Real-world Applications of Common Decision Tree Algorithms: A Short Review

Kiperband Morin[1], Cohen Adi[1], Talmid Menachem[1], Rosenberg Amitai[1]

[1] Department of Technology Management, Bar Ilan University, Ramat Gan, Israel

Abstract

Today, data mining is part of a new norm and is used everywhere in research and practical applications. A significant field in data mining is decision trees. These can be used to classify complex datasets and perform future predictions based on data processing. Two of the most used decision tree algorithms include ID3 and C4.5, which can be used relatively easily in popular machine learning software applications like Weka. After reviewing the research regarding these algorithms, we wanted to make their understanding more accessible to people without prior knowledge of the subject. In this article, we touch the surface of these algorithms with a simple explanation of how they work. We continue with a review of two practical and relevant case studies of how ID3 and C4.5 are used for essential findings and can have an enormous impact on important matters. We are excited to see where the research and science are heading with these algorithms and eager to share them with the world.

1. Introduction

In recent years, technological advancement has changed our approach to handling data. Data mining is being applied to learning hidden patterns and making appropriate predictions. It refers to a collection of techniques used to extract hidden knowledge, such as patterns, relationships, or rules, from large datasets[1].

Data mining is an interdisciplinary field that derives its methods from machine learning, artificial intelligence, and statistics, among others.

Data mining algorithms can be used to build models built upon existing data. These models can be applied to solve classification, regression, clustering, and optimization problems.

Machine learning applications via data mining can be found in many fields, including retail, banking, education, health sectors, and more. To process the extensive data emanating from the various sectors, researchers are developing different algorithms using expertise from several fields and knowledge of existing algorithms.

Decision Tree is a supervised learning approach used in statistics, data mining, and machine learning. In data mining and machine learning, a classification or regression decision tree is used as a predictive model to draw conclusions about a set of observations.

The Decision Tree method divides a dataset into subsets based on the most significant attributes in the dataset. The way the decision tree identifies this attribute and how this splitting is done is decided by each specific algorithm.

The most significant predictor in a decision tree is designated as the root node, and dividing is done to form branches that contain sub-nodes called decision nodes. The nodes that do not split further are terminal or leaf nodes.

Given their intelligence and simplicity, decision trees are among the most popular machine learning algorithms[2].

A decision tree follows a top-down approach, also called a gluttonous system, as it only considers the current node between the worked on without focusing on the future nodes.

In 1986, a machine learning researcher named J. Ross Quinlan developed a decision tree algorithm known as ID3. Later, he presented C4.5, which was the successor of ID3. ID3 and C4.5 adopt a greedy approach. Decision tree algorithms include ID3, C4.5, C5.0, and CART, where ID3 and C4.5 are primarily used in classification problems.

ID3 and C4.5 are two of the most used decision tree algorithms[3]. Hence, we will focus on them in our article.

In this article, we will explain these algorithms and present relevant real-world use cases for practical application, which are at the forefront of research.

The article is intended for an overhead overview of decision trees, to understand in summary where the research stands on this topic and how these algorithms can be used for essential and practical applications.

We hope this article can be a gateway for anyone wishing to attain knowledge in decision trees and consider using it in their studies.

2. Review of Literature

ID3

ID3 was invented by Ross Quinlan in 1986[4], to generate a decision tree from a dataset.

ID3 uses two metrics named Entropy and Information Gain; both will be explained later in the appropriate section.

In an ID3-based decision tree, only categorical type features can be used, and numerical type features cannot be applied.

Some disadvantages of ID3 include overfitting (the model performs significantly better for training data than new data) and long computation time.

Several researchers have attempted to improve ID3 using different methods: Yi-bin et al., 2017[5] proposed an improved ID3 algorithm that simplifies the formula and thus makes the decision tree building process faster. Wang et al., 2017[6] and Soni and Pawar, 2017[7] proposed novel approaches that mitigate ID3's bias toward multi-valued attributes.

ID3 has been implemented in many studies, including food health[8], medicine[9], and performance evaluation[10], to name a few.

C4.5

Later, in 1993, Quinlan developed C4.5 to overcome some of ID3's limitations[11]. In 2011, developers of the Weka machine learning software described the C4.5 algorithm as "a landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date"[12]. It became quite popular after ranking first in the Top 10 Algorithms in Data Mining pre-eminent paper published by Springer LNCS in 2008[13].

The improvements in C4.5 include using the information gain ratio metric (IGR, explained later) rather than Information Gain, as in ID3. Secondly, pruning (removal of those branches

in our decision tree which do not contribute significantly to our decision process) can be done during or after the tree's construction. Thirdly, C4.5 can handle attributes that contain continuous features. Also, it can handle missing data[14].

However, even the improved C4.5 algorithm came with some limitations, which researchers are trying to solve. For instance: It constructs empty branches with zero values, and overfitting, although better than ID3, still occurs when the algorithm picks up data with unusual features, especially noisy data[15].

In (Chen et al., 2013)[16], researchers proposed an improved C4.5 decision tree algorithm based on sample selection to improve the classification accuracy, reduce the training time of a large sample, and find the best training set. Muslim et al. (2018)[17] conducted research to improve the accuracy of the C4.5 algorithm.

In (Cherfi et al., 2018)[18], a novel algorithm for building decision trees was proposed. This algorithm, named VFC4.5 (Very Fast C4.5), is an improvement of C4.5.

Lastly, several researchers have attempted to improve decision tree algorithms, including (Yuan and Wang, 2016)[19], (She et al., 2017)[20], (Chandrasekar et al., 2017)[21], and others.

The C4.5 algorithm has been implemented in various cases. For instance, one study[22] studied data mining in identifying determinant factors related to customer satisfaction in a fast-food restaurant with system accuracy results of over 80 %. Another study[23] used classification methods (Naïve Bayes, C4.5, SVM) in classifying text in Arabic. A third[24] measured student performance through mining the information hiding inside the student scores.

3. ID3

Introduction

ID3 stands for Iterative Dichotomiser 3 and is named such because the algorithm iteratively (repeatedly) dichotomizes (divides) features into two or more groups at each step until the end of the process.
This algorithm divides the dataset into training data (which develops and trains the model) and testing data (which validates the model).
The algorithm is a learning algorithm that uses the top-down greedy approach. This means that the tree is built from top to bottom, and in each iteration, the best attribute at that moment is chosen to create a node.
The purpose of the resulting tree is a classification of future samples.

ID3 is considered a decision tree algorithm that creates a simple and efficient tree with the smallest depth[25]. The data type of the ID3 algorithm dataset is categorical only, and it cannot use a continuous dataset for simulation[26].

ID3 uses two metrics to build the tree:

1. **Entropy** – Entropy is a measure of uncertainty in a dataset. The higher the entropy, the higher the uncertainty. In decision trees, the metric determines how informative a node is[27].

$$entropy(Set) = I(Set) = -\sum_{i=1}^{k} P(value_i) \cdot \log_2\left(P(value_i)\right)$$

Where $P(value_i)$ Is the probability of getting the i^{th} value when randomly selecting one from the set.

2. **Information Gain** – The Information Gain metric uses entropy to determine the best feature for creating a split of the tree[25]. The metric shows how much uncertainty of the dataset is reduced by splitting on a particular attribute.

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Where $Values(A)$ is the all possible values for attribute A, and S_v is the subset of S for which attribute A has value v.

In the first step, the algorithm calculates the entropy for the entire dataset. In the second step, the algorithm calculates the entropy for each attribute for its categorical values (features of the attribute) and the Information Gain for the attribute. In the third step, the algorithm chooses the attribute with the largest information gain to be the node. The algorithm repeats these three steps until the final tree is obtained.

Application Example No. 1 - Predicting the early sign of diabetes using ID3 as a data model

Background on Diabetes

Diabetes is a chronic metabolic disease characterized by high blood sugar (glucose) concentration. According to the Centers for Disease Control and Prevention (CDC), the complications of diabetes can cause heart disease, nerve damage, oral health damage, vision loss, chronic kidney disease, leg damage, hearing loss, and mental health damage.

It is crucial to identify it early to optimize its treatment and prevent its development and the appearance of collateral damage.

According to the World Health Organization (WHO), the number of patients with chronic diseases is increasing rapidly, and these diseases are among the leading causes of death worldwide. Among the deaths from chronic diseases, diabetes is a major factor[28]. According to the International Diabetes Federation (IDF), 1 out of 10 adults aged 20-79, who make up 537 million people live with diabetes; in 2021, diabetes was the cause of about 6.7 million deaths (on average, one every 5 seconds) and from an economic point of view - the financial expenses following the disease of diabetes reached about 966 billion dollars, an increase of 316% in the last 15 years.

We live in a developing technological era based on vast amounts of data, which allows us innovative ways to deal with challenges in various fields. In this part of the article, we will demonstrate the use of the ID3 algorithm for classification and prediction, which will help identify signs of diabetes in its early stages and improve its treatment, as presented in the article "Predicting the early sign of diabetes using ID3 as a data model"[29].

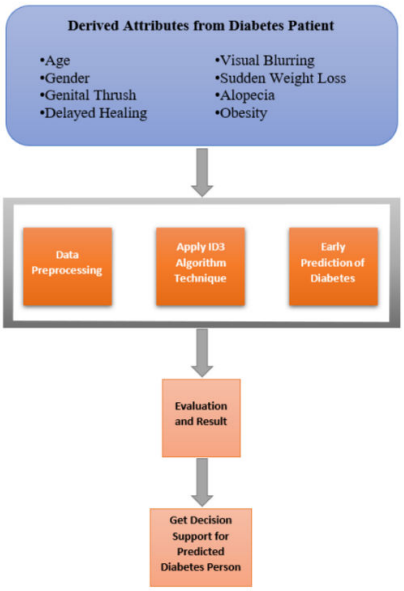
Research Method

This study combined data mining techniques and the prediction model of the early signs of diabetes. For predicting and analyzing the result, the researcher used the ID3 algorithm. The dataset contains 520 patient records that were divided into two parts: training - to build the model and testing - to evaluate the degree of accuracy of the built model. The dataset used for

this study was taken from Kaggle, an online community of data scientists and machine learning practitioners that allows users to find and publish datasets.

In decision trees, each node represents an attribute. To diagnose diabetes, one must consider several attributes according to which the analysis and prediction are performed. The article's author considered the following attributes: age, gender, sudden weight loss, genital thrush, visual blurring, delayed healing, baldness (Alopecia), and obesity. The structure of the system used for this study contains three parts:

- **Data pre-processing** - in this part, to achieve good results, several processes must be performed: data cleaning, data reduction, and data transformation. In the data transformation, the researchers turned the single figure of age into a figure of range.
- **Application of the ID3 algorithm** - building the decision tree by selecting the best features with the help of entropy and information gain indices.
- **Prediction** - evaluating the features based on the ID3 algorithm after the calculations. At this stage, the researchers predict results based on training and testing data stored as CSV files in their java folder. Using the java program, they type a line of code and receive another CSV file with the predicted results (output).



The architecture of the diabetes prediction system.

Results

The output file with the predicted results indicates the following statements:

- Of the test data, it is likely that six patients will develop diabetes.
- Patients in the 31-40 and 41-50 age groups are at a higher risk of developing diabetes
- Sudden weight loss and delayed healing are the prominent features that can be signs of diabetes and were selected to be the top nodes in the tree.

Test Data

Age	Gender	sudden weight loss	Genital thrush	visual blurring	delayed healing	Alopecia	Obesity	class
41-50	Male	Yes	No	No	Yes	Yes	Yes	
31-40	Male	No	No	No	Yes	No	No	
>70	Male	No	Yes	No	Yes	Yes	No	
51-60	Female	Yes	Yes	No	No	Yes	No	
41-50	Female	No	No	Yes	Yes	No	Yes	
31-40	Male	No	Yes	No	No	Yes	No	
31-40	Male	No	No	No	Yes	Yes	No	
31-40	Male	Yes	No	No	Yes	No	Yes	
41-50	Female	Yes	No	No	Yes	No	No	
20-30	Female	No	No	No	Yes	No	No	

Prediction Result

Negative
Positive
Negative
Positive
Negative
Positive
Negative
Positive
Positive
Positive

Test data and prediction results.

In this article, the researcher did not provide numerical data regarding the accuracy percentages of the built model but compared it to other methods used in the research of diabetes prediction[30] and noted that the predicted result is more accurate using the ID3 decision tree. With the help of the predicted results, medical professionals can make important decisions related to the diagnosis and treatment of diabetes.

4. C4.5

Introduction

The C4.5 algorithm ("C" – written in the C programming language, "4.5" – spec version) was developed by Ross Quinlan as an extension of the ID3 algorithm to accommodate both numerical and categorical data. In addition, the C4.5 Algorithm can also handle noise and missing data problems. The formation of rules in the C4.5 algorithm starts with calculating the Entropy and Information Gain, as in ID3. Where it differs from ID3 is by calculating the Information Gain Ratio and SplitInfo metrics.

The IGR evaluates the various attributes' information values and determines the best split attribute. Then the decision tree is generated with a depth-first strategy. Similarly, every tree node initializes its information and produces child nodes (Peng et al., 2017).

The IGR is a modification of information gain to reduce feature bias towards attributes with many branches. The gain ratio is large if the data is spread evenly, and the value will be small if all data enters one branch.

$$SplitInfo\ (S, S_i) = -\sum_{i=1}^{k} p\ (v_i \mid S)\ log_2\ p\ (v_i \mid S)$$

$$\text{GainRatio(A)} = \frac{\text{Gain(A)}}{\text{SplitInfo(A)}}$$

When the above calculations are performed, the attribute with the highest IGR is selected as the root or split attribute. The process is repeated for every branch until all the branches' cases have the same class.

Application Example No. 2 - Rule Formation Application based on C4.5 Algorithm for Household Electricity Usage Prediction[31]

Background on Electricity Usage

The growth of the population, coupled with the rapid technological advances in all fields, causes an increase in energy consumption, specifically electrical energy. Electricity is the

most needed energy because it is easy to distribute and convert into other energies and hence is vital for modern society.

In Indonesia, household electricity use has increased significantly from 52 % in 2001 to 95 % in 2017. Since raw, electricity-producing materials are limited, Indonesia's government is holding an energy-saving culture campaign to reduce consumption.

In addition to saving electricity, an energy management system (EMS) combined with the Plan-Do-Check-Act (PDCA) model has been applied to optimize the energy used in small and medium-sized enterprises. The PDCA concept utilizes existing models in data mining or machine learning to build a decision support system for energy management.

This study develops a rule-forming software application based on the C4.5 algorithm to predict the electricity usage of **personal households** in the city. The study's criteria for predicting household electricity consumption are family size, monthly income, and electrical power.

Research Methodology

Data – sample data of 3,800 household electricity users in Ternate, Indonesia.
This data consists of the following features: The number of electronic equipment, the number of family members, electric power, house area, and monthly income (marked as X#). The class represents the household electricity kWh usage.

The data is then transformed into categories as follows:

No.	Criteria	Criteria value	Transformation results
1	X1	≤ 3	Little
		4 – 6	Medium
		> 6	Many
2	X2	≤ 80	Small
		81 – 150	Are
		> 150	Great
3	X3	≤ *Rp.* 1.500.000	Low
		Rp. 1.500.001 – *Rp.* 2.500.000	Are
		Rp. 2.500.001 – *Rp.* 3.500.000	High
		> 3.500.000	Very high
4	X4	450 VA	Low
		900 VA	Are
		1300 VA	High
		2200 VA	Very high
5	X5	≤ 3	Little
		4 – 6	Medium
		> 6	Many

Table 4 Transformed class or target.

Class/label	Electrical Power	Class grades	Transformation results
Electricity Usage	450 VA	≤ 75kWh	electricity saving
		> 75 kWh	non-electricity saving
	900 VA	≤ 115 kWh	electricity saving
		> 115 kWh	non-electricity saving
	1300	≤ 201 kWh	electricity saving
		> 201 kWh	non-electricity saving
	2200	≤ 358 kWh	electricity saving
		> 358 kWh	non-electricity saving

Rule Formation - The first stage is finding the **entropy** value. After that, the **gain value** of all the features in the household electricity usage data is calculated. The feature with the highest gain value is defined as the **root node**.

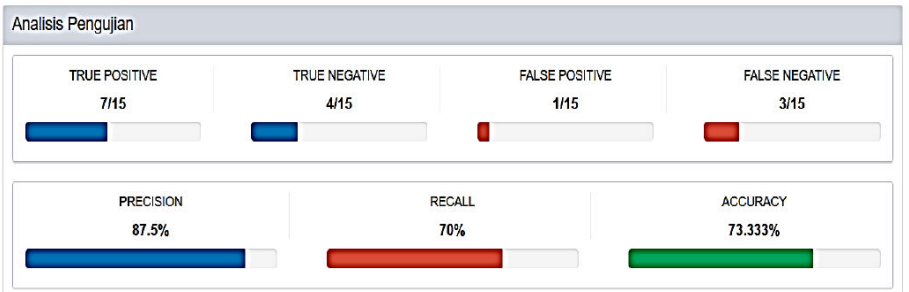
After the root node is determined by calculating the overall gain of the features, the **SplitInfo** and gain ratio is further estimated to define the tree branches.

Based on the results of the calculation, the highest gain value is on the **X5** criterion (i.e., **the number of electronic equipment**) with a value of 0.153, and hence, the root node is electronic equipment. To determine the next node, the calculation is done each time again.

Table 6 Formation of rules.

Number	Rule
1	If (Number of electronics = little and number of family members = little) Then electricity saving
2	If (Number of Electronics = little and number of family members = medium and electric power = low) Then electricity saving
3	If (Number of Electronics = medium and electric power = very high) Then non-electricity saving
4	If (Number of Electronics = medium and electric power = low and number of family members = little) Then non-electricity saving
5	If (Number of Electronics = medium and electric power = low and number of family members = medium) Then electricity saving
6	If (Number of Electronics = medium and electric power = low and number of family members = many) Then electricity saving
7	If (Number of Electronics = medium and electric power = high and monthly income = are) Then electricity saving
8	If (Number of Electronics = medium and electric power = high and monthly income = high) Then non-electricity saving
9	If (Number of Electronics = medium and electric power = high and monthly income = very high) Then electricity saving
10	If (Number of Electronics = medium and electric power = high and monthly income = low and number of family members = low) Then non-electricity saving
11	If (Number of Electronics = medium and electric power = high and monthly income = low and number of family members = medium) Then electricity saving
12	If (Number of Electronics = many and electric power = low) Then non-electricity saving
13.	If (Number of Electronics = many and electric power = are) Then non-electricity saving
14.	If (Number of Electronics = many and electric power = high) Then non-electricity saving

Algorithm Performance Testing - After a rule is formed and calculated by the C4.5 algorithm, this system also conducts tests to see the performance of the C4.5 algorithm by implementing a confusion matrix.



Performance Results

The test was carried out five times to see the differences and performance of the test model with different data amounts.

Table 6 system performance test comparison results.

No	Amount of training data	Number of rules	Number of test data	System performance results (%) Precision	Recall	Accuracy
1	15	6	55	69.44 %	67.56 %	58.2 %
2	25	13	45	64.29 %	69.231%	60 %
3	35	18	35	72.73 %	69.57 %	62.86 %
4	45	10	25	65 %	86,675	64 %
5	55	14	15	87.5 %	70 %	73.33 %
6	1000	14	2800	90.3 %	74.4 %	76.22 %

Authors Conclusions

An important conclusion that the authors wrote was that the amount of applied training data strongly influenced the number of formed rules by applying the C4.5 algorithm.

5. Conclusions

In this article, we presented an overview of decision tree classification and prediction algorithm, alongside application examples in different areas of life, such as medicine and energy, which achieved good prediction results. We can see that many researchers have used different algorithms to predict and improve performance in various fields of life, where research provides them with enormous opportunities. An essential practical conclusion from the articles presented is that a larger set of training data is important for the prediction's accuracy. The subject of machine learning has improved multiple aspects of our lives with the help of practical ways to improve effectiveness, decision-making, cost and time savings, accurate prediction based on data sets, and finding new patterns and relationships in large data sets. Decision tree algorithms such as ID3 and C4.5 remain two of the most common classification algorithms. Many researchers continue to improve and develop these algorithms to optimize them further. These algorithms still face various limitations, so it is necessary to continue studying the subject to improve them and achieve even better results. We hope that we have made the information accessible and opened a window for anyone who wants to get to know the subject and enter the field.

6. Future Work

Future research can continue to explore how the use of these algorithms can be expanded. To continue optimizing and improving the prediction results, increasing the depth and number of data sets used to build the models is necessary. We also think it should be examined whether there is a relationship between the number of features in the data set used for classification and between the percentage of data distribution for training and testing that affects the accuracy percentages of the various algorithms.

7. References

[1] A.M. Almasoud, H.S. Al-Khalifa, A. Al-Salman, "Recent developments in data mining applications and techniques," in 2015 Tenth International Conference on Digital Information Management (ICDIM) (2015), pp. 36-42

[2] Wu, Xindong; Kumar, Vipin; Ross Quinlan, J.; Ghosh, Joydeep; Yang, Qiang; Motoda, Hiroshi; McLachlan, Geoffrey J.; Ng, Angus; Liu, Bing; Yu, Philip S.; Zhou, Zhi-Hua (2008-01-01). "Top 10 algorithms in data mining". Knowledge and Information Systems

[3] C. Anuradha, T. Velmurugan A data mining based survey on student performance evaluation system 2014 IEEE International Conference on Computational Intelligence and Computing Research (2014), pp. 1-4

[4] Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81–106

[5] L. Yi-bin, W. Ying-ying, R. Xue-wen, Improvement of ID3 algorithm based on simplified information entropy and coordination degree in 2017 Chinese Automation Congress (CAC) (2017), pp. 1526-1530

[6] Z. Wang, Y. Liu, L. Liu, A new way to choose splitting attribute in ID3 algorithm, in 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (2017), pp. 659-663

[7] V.K. Soni, S. Pawar, Emotion based social media text classification using optimized improved ID3 classifier, in 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (2017), pp. 1500-1505

[8] Bernal, Iván & Herrera Miranda, Israel & Hernández Hernández, Jose Luis & Molina Angel, Felix. (2019). Implementation of the ID3 algorithm for the generation of a decision tree with food health data from the State of Guerrero, Mexico. 10.35429/JMQM.2019.4.3.1.8.

[9] Shuo Yang, Jing-Zhi Guo, Jun-Wei Jin, An improved Id3 algorithm for medical data classification, Computers & Electrical Engineering, Volume 65, 2018, Pages 474-487

[10] Yan Li, Dandan Jiang, Fachao Li, The Application of Generating Fuzzy ID3 Algorithm in Performance Evaluation, Procedia Engineering, Volume 29, 2012, Pages 229-234
[11] Quinlan, J. R. C4.5: *Programs for Machine Learning*. Morgan Kaufmann Publishers, 1993.
[12] Ian H. Witten; Eibe Frank; Mark A. Hall (2011). "Data Mining: Practical machine learning tools and techniques, 3rd Edition". Morgan Kaufmann, San Francisco. p. 191.
[13] Umd.edu - Top 10 Algorithms in Data Mining
[14] K. Adhatrao, A. Gaykar, A. Dhawan, R. Jha, V. Honrao, Predicting Students' Performance using ID3 and C4.5 Classification Algorithms, International Journal of Data Mining & Knowledge Management Process (IJDKP), 3 (5) (2013)
[15] Jothikumar, R. & Balan, Siva. (2016). C4.5 classification algorithm with back-track pruning for accurate prediction of heart disease. 2016. S107-S111.
[16] F. Chen, X. Li, L. Liu, Improved C4.5 decision tree algorithm based on sample selection, in 2013 IEEE 4th International Conference on Software Engineering and Service Science (2013), pp. 779-782
[17] M.A. Muslim, S.H. Rukmana, E. Sugiharti, B. Prasetiyo, S. Alimah, Optimization of C4.5 algorithm-based particle swarm optimization for breast cancer diagnosis, J, Phys.: Conf. Ser., 983 (1) (2018), p. 012063
[18] A. Cherfi, K. Nouira, A. Ferchichi, Very Fast C4.5 Decision Tree Algorithm, Journal of Applied Artificial Intelligence, 32 (2) (2018), pp. 119-139, 2018
[19] Z. Yuan, C. Wang, An improved network traffic classification algorithm based on Hadoop decision tree, Presented at the 2016 IEEE International Conference of Online Analysis and Computing Science (ICOACS), (2016), pp. 53-56
[20] X. She, T. Lv, X. Liu, The Pruning Algorithm of Parallel Shared Decision Tree Based on Hadoop, Presented at the 2017 10th International Symposium on Computational Intelligence and Design (ISCID) (2017), pp. 480-483
[21] P. Chandrasekar, K. Qian, H. Shahriar, P. Bhattacharya, Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing, 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC) (2017), pp. 481-484
[22] BA Tama. Data mining for predicting customer satisfaction. J. Theor. Appl. Inf. Technol. 2015; 75, 3-7.
[23] AH Mohammad. Comparing two feature selection methods (Information gain and gain ratio) on three different classification algorithms using Arabic dataset. J. Theor. Appl. Inf. Tech. 2018; 96, 1561-9.
[24] P Gu and Q Zhou. Student performances prediction based. Emerg. Comput. Inf. Technol. Educ. 2012; 146, 1-8.
[25] DECISION MAKING USING ID3 ALGORITHM Mary Slocum* M.S. Program in Computer Science, Rivier University
[26] Study and Analysis of Decision Tree Based Classification Algorithms, Harsh Patel, California State University, Fullerton, Purvi Prajapati Charotar University of Science and Technology
[27] Building Classification Models: ID3 and C4.5
[28] A Review of Recent Advances for Preventing, Diagnosis and Treatment of Diabetes Mellitus using Semantic Web

[29] Predicting the Early Sign of Diabetes using ID3 as a Data Model
[30] For example, KNN, and Naïve Bayes - D. Shetty, K. Rit, S. Shaikh, and N. Patil, "Diabetes disease prediction using data mining," in 2017 International Conference on Innovations in Information, Embedded and Communication Systems, (ICIIECS), 2017, pp. 1–5, doi: 10.1109/ICIIECS.2017.8276012.
[31] Firman Tempola*, Miftah Muhammad, Abdul Kadir Maswara and Rosihan Rosihan, Department of Informatics, Khairun University, North Maluku 97719, Indonesia, June 2021