, civil services, at home or other	Nominal
11	Reason	Reason to select this school: Near to home, reputation of school, course preference or other	Nominal
12	Guardian	Students' Guardian: mother, father or other	Nominal
13	Traveltime	Time taken to reach the school: 1 to 4 (1 - <15 min., 2- 15 to 30 min., 3- 30 min. to 1 hour, or 4 - >1 hour)	Numeric
14	Studytime	Weekly study time: 1 to 4 (1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)	Numeric
15	Failures	Past class failure count: (n if 1<=n<3, else 4)	Numeric
16	Schoolsup	Extra support to education (yes or no)	Binary
17	Famsup	Family support on education (yes or no)	Binary
18	Paid	Extra classes paid under the course subject: Portuguese or Math (yes or no)	Binary
19	Activities	Extra-curricular activities (yes or no)	Binary
20	Nursery	Nursery school completed (yes or no)	Binary
21	Higher	Preference to study higher education (yes or no)	Binary
22	Internet	Internet facility at home (yes or no)	Binary

23	Romantic	Romantic relationship status (yes or no)	Binary
24	Famrel	Family relationships quality: 1 to 5 (1- very bad to 5- excellent)	Numeric
25	Freetime	After school free time (1- very low to 5 - very high)	Numeric
26	Gout	Going along with friends (1- very low to 5- very high)	Numeric
27	Dalc	Alcohol consumption on workdays (1- very low to 5- very high)	Numeric
28	Walc	Alcohol consumption on weekends (1- very low to 5- very high)	Numeric
29	Heath	Present health status (1- very bad to 5- very good)	Numeric
30	Absences	School absences count: 0 to 93	Numeric
31	G1	1st period grade: 0 to 20	Numeric
32	G2	2nd period grade: 0 to 20	Numeric
33	G3	Final grade: 0 to 20	Numeric

5.2 Data Preparation

After the data collection, the numeric data are transformed into the required form. For dataset1, Data are classified with fixed sets ranking by the final score obtained by the students.

Marks above 79 ➔ Very Good

Marks between 60 and 79 ➔ Good

Marks between 40 and 59 ➔ Satisfactory

Marks below 40 ➔ Poor

The result is in binary representing,

1 ➔ Pass

0 ➔ Fail

For dataset2, only the final GRADE Score (0-20) is transformed into binary representing 1(Pass) and 0(Fail).

Marks above 9 ➔ 1

Marks below 10 ➔ 0

The data does not contain null values. All the data are converted in the number format and transformed to a similar scale using the scalar function.

5.3 Implementation of the mining model

The results are constructed using data a mining model. The algorithms used for prediction are:

- **ID3**: J. Ross Quinlan originally developed ID3 (Iterative DiChaudomiser 3) at the University of Sydney[1].
- **C4.5**: This algorithm was proposed in 1993, again by Ross Quinlan [1]. It allows to calculate the gain ratio and calculates based on information gain[1].
- **CART**: Classification and Regression Trees (CART) allows to define the forecast value after changing the value of another variable[1]. It builds trees based on the Gini index[3].
- **Random Forest**: A random forest (RF) uses many decision trees and predicted the output that has the maximum vote [6].
- **K-nearest Neighbour**: The K-nearest neighbor (KNN) algorithm is one of the simplest and earliest classification algorithms[6].
- **Support Vector Machine**: The Support vector machine (SVM) is a strong machine learning algorithm that can be used with both linear and non-linear data for prediction [6].

- **Artificial Neural Networks**: Artificial neural networks (ANNs) are a set of machine learning algorithms that are inspired by the functioning of the neural networks of the human brain[6].

6. Predictive Analytics

The execution time and accuracy are the major basis for the analysis of the result. The accuracy, precision, recall, f1-score, AUC, and execution time are used to measure the overall performance.

The confusion matrix uses the formula [6]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Here,

TP – True Positive

TN – True Negative

FP – False Positive

FN – False Negative

After that, the values are displayed using other performance measures: precision, f1 score, and recall. The result is visualized using the ROC curve and displayed in the AUC area for the accuracy of the results. In addition, the post-processing techniques Bagging and Boosting are used to check the accuracy of the decision tree classifier. The 10-fold cross-validation is performed to validate the accuracy results.

7. Results and Discussion

This research project comparatively discusses the performance of the different Decision tree algorithms: ID3, C4.5, and CART and other classification algorithms: Random Forest (RF), K-nearest Neighbour (KNN), Support Vector Machine (SVM), and Artificial Neural Network (ANN). Finally, helps to build a model for predicting students' results based on the past data using the aforementioned prediction algorithms. The results were obtained after the execution of coding. The ROC curve shows the distance from the actual to the predicted score.

In the first database C4.5, CART, RF, and KNN showed 100% accuracy CART and KNN have the lowest execution time compared to other. ANN also showed 100% accuracy but due to maximum iteration time, it showed very slow performance compared to the other algorithms.

According to the result, ID3 showed a 100% positive ratio over total positive but due to low (below 0.5) 0.37 positive ratios over the actual group it had overall 0.53 combined (positive + negative) ratios with 85% accuracy.

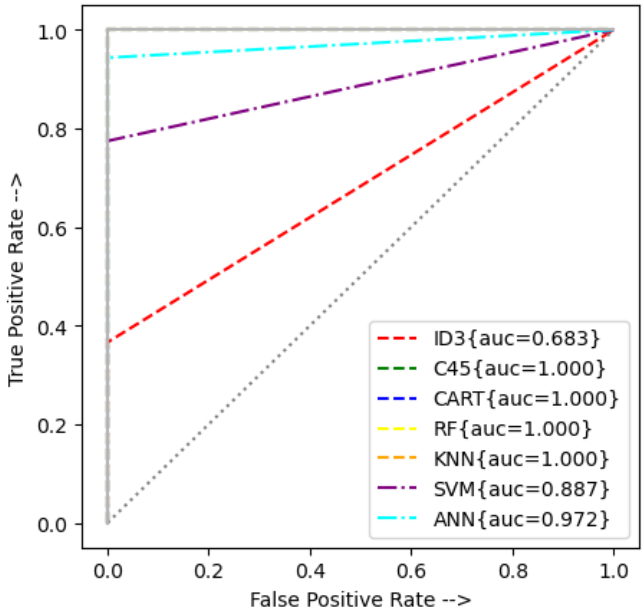

Figure 1. ROC using dataset1

Table 3. Output of dataset1

S. No.	Algorithm	Accuracy (%)	Precision	Recall	F1-score	Execution time (sec.)	AUC area
1	ID3	85.00	1.0	0.37	0.54	0.04	0.683
2	C4.5	100	1.0	1.0	1.0	0.17	1.000
3	CART	100	1.0	1.0	1.0	0.01	1.000
4	RF	100	1.0	1.0	1.0	0.30	1.000
5	KNN	100	1.0	1.0	1.0	0.03	1.000
6	SVM	94.67	1.0	0.78	0.87	0.03	0.887
7	ANN	100	1.0	1.0	1.0	10.92	0.972

However, the ID3 showed 80% of minimum and 86% maximum accuracy after 10-fold cross-validation. The average of 83% with ±1.90 S.D. is even lower than the previous accuracy. Except for ID3 and SVM, all others showed 100% accuracy which is suitable for similar datasets.

Table 4. 10-fold cross-validation summary (database1)

S.No.	Algorithm	Min (%)	Max (%)	Average (%)	Standard deviation
1	ID3	80	86	83	1.90
2	C4.5	100	100	100	0.00
3	CART	100	100	100	0.00
4	RF	100	100	100	0.00
5	KNN	100	100	100	0.00
6	SVM	95	100	97	1.5
7	ANN	100	100	100	0.00

In the second database, CART showed the highest accuracy and lowest execution time compared to other algorithms. However, other decision tree algorithms C4.5 and Random Forest also provided high accuracy results. But, C4.5 and RF took more time to execute than KNN and SVM. Here, the ID3 algorithm has moderate accuracy and execution time. According to the result, KNN has the lowest accuracy and the other two algorithms ANN and SVM have moderate accuracy with no significant difference from ID3.

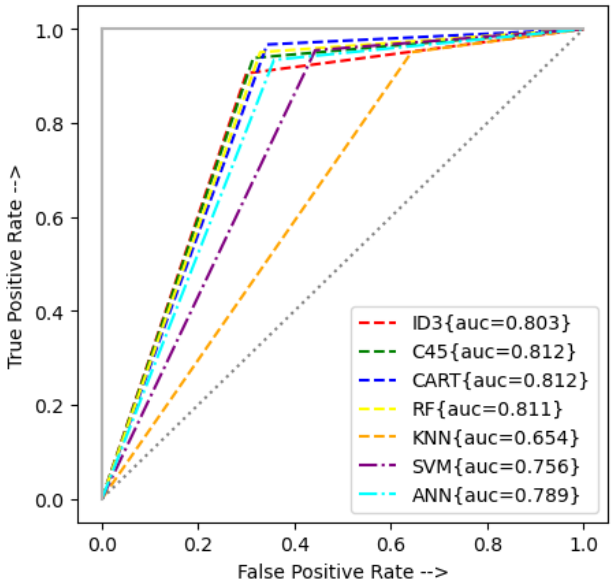

Figure 2. ROC using dataset2

Table 5. Output of dataset1

S. No.	Algorithm	Accuracy (%)	Precision	Recall	F1-score	Execution time(sec.)	AUC area
1	ID3	85.99	0.91	0.91	0.91	0.54	0.803
2	C4.5	88.22	0.91	0.94	0.93	2.06	0.812
3	CART	89.81	0.91	0.97	0.94	0.01	0.812
4	RF	88.85	0.91	0.95	0.93	0.37	0.811
5	KNN	81.85	0.84	0.95	0.89	0.06	0.654
6	SVM	86.62	0.88	0.95	0.91	0.04	0.756
7	ANN	86.94	0.90	0.93	0.92	2.49	0.789

After applying the Bagging technique, the CART showed 91% accuracy but after boosting showed only 87.65% accuracy. Therefore, for small datasets Bagging can improve the accuracy of the result.

Table 6. 10-fold cross-validation summary (database2)

S.No.	Algorithm	Min (%)	Max (%)	Average (%)	Standard deviation
1	ID3	81.90	90.48	86.10	2.42
2	C4.5	84.76	93.33	88.29	2.63
3	CART	83.65	92.31	88.41	2.35
4	RF	86.67	97.12	91.20	3.15
5	KNN	86.67	97.12	91.20	3.15
6	SVM	82.86	95.19	89.19	3.52
7	ANN	84.76	92.31	89.18	2.67

With 10-fold cross-validation, the algorithms RF and KNN showed the highest Minimum and Maximum accuracy (%) with the highest average of 91.20% and ±3.15 S.D. But, C4.5, CART, SVM, and ANN have no significant difference having an average from 88.29% to 89.19% with ±2.35% to ±3.52%.

8. Conclusion

Finally, with the smaller dataset having fewer or more attributes Decision Tree algorithms can provide good performance for the prediction of students' results. According to the results, obtained CART showed the highest performance and accuracy compared to all other algorithms. Decision Trees using C4.5 and Random Forest also showed good accuracy but with more execution time.

References

[1] Hssina, Badr, Abdelkarim Merbouha, Hanane Ezzikouri, and Mohammed Erritali. "A comparative study of decision tree ID3 and C4. 5." International Journal of Advanced Computer Science and Applications 4, no. 2 (2014): 13-19.

[2] Yadav, Surjeet Kumar, Brijesh Bharadwaj, and Saurabh Pal. "Data mining applications: A comparative study for predicting student's performance." arXiv preprint arXiv:1202.4815 (2012).

[3] Yadav, Surjeet Kumar, and Saurabh Pal. "Data mining: A prediction for performance improvement of engineering students using classification." arXiv preprint arXiv:1203.3832 (2012).

[4] Aksu, Gökhan, and Nuri Dogan. "Comparison of Decision Trees Used in Data Mining= Veri madenciliginde kullanilan karar agaçlarinin karsilastirilmasi." Pegem Journal of Education and Instruction 9, no. 4 (2019): 1183-1208.

[5] Singh, N. P., and Nakul Gupta. "Comparative analysis of data mining models for classification for small data set." In 2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT), pp. 1-5. IEEE, 2018.

[6] Uddin, Shahadat, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. "Comparing different supervised machine learning algorithms for disease prediction." BMC medical informatics and decision making 19, no. 1 (2019): 1-16.

Author’s biography

Narayan Prasad Dahal is pursuing his final semester of M. Phil. in ICT at the Faculty of Science, Health and Technology, Nepal Open University, Lalitpur, Nepal. He received the Masters of Science in Information and Technology (M. Sc. IT) from Sikkim Manipal University, Gantok, India in 2013. He works in different colleges as an assistant lecturer in Computer Science and Information Technology.

Subarna Shakya received MSc and Ph.D. degrees in Computer Engineering from Lviv Polytechnic National University, Ukraine, in 1996 and 2000 respectively. He is a Professor of Computer Engineering, Pulchowk Campus, Institute of Engineering, Tribhuvan University, Nepal. He has served as Executive Director at the National Information Technology Center, Government of Nepal.