Indonesian Journal of Artificial Intelligence and Data Mining (IJAIDM)
Vol. 5, No. 2, September 2022, pp. 66 – 75
p-ISSN: 2614-3372 | e-ISSN: 2614-6150

Potential for Improvement of Student's English Language with the C4.5 Algorithm

[1]Cyntia Lasmi Andesti, [2]Fitria Lonanda, [3]Nur Azizah
[1]Departement of Informatics, STMIK Indonesia Padang
[2]Departement of Information System, STMIK Indonesia Padang
[3]Departement of Digital Business, STMIK Indonesia Padang
Email: [1]cyntia@stmikindonesia.ac.id, [2]fitria@stmikindonesia.ac.id, [3]nur.azizah@stmikindonesia.ac.id

Article Info
Article history:
Received Jun 08th, 2022
Revised Aug 30th, 2022
Accepted Sep 20th, 2022

Keyword:
C4.5 Algoritm
Classification
Data Mining
English Language
Weka

ABSTRACT
Proficiency in English is not a barrier for the Millennial Generation today. Sophisticated technology can also help to increase proficiency in English. However, there are still many who do not use this technology to support English proficiency. Apart from not using technology, the millennial generation is also lacking in practicing English in everyday life. There are several factors that can predict the potential for increasing proficiency in English, namely Reading (C1), Practice (C2), Pronunciation (C3), Environment (C4), Technology (C5), English Club (C6), and Listening (C7). These factors become parameters in solving problems that occur. These parameters are used in Data Mining method, namely Classification C4.5 or what is often called C4.5 Algorithm. This study aims to determine the potential for increasing proficiency in English. The data processed in this study were 90 respondents from the results of questionnaire data distributed. The software used in the processing is WEKA 3.8.6 Software. The processing steps are to calculate the Entropy value and Gain value of each attribute, form the root node (node) based on the highest gain value and form a decision tree. The results of the discussion on the Weka 3.8.6 software, the data accuracy rate is 90 % or 81 data and the error rate is around 10 % or 9 Data. From the data of 90 respondents, the factors that influence the potential for increasing proficiency in English are Practice (C2).

Corresponding Author:
Cyntia Lasmi Andesti,
Departement of Informatics,
STMIK Indonesia Padang,
Jl. Khatib Sulaiman Dalam No.1 Padang, Sumatera Barat.
Email: cyntia@stmikindonesia.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v5i2.17333

1. INTRODUCTION

The nation's Millennial generation is now increasingly advanced. Various facilities that can be used in communication. In communicating, a good and correct language of instruction is needed. Besides Indonesian as the language of instruction, English should also be adopted more fluently by today's generation. English is also the most dominant international language. In addition, English is also one of the compulsory subjects in every State and Private Universities. This shows that English is one of the indicators of one's success in supporting a career in the world of work. The results of observations found in the field, are still minimal in the use of English to support communication. Therefore, it is necessary to investigate what factors can increase the potential for proficiency in English. One of the causes of the lack of use of English is the arrangement between English and Indonesian. Of course, making English the biggest scourge in its use.

Various branches of computer science that can be used in solving complex problems such as Data Mining [1], Expert Systems [2], Fuzzy Logic [3], Artificial Neural Networks [4] and others [5]. Based on the problems being faced, the researcher uses a classification algorithm in solving problems to predict the potential

for increasing English language for STMIK Indonesia Padang students, especially for class 1 INF, 4 SI A and 4 SI B. The predictions are made using the data mining classification method, namely the C45 algorithm [6]. The C4.5 algorithm data is taken based on the respondent's data to classify which factors have the most influence on the potential for increasing English language in students. The use of C4.5 Classification is in the form of levels from Root to Leaf in a Hierarchical manner. Many related researchers have used the C4.5 Algorithm to solve problems faced by researchers, among others, which was carried out with the research title A Modified C4.5 Classification Algorithm: With The Discretization Method in Calculating the Goodness Score Equivalent which is done by author J. S. Mapa, A. Sison, and R. P. Medina . In this study it is explained that The experimental results reveal that the C4.5 algorithm is an effective solution in calculating the goodness score as well as being the best solution compared to the traditional search score formula [7].

In a study entitled Classification of Diabetes Dataset with Data Mining Techniques by Using WEKA Approach which is done by author K. Alpan and G. S. Ilgi explained about This study discusses that there are many health problems that can be solved by the classification method because the classification method produces knowledge accurately and efficiently [8]. The results show that all data mining methods work with an accuracy rate of 86% and K-NN performs the highest accuracy of 98.07%, thus an effective classification method for solving the problem. In a study about C4.5 Decision Tree against Neural Network on Gait Phase Recognition for Lower Limp Exoskeleton Two IMUs and two FSR sensors are used for detecting human leg motion. The algorithms are trained by 3 different sizes of training pattern. The test dataset of 8 different walking speeds are used to verify the algorithms. The average percentages of CSR are compared among training dataset and algorithms[9]. The results show that C4.5 algorithm yield higher success rate than NARX at 10,000 training patterns and larger. Based on this, it is hoped that the goal can be achieved, namely to analyze the prediction of the potential for English Language improvement that can contribute to students in analyzing the attributes that most influence the improvement in English.

2. RESEARCH METHOD

2.1 Literatur Review

2.1.1 Data Mining

Data mining is a procedure designed to extract added value from a data set in the form of: previously unknown information. Therefore, data mining becomes the basic technique turn the data into information. Data mining is a data analysis technique based on statistical application which aims to extract information. According to Han and Kamber (2006), data mining is defined as "analysis" of observational data sets to find unexpected relationships and to summarize data in new ways that are understandable and useful to data owners. Data mining techniques fall into two categories: supervised learning and unsupervised learning, and data mining algorithms can be grouped into five categories: clustering, association rules, outlier detection, classification, and prediction[10]. With data mining, large amounts of data can be used for other useful things information. Information and knowledge obtained can be used in many fields, such as business management, education, health and so on[11]. Data mining can perform work such as estimating and classifying group data[12].

Data mining is also known as pattern recognition which is an algorithm used for new knowledge that comes from old data, the result of data processing in order to find hidden patterns from the processed data. Data mining has several functions, namely[5]:

a) Function description (Description)
b) Functiom estimation (estimation)
c) Function prediction (Prediction)
d) Function classification (Classification)
e) Function groupping (Classification)
f) Function assocation (Association)

Data mining actually has long roots from fields of science such as artificial intelligence, machine learning, statistics and databases [11]. The various data mining techniques are [11] (Figure 1).

There are many techniques in data mining that can provide solutions in solving a problem in various fields. In this study, the data mining technique used is predictive. which predicts the potential for improving students' English skills.

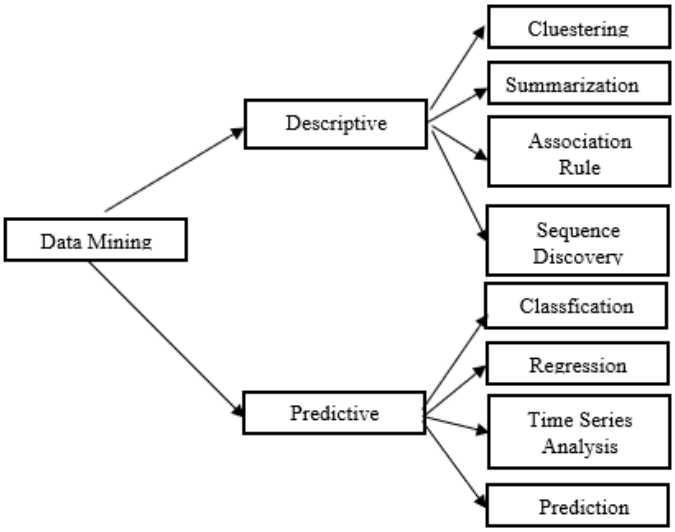

Figure 1. Data Mining Techniques

2.1.2 Knowledge Discovery In Databases (KDD)

Knowledge Discovery In Databases (KDD) is the entire non-trivial process of finding and identify patterns (patterns) in the data, where patterns found are valid, new, can be useful and understandable[11]. KDD is related to technique integration and scientific discovery, interpretation and visualization of the patterns of a number of data sets[11]. The following are the stages of KDD [11]:

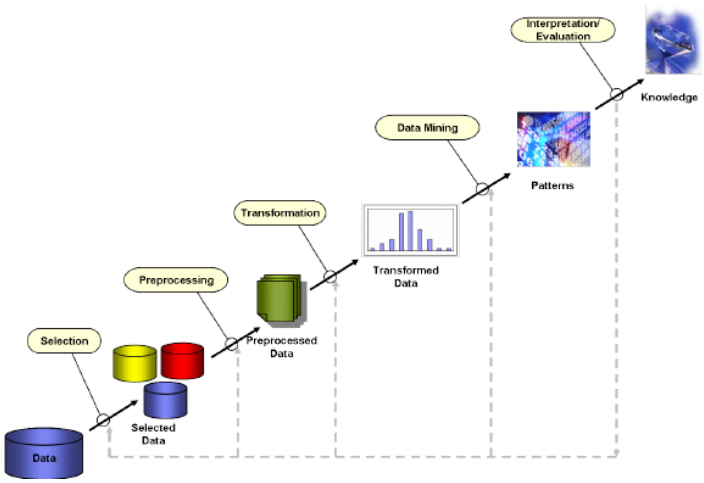

Figure 2. KDD Stages

The most difficult and time consuming part of the KDD system is the data preparation process. In one study it was shown in a study that 60% of the time for compiling a KDD system goes into preparing data for mining while the actual mining process consumes only 10%[10].

2.1.3 C4.5 Algorithm

The C4.5 algorithm is considered a very helpful algorithm in classifying data because the characteristics of the classified data can be obtained clearly, both in the form of structure decision trees and if-then rules, making it easier for users to diginformation on the relevant data. The selection of attributes also greatly influences the Algorithm processing C4.5 because the decision is very dependent on the selected attribute [13]. The C4.5 algorithm uses the J48 rule which determines the related attributes in dealing with the problem and at the same time managing the pruning. Utilization of the C4.5 Algorithm using the Divide and Conquer method to determine the decision tree in the application of the binaryization process to the data so that there is an increase in the estimated error and speed [14]. Decision Tree is used as a reasoning procedure to get a solution to the problem at hand. The tree structure in the C4.5 Algorithm contains internal nodes that describe the attributes, each branch represents the results of the tested attributes and the leaf describes the class.

Many researchers have used the C4.5 algorithm to solve the problems they faced, including those carried out with the research title RapidMiner and Weka Analysis in Predicting the Quality of Employee Performance Using the C4.5 Algorithm Method. This researcher explains that the C4.5 algorithm is expected to be able to provide solutions and can help predict the quality of employee performance[15]. The use of the c45 algorithm has the concept of entropy which selects attributes recursively until no branches are tested [16]**.**

The Entropy formula in the C4.5 Algorithm is as follows [17]**:**

$$Entropy\ (S) = \sum_{i=0}^{n} - pi * \ log_2\ p \quad (1)$$

Explanation:

S	: Case Collection
N	: Number of Partitions S
Pi	: Proportion of Si to S

The formula to find the Gain value is as follows [17]:

$$Gain(S, A) = Entropy\ S - \sum_{i=0}^{n} \frac{|Si|}{|S|} * Entropy(s) \quad (2)$$

Explanation:

S	: Case Collection
A	: Attributes
N	: Number of Attribute Partitions A
| Si |	: Number of Cases on partition i
| S	: Number of Cases in S
Pi	: Proportion of Si to S

2.1.3 WEKA

Weka is an open source data mining tool which has the advantage that there are algorithms and machine learning, it is not too complicated to use and is up to date so that the Weka application helps companies in the business world, academics and health sector agencies. The success of data mining is based on high-quality data collection, the use of appropriate models and algorithms [18]**.** Weka is able to solve real-world data mining problems, particularly the classification underlying machine learning approaches[11].

2.2 Methodology

The Research Method Is A Way or Procedure With Clear And Systematic Stages With The Purpose Of Carrying Out The Process Of Solving The Problem That Is Being Researched. The stages carried out in this research can be seen in the following Figure 3 [14].

The description below will explain the Figure 3:

a. Data Collection
   Data were obtained from the results of questionnaire respondents distributed in STMIK Indonesia Padang, especially 2SIA and 1INFA students. The tables included depend on the data to be used in this study, namely Reading, Practice, Pronounciation, Environment, Technology, English Club, and Listening.

b. Pre-Processing
   At this stage, after the data is collected, the preprocessing process is carried out which will be used for testing with the WEKA Application.

c. Data Cleaning
   Data cleaning is carried out with the aim of cleaning data so that when all attributes are selected they do not experience a missing value, missing value or redundancy value because the results of the data mining process can be clean from errors in data processing.

d. Data Integration
   After cleaning the data, the cleaned data is integrated into one file as a dataset that will be used in the data mining analysis process, namely the predictive potential factor data for increasing English

Language. In this study, data integration from a number of tables involved was carried out using Structured Query Language (SQL) and the results were stored in a single file in CSV format.

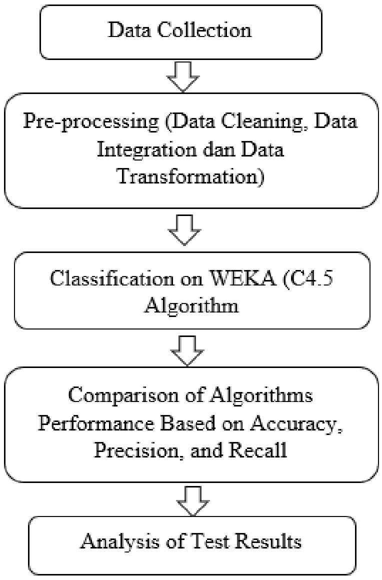

Figure 3. Research Method

e. Data Transformation

The data obtained is based on the respondents' answers and then analyzed so that it can produce information. The information is converted into a data transformation. Data transformation is done to convert data into values with a certain format. Data that has been in a certain format is divided into small groups. Reading data is divided into 3 groups, namely "Very Good" and "Good". Practice is divided into “Can” and “Can't”. Pronounciation is divided into “Influential” and “Less Influential”. Environment is divided into “Yes” and “No”. Media is divided into “Media” and “Not using media”. English Club is divided into “Following” and “Unfollowing” and Listening is divided into “Yes” and “No”.

3. RESULTS AND ANALYSIS

Transformation data that has been changed is data that has been converted through several stages, starting from Data collect, Data Preprocessing, Data Cleaning, Data Integration to the Data Transformation stage. The classification process uses data mining tools WEKA 3.8.6. The data tested using data from STMIK Indonesia Padang students, especially students of 2SI A, 2 SIB and INF1. The training model for work estimation is k-fold cross-validation with a 10-fold value. The J48 method used in the C4.5 algorithm classicization process. The test is carried out using a confusion matrix, which is based on accuracy, precision, and recall value. The data carried out in the training process are 90 records. Using 7 criteria, among others Reading (C1), Practice (C2), Pronunciation (C3), Environment (C4), Tecnology (C5), Listening (C6), and English Club (C7). The results of the performance measurement of each classifier are obtained as follows:

Table 1. Accumulated Data

Reading	Practice	Pronunciation	Environment	Teknologi	Listening	Englishclub	Potency
Very Good	Can	Influential	Conducive	Media	Yes	Follow	Increase
Very Good	Can	Influential	Conducive	Media	Yes	Follow	Increase
Very Good	Can	Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can	Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can	Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can	Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can	Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can	Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can	Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can	Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can	Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can	Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can	Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can	Influential	No kondusif	No Media	Yes	No Follow	Decrease

Reading	Practice	Pronunciation	Environment	Teknologi	Listening	Englishclub	Potency
Good	Can	Influential	Conducive	No Media	Yes	No Follow	Increase
Good	Can	Influential	Conducive	No Media	Yes	No Follow	Increase
Good	Can	Influential	Conducive	No Media	Yes	No Follow	Increase
Good	Can	Influential	Conducive	No Media	Yes	No Follow	Increase
Good	Can	Influential	Conducive	No Media	Yes	No Follow	Increase
Good	Can	Influential	Conducive	No Media	Yes	No Follow	Increase
Good	Can	Influential	Conducive	No Media	Yes	No Follow	Increase
Good	Can	Influential	Conducive	No Media	Yes	No Follow	Increase
Good	Can	Influential	Conducive	No Media	Yes	No Follow	Increase
Good	Can’t	Influential	Conducive	No Media	Yes	No Follow	Decrease
Good	Can’t	Influential	Conducive	No Media	Yes	No Follow	Decrease
Good	Can’t	Influential	Conducive	No Media	Yes	No Follow	Decrease
Good	Can’t	Influential	Conducive	No Media	Yes	No Follow	Decrease
Good	Can’t	Influential	Conducive	No Media	Yes	No Follow	Decrease
Good	Can’t	Influential	Conducive	No Media	Yes	No Follow	Decrease
Good	Can’t	Influential	Conducive	No Media	Yes	No Follow	Decrease
Good	Can’t	Influential	Conducive	No Media	Yes	No Follow	Decrease
Good	Can’t	Influential	Conducive	No Media	Yes	No Follow	Decrease
Good	Can’t	Influential	Conducive	No Media	Yes	No Follow	Decrease
Very Good	Can’t	Less Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can’t	Less Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can’t	Less Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can’t	Less Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can’t	Less Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can’t	Less Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can’t	Less Influential	No kondusif	No Media	Yes	No Follow	Decrease
Very Good	Can’t	Less Influential	No kondusif	No Media	Yes	No Follow	Increase
Very Good	Can’t	Less Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can’t	Less Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can’t	Less Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can’t	Less Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can’t	Less Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can’t	Less Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can’t	Less Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can’t	Less Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can	Influential	No kondusif	Media	Yes	Follow	Increase
Very Good	Can	Influential	No kondusif	Media	Yes	Follow	Increase
Very Good	Can	Influential	No kondusif	Media	Yes	Follow	Increase
Very Good	Can	Influential	No kondusif	Media	Yes	Follow	Increase
Very Good	Can	Influential	No kondusif	Media	Yes	Follow	Increase
Very Good	Can	Influential	Conducive	Media	Yes	Follow	Increase
Very Good	Can	Influential	Conducive	Media	Yes	No Follow	Increase
Very Good	Can	Influential	Conducive	Media	Yes	No Follow	Decrease
Very Good	Can	Influential	Conducive	Media	No	Follow	Increase
Very Good	Can	Influential	Conducive	Media	No	Follow	Decrease
Very Good	Can	Influential	Conducive	No Media	Yes	Follow	Increase
Very Good	Can	Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can	Influential	No kondusif	Media	Yes	Follow	Increase
Very Good	Can	Influential	No kondusif	Media	Yes	Follow	Decrease
Very Good	Can	Less Influential	Conducive	Media	Yes	Follow	Decrease
Very Good	Can	Less Influential	Conducive	Media	Yes	Follow	Decrease
Very Good	Can’t	Influential	Conducive	Media	Yes	Follow	Increase
Very Good	Can’t	Influential	Conducive	Media	Yes	Follow	Decrease
Good	Can	Influential	Conducive	Media	Yes	Follow	Increase
Good	Can	Influential	Conducive	Media	Yes	Follow	Decrease
Very Good	Can	Influential	Conducive	Media	Yes	Follow	Increase
Very Good	Can’t	Less Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can’t	Less Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can’t	Less Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can	Influential	Conducive	Media	Yes	Follow	Increase
Very Good	Can’t	Less Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can	Influential	Conducive	Media	Yes	Follow	Increase
Very Good	Can’t	Less Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can	Influential	Conducive	Media	Yes	Follow	Increase
Very Good	Can’t	Less Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can	Influential	Conducive	Media	Yes	Follow	Increase
Very Good	Can	Influential	Conducive	Media	Yes	Follow	Increase
Very Good	Can	Influential	Conducive	Media	Yes	Follow	Increase
Very Good	Can	Influential	Conducive	Media	Yes	Follow	Increase
Very Good	Can	Influential	Conducive	Media	Yes	Follow	Increase
Very Good	Can	Influential	Conducive	Media	Yes	Follow	Increase
Very Good	Can’t	Less Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can	Influential	Conducive	Media	Yes	Follow	Increase

Reading	Practice	Pronunciation	Environment	Teknologi	Listening	Englishclub	Potency
Very Good	Can't	Less Influential	Conducive	No Media	Yes	Follow	Decrease
Very Good	Can	Influential	Conducive	Media	Yes	Follow	Increase
Very Good	Can't	Less Influential	Conducive	No Media	Yes	Follow	Decrease

The preprocessing results obtained are then processed into a decision tree using WEKA software to determine the potential for improving students' English skills, as for the decision tree image as Figure 4.

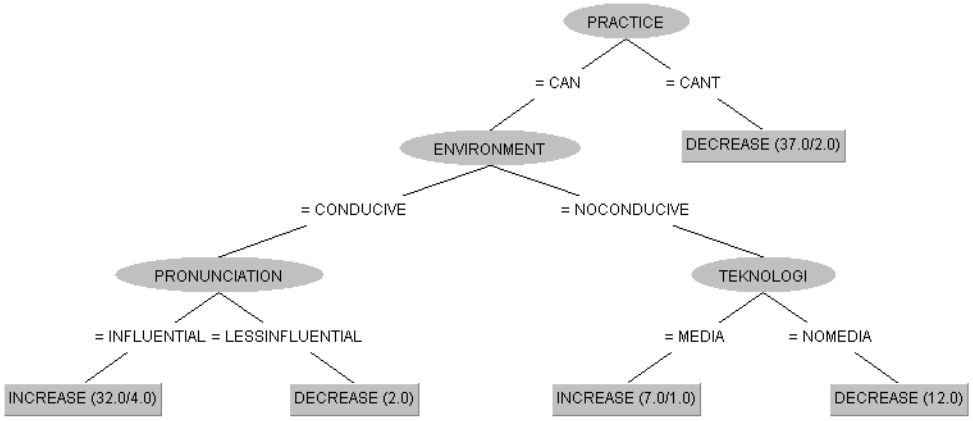

Figure 4. English Improvement Potential Decision Tree

Based on the results of the decision tree, the classification of the potential for improving students' English skills, that the attribute that has the main influence in obtaining predictions of the potential for increasing English proficiency is the practice variable (C2) which occupies the root node as shown in the following figure with The results of the validation using the C4.5 algorithm with the weka tools are shown in the Summary section. The rules that can be formed from the decision tree in Figure 4 are as follows:

"IF Practice= Can AND Environment= Conducive AND Pronounciation= Influential THEN class= Increase"
"IF Practice= Can AND Environment= Conducive AND Pronounciation= Less Influential THEN class= Decrease"
"IF Practice= Can't, THEN class= Decrease"
"IF Practice= Can AND Environment= No Conducive AND Teknologi= Media THEN class= Increase"
"IF Practice=Can AND Environment= No Conducive AND Teknologi= No Media THEN class= Decrease"

From the sample data with 90 records the number of rules formed is 6 rules. The next stage is the processing of data from the results obtained from the C4.5 algorithm. The results of the classification can be seen as Figure 5.

```
Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0.12 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          81               90      %
Incorrectly Classified Instances         9               10      %
Kappa statistic                          0.7964
Mean absolute error                      0.1577
Root mean squared error                  0.3001
Relative absolute error                 32.7757 %
Root relative squared error             61.173  %
Total Number of Instances               90

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.944     0.130     0.829      0.944    0.883      0.802    0.881     0.776     INCREASE
               0.870     0.056     0.959      0.870    0.913      0.802    0.881     0.916     DECREASE
Weighted Avg.  0.900     0.085     0.907      0.900    0.901      0.802    0.881     0.860

=== Confusion Matrix ===
```

Figure 5. Output Classification

Based on these results, the researcher conducted data testing using the WEKA software to see the suitability of the method used with the case being studied. Based on the calculation results obtained an accuracy of 90%. This can explain that the C4.5 method can be applied to the potential for improving students' English skills. The following are the results of calculations using the WEKA software.

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          81               90      %
Incorrectly Classified Instances         9               10      %
Kappa statistic                          0.7964
Mean absolute error                      0.1577
Root mean squared error                  0.3001
Relative absolute error                 32.7757 %
Root relative squared error             61.173  %
Total Number of Instances               90
```

Figure 6. Accuracy Weka Software

4. DISCUSSION

The results shown in Figures 4 and 5 show 90 processed data where the accuracy obtained for all data is 90%. Weka is a software that provides data miners to see the predictions of the potential for increasing English proficiency in students. In addition to the overall accuracy percentage of the classifier, the model output also provides detailed accuracy by class. The detailed accuracy description includes the true positive rate (TP Rate). False positive rate (FP Rate), precision (percentage of instances actually divided by total classified instances), recall (proportion of instances classified as a particular class divided by actual total), F-measure (a measure that combines precision and recall value)). As mentioned earlier, the overall average accuracy of the model is 90%, because there are 2 classes in predicting the potential for improving English proficiency, it is necessary to explore the accuracy of each class. The TP level for the "increasing" class is 94.4% while the TP level for the "decreasing" class is 87% (Table 2 and Table 3).

Table 2. Detailed Accuracy By Class

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.944	0.130	0.829	0.944	0.883	0.802	0.881	0.776	INCREASE
	0.870	0.056	0.959	0.870	0.913	0.802	0.881	0.916	DECREASE
Weighted Avg	0.900	0.085	0.907	0.900	0.901	0.802	0.881	0.860	

Table 3. Confusion Matrix

a	b	Classified as
34	2	INCREASE
7	47	DECREASE

From the information in Table 2 and Table 3, the process of calculating the average percentage of success accuracy and error rate will then be carried out as follows:

$$\text{Accuracy} = \frac{\text{Many Correct Predictions}}{\text{a total of Many Predictions}} = \frac{f11+f00}{f11+f10+f01+f00} \quad (3)$$

$$\text{Error Rate} = \frac{\text{Many Wrong Prediction}}{\text{a total of many wrong predictions}} = \frac{f10+f01}{f11+f10+f01+f00} \quad (4)$$

$$\text{Accuracy} = \frac{34+47}{34+2+7+47} = 0{,}9$$

$$\text{Persentase accuracy} = 0{,}9 \times 100\% = 90\ \%$$

$$\text{Error Rate} = \frac{2+7}{34+2+7+47} = 0{,}1$$

$$\text{Persentase accuracy} = 0{,}1 \times 100\% = 10\ \%$$

Data mining analysis with the C4.5 algorithm can be used on data sets predicting the potential for improving students' English skills which have an accuracy value of 90% while the error rate for the data set is

10%.. This shows that the C4.5 algorithm can be used on the data set of potential students' English proficiency. This research provides 3 contributions to existing research. The first is to apply the KDD model in applying the prediction of the potential for increasing English proficiency in students. The second is to apply 7 attributes as model input. The previous study only used 5 attributes as input for a predictive model for the potential for improving students' English proficiency. The third contribution is to apply three steps of data mining techniques, namely prediction feature selection and classification.

5. CONCLUSION

The research uses a classification algorithm, namely the C4.5 algorithm using a dataset of Informatics and Information Systems students at STMIK Indonesia Padang. This research is only limited to statistical analysis or based on data in the field, so external factors are needed that can be used as additional attributes. In addition, it is necessary to re-identify other attributes that affect the potential factor for increasing proficiency in English in the form of a questionnaire distribution. Based on the results of the research conducted, it can be concluded that the application of the C4.5 classification to the improvement of English proficiency attributes that are used as potential improvements include: Reading, Practice, Pronounciation, Environment, Technology, English Club, and Listening. The results of the calculation of the attributes that have the most influence on the potential for improving English are Practice (C2). Testing was also carried out using the WEKA software and obtained an accuracy of 90 %.

REFERENCES

[1] J. S. Lee, "AUC4.5: AUC-Based C4.5 Decision Tree Algorithm for Imbalanced Data Classification," *IEEE Access*, vol. 7, pp. 106034–106042, 2019, doi: 10.1109/ACCESS.2019.2931865.

[2] S. Chandra, S. Sumijan, and E. P. W. Mandala, "Expert System For Diagnosing Hemophilia In Children Using Case Based Reasoning," *Indones. J. Artif. Intell. Data Min.*, vol. 2, no. 1, pp. 45–51, 2019, doi: 10.24014/ijaidm.v2i1.6681.

[3] J. Nugroho, L. Linawati, and T. Mahatma, "International Journal of Active Learning Analysis of Lecturers Competency Performance Evaluation using Fuzzy Modeling," *Int. J. Act. Learn.*, vol. 4, no. 2, pp. 99–113, 2019, [Online]. Available: https://journal.unnes.ac.id/nju/index.php/ijal/article/view/20025

[4] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. E. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, p. e00938, 2018, doi: 10.1016/j.heliyon.2018.e00938.

[5] Irnanda, "Penerapan Klasifikasi C4.5 Dalam Meningkatkan Kecakapan Berbahasa Inggris dalam Masyarakat," *Semin. Nas. Teknol. Komput. Sains*, pp. 304–308, 2020.

[6] Y. I. Kurniawan, A. Fatikasari, M. L. Hidayat, and M. Waluyo, "Prediction for Cooperative Credit Eligibility Using Data Mining Classification With C4.5 Algorithm," *J. Tek. Inform.*, vol. 2, no. 2, pp. 67–74, 2021, doi: 10.20884/1.jutif.2021.2.2.49.

[7] J. S. Mapa, A. Sison, and R. P. Medina, "A Modified C4.5 Classification Algorithm: With the Discretization Method in Calculating the Goodness Score Equivalent," *ICETAS 2019 - 2019 6th IEEE Int. Conf. Eng. Technol. Appl. Sci.*, pp. 4–7, 2019, doi: 10.1109/ICETAS48360.2019.9117309.

[8] K. Alpan and G. S. Ilgi, "Classification of Diabetes Dataset with Data Mining Techniques by Using WEKA Approach," *4th Int. Symp. Multidiscip. Stud. Innov. Technol. ISMSIT 2020 - Proc.*, 2020, doi: 10.1109/ISMSIT50672.2020.9254720.

[9] A. Thongsook, T. Nunthawarasilp, P. Kraypet, J. Lim, and N. Ruangpayoongsak, "C4.5 Decision Tree against Neural Network on Gait Phase Recognition for Lower Limp Exoskeleton," *2019 1st Int. Symp. Instrumentation, Control. Artif. Intell. Robot. ICA-SYMP 2019*, pp. 69–72, 2019, doi: 10.1109/ICA-SYMP.2019.8646253.

[10] M. M. Ghazal and A. Hammad, "Application of knowledge discovery in database (KDD) techniques in cost overrun of construction projects," *Int. J. Constr. Manag.*, vol. 0, no. 0, pp. 1–15, 2020, doi: 10.1080/15623599.2020.1738205.

[11] C. Elma Purnomo, "Penerapan Metode C4.5 Untuk Klasifikasi Warga Miskin Pada Desa Mengandung Sari," *J. Teknol. dan Sist. Inf.*, vol. 2, no. 3, pp. 14–25, 2021, [Online]. Available: http://jim.teknokrat.ac.id/index.php/JTSI

[12] Nadiah, S. Soim, and Sholihin, "Implementation of Decision Tree Algorithm Machine Learning in Detecting Covid-19 Virus Patients Using Public Datasets," vol. 5, no. 1, pp. 37–43, 2022.

[13] N. Azwanti, "Algoritma C4.5 Untuk Memprediksi Mahasiswa Yang Mengulang Mata Kuliah (Studi Kasus Di Amik Labuhan Batu)," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 9, no. 1, pp. 11–22, 2018, doi: 10.24176/simet.v9i1.1627.

[14] Gunawan, Hanes, and Catherine, "C4 . 5 , K-Nearest Neighbor , Naïve Bayes and Random Forest Algorithms Comparison to Predict Students ' On Time Graduation," vol. 4, no. 2, pp. 62–71, 2021.

[15] A. Penyareng *et al.*, "Analisis RapidMiner Dan Weka Dalam Memprediksi Kualitas Kinerja Karyawan Menggunakan Metode Algoritma C4.5," vol. 9, no. 2, pp. 1655–1665, 2022.

[16] D. Istiawan and L. Khikmah, "Implementation of C4.5 Algorithm for Critical Land Prediction in Agricultural Cultivation Areas in Pemali Jratun Watershed," *Indones. J. Artif. Intell. Data Min.*, vol. 2, no. 2, p. 67, 2019, doi: 10.24014/ijaidm.v2i2.7569.

[17] J. Eska, "Penerapan Data Mining Untuk Prekdiksi Penjualan Wallpaper Menggunakan Algoritma C4.5 STMIK

Royal Ksiaran," *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. 2, pp. 9–13, 2016.
[18] S. Turnip and P. Siltionga, "Analisis Pola Penyebaran Penyakit dengan Menggunakan Algoritma C4.5," *J. Tek. Inform. Unika St. Thomas*, vol. 03, no. 479, pp. 3–7, 2018.

BIBLIOGRAPHY OF AUTHORS

Cyntia Lasmi Andesti was born in Padang and currently a Lecturer from Department of Informatics at STMIK Indonesia Padang. She actively teaches in the fields of Informatic Logics and Digital System. Received Bachelor's Degree in Informatic Engineering Department in Universitas Putra Indonesia "YPTK" Padang and Master's Degree in Informatic Engineering in Universitas Putra Indonesia "YPTK" Padang. The focus of her research is data mining and Artificial Intelligence.

Fitria Lonanda, is currently a Lecturer from Department of Information System at STMIK Indonesia Padang. She teaches actively English for Information Systems, English for Informatics, and English for Business. Received Bachelor's Degree in English Department in Universitas Andalas and Master's Degree in Linguistics in Universitas Gadjah Mada. The focus of her research is Linguistics.

Nur Azizah, is currently a Lecturer from Department of Digital Business at STMIK Indonesia Padang. Received Bachelor's Degree in Information System Department in STMIK Indonesia Padang and Master's Degree in Information System in Universitas Putra Indonesia "YPTK" Padang. The focus of her research is data mining.