Journal of Applied Engineering and Technological Science
Vol 3(2) 2022 : 90-97

APPLICATION OF C5.0 ALGORITHM IN PREDICTION OF LEARNING OUTCOMES IN CALCULUS SUBJECT

Fida Nafisah Giustin1*, Betha Nurina Sari2, Tesa Nur Padilah3
Program Studi Teknik Informatika, Universitas Singaperbangsa Karawang
fida.nafisah17099@student.unsika.ac.id[1*], betha.nurina@staff.unsika.ac.id [2], tesa.nurpadilah@staff.unsika.ac.id[3]

Received : 01 May 2022, Revised: 19 June 2022 , Accepted : 20 June 2022
**Corresponding Author*

ABSTRACT
Calculus is one of the basic subject that must be studied at the computer science faculty of the informatics engineering study program. For some students, especially in the Faculty of Informatics Engineering, calculus is a subject that is considered quite difficult, even though this subject is important for them. And the resulted for some students having to repeat this subject. For this reason, predictions of calculus learning outcomes are carried out by applying the data mining process and using the C5.0 method for the prediction process based on the classification concept that will be carried out. This study applies the Cross Industry Standard Process for Data Mining (CRISP – DM) methodology with the C5.0 algorithm. The results are in the form of a decision tree (Decision tree) and the rules in it using the attributes of guardian, number of family members, status of residence, internet, activity, desire to continue study, the last education of parents (father and mother), parents' occupations, grades on assignments, UAS, and UTS. The C5.0 algorithm is able to predict the results of learning calculus. The evaluation results show that the applied C5.0 algorithm has an accuracy of 95%.

Keywords : *data mining, prediction, C5.0 algorithm, CRISP-DM, student achievements*

1. Introduction

Education is a conscious and planned effort to create a learning atmosphere and learning process so that students actively develop their potential to have religious spiritual strength, self-control, personality, intelligence, noble character and skills needed by themselves, society, nation and state (UU National Education System No. 20 of 2003). Educational institutions are places where learning activities take place with the aim of changing individual behavior for the better through interaction with the surrounding environment (Rahman, K, 2018). One type is formal educational institutions such as universities. Currently there are many universities in Indonesia. One of them is the Singaperbangsa Karawang University.

Singaperbangsa Karawang University is one of the universities located in West Java. Founded on February 2, 1982 by the Pangkal Perjuangan Higher Education Foundation and became a state university on October 6, 2014. There are several study programs running at Unsika, one of which is Informatics Engineering. Informatics Engineering is one of the study programs that focuses on dealing with the problem of transformation or data processing by making optimal use of computer technology through logical processes. In it, there are basic courses that must be studied, namely calculus. For some students, especially in the Faculty of Informatics Engineering, calculus is a subject that is considered quite difficult, even though this course is very important for them. This resulted in some students having to repeat this course.

At computer science faculty Unsika itself, from 2016 -2019 there were still students who got a final score of less than 5 or a quality score below D and repeated the course to improve their calculus scores. In 2016 there were 5 students who repeated the calculus course. In 2017, 6 students repeated the calculus course. In 2018 the number of computer science faculty students who repeated was 4 people and in 2019 as many as 6 students repeated calculus courses from a total of 684 computer science students.

From several previous studies, according to P. Sokkhey & T. Okazaki there are 10 important factors that influence student learning outcomes (Sokkhey, P., Navy, S., Tong, L., & Okazaki, T, 2020). According to Benedict et al, prediction of student academic performance using the decision tree method achieved an accuracy of 71.661% (Benediktus, N., & Oetama, R. S, 2020). According to Zhang X, by using a decision tree on the student dataset, a pattern was found where current learning outcomes depend on previous results (Zhang, X., Xue, R., Liu, B., Lu, W., & Zhang, Y, 2018). According to Ulfi et al, using Decision Tree C5.0 to perform early detection of drop out students with GPA and attendance as attributes for classifying. The results obtained are 93% accuracy with an error rate of 5% (Aesyi, U. S., Lahitani, A. R., Diwangkara, T. W., & Kurniawan, R. T., 2021). Sari predicts student learning outcomes in mathematics, there are 25 variables that are effectively used Sari B. N, 2017). So this research was conducted to predict the calculus learning outcomes of the faculty of computer science in the Informatics Engineering study program by applying the data mining process and using the C5.0 method for the prediction process based on the classification concept to be carried out. The difference between this research and previous research is that the object of research uses the latest data from the calculus course and uses the C5.0 method(Damanik, et al., 2019; Cherfi, et al., 2018; Fabriantono, et al., 2020).

2. Research Method

This research applies the Cross Industry Standard Process for Data Mining (CRISP – DM) methodology. The CRISP – DM method is a standardized data mining methodology compiled by three data mining market initiators, namely Daimler Chrysler (Daimler-Benz), SPSS (ISL) and NCR (Hanin, N. A., 2022). There are 6 stages of this methodology which are shown in Figure 1.

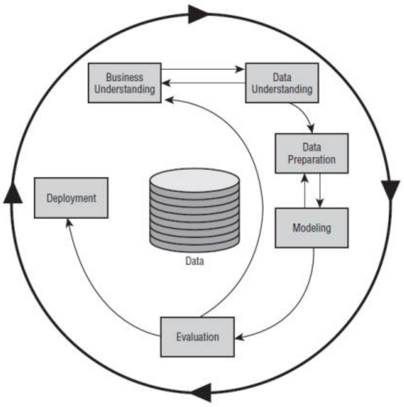


Fig. 1. Research Methodology flow

2.1 Business Understanding

At this stage an understanding of the substance of the data mining activities that will be carried out is carried out as well as determining goals and preparing strategies to achieve these goals (Hanin, N. A., 2022).

2.2 Data Understranding

Data understanding is the process of collecting initial data and studying the data to understand what the data can do. For this reason, student data was collected, both by distributing questionnaires and final grade data from the university.

2.3. Data Preparation

Data preparation is to create a new database that will be used for the data mining process. The data obtained is still in the form of unstructed data, in which the contents of the data still contain noise. Therefore, in the data preparation process, data cleaning is carried out including eliminating data duplication, and correcting errors in data (Fahmi, R. N, 2021).

2.4 Modelling

Modeling is the stage of selecting and applying various modeling techniques and some of the parameters will be adjusted to get the optimal value (Renaldi, D, 2020). The data modeling process uses the C5.0 method.

2.5 Evaluation

Evaluation model is the stage of determining whether the model built is in accordance with the objectives set in the initial phase (Business understanding). At this stage, the evaluation of the model is done using a confusion matrix. Confusion matrix is a method that is usually used to perform accuracy calculations on data mining concepts or Decision Support Systems (Han, J., Kamber, M., & Mining, D, 2006).

2.6 Deployment

Make a report about the knowledge gained from or pattern recognition in the data mining process which is presented in the form of graphs or descriptions that are easy to understand (Renaldi D, 2020).

3. Result and Discussion

The implementation of the C5.0 model is carried out on the data of computer science students from the informatics engineering study program who have taken calculus subject.

3.1 Business Understanding

At this stage, search for material by reading journals related to decision trees with the C5.0 method, journals related to educational data mining, and processing questionnaire data.

3.2 Data Understanding

At the data understanding stage, questionnaires were distributed and final grade data were collected from lecturers who were in charge of the calculus course.

3.3 Data Preparation

Data preparation is the stage of cleaning raw data from noise, cleaning duplicate data, correcting data errors, and data transformation. In this process, the obtained questionnaire data is cleaned and the selected attributes are taken during the validity and reliability testing process. The initial dataset is shown in table 1.

Table 1 – Initial Dataset

Guardian	Family size	Pstatus	Internet	Desire to continue studies
Father	More than 3	Living together	Yes	Yes
Father	More than 3	Living together	Yes	Yes
Mother	Less than 3	Living together	No	No
Father	More than 3	Living together	Yes	No
...	...	...	...	...
Mother	More than 3	Living together	No	No

Activity	Medu	Fedu	Mjob	Fjob
Yes	Associate Degree / Bachelor	Associate Degree / Bachelor	Civil servant / armed forces	Entrepreneur
No	High school equal	Associate Degree / Bachelor	Civil servant / armed forces	Civil servant / armed forces
No	High school equal	High school equal	Housewife	Other
No	Associate Degree / Bachelor	Associate Degree / Bachelor	Civil servant / armed forces	Other
....	....	....	....	....
No	High school equal	Associate Degree / Bachelor	Housewife	Civil servant / armed forces

Assigment	Midterm exam	Exam	NA	Description
53.3	31.0	5.0	27.9	Fail
67.3	28.0	9.0	28.5	Fail
64.7	22.0	12.0	29.3	Fail
59.3	48.0	5.0	31.9	Fail
....	....	....	....	....
77.0	28.0	13.0	33.8	Fail

It can be seen in table 1, there are several attributes that have values with data types that do not match what is needed, for this reason, changes are made to data types and deletion of attributes that are not used in the dataset so as to produce a new dataset that is ready to be used in the modeling process. The latest dataset can be seen in table 2.

Table 2 – Data After Processing

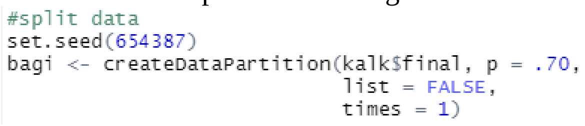
Guardian	Family size	Pstatus	Internet	Desire to continue studies
Father	GT3	T	Yes	Yes
Father	GT3	T	Yes	Yes
Mother	LE3	T	No	No
Father	GT3	T	Yes	No
...	....	....	....	....
Mother	GT3	T	Yes	No

Activity	Medu	Fedu	Mjob	Fjob
Yes	5	5	Civil servant / armed forces	Entrepreneur
No	4	5	Civil servant / armed forces	Civil servant / armed forces
No	4	4	Housewife	Other
No	5	5	Civil servant / armed forces	Other
....	....	....	....	....
No	4	5	Housewife	Civil servant / armed forces

Assigment	Midterm exam	Exam
53.3	31.0	5.0
67.3	28.0	9.0
64.7	22.0	12.0
59.3	48.0	5.0
....	....	....
77.0	28.0	13.0

Table 2 is the final stage of the data preparation process. It can be seen, there are several attributes, namely NA (final grade) and description (pass / repeat) which are deleted and the data type of data on the attributes of Medu (mother's last education) and Fedu (father's last education) is changed from data of type factor to number.

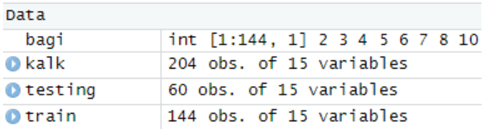
After the dataset is created, the next step is to divide the data into training data and testing data. Splitting dataset is done with the input shown in Figure 2.



```
#split data
set.seed(654387)
bagi <- createDataPartition(kalk$final, p = .70,
                            list = FALSE,
                            times = 1)
```

Fig. 2. Input Splitting Dataset

After inputting into Rstudio, it will produce the output shown in Figure 3.



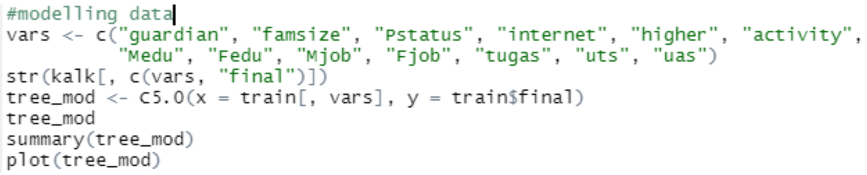
```
Data
  bagi       int [1:144, 1] 2 3 4 5 6 7 8 10
  kalk       204 obs. of 15 variables
  testing    60 obs. of 15 variables
  train      144 obs. of 15 variables
```

Fig. 3. Output Splitting Dataset

In Figure 3, the distribution of data is carried out with a ratio of 70: 30 where the train data has 144 observations with 15 attributes and the testing data has 60 observations.

3.4 Modelling

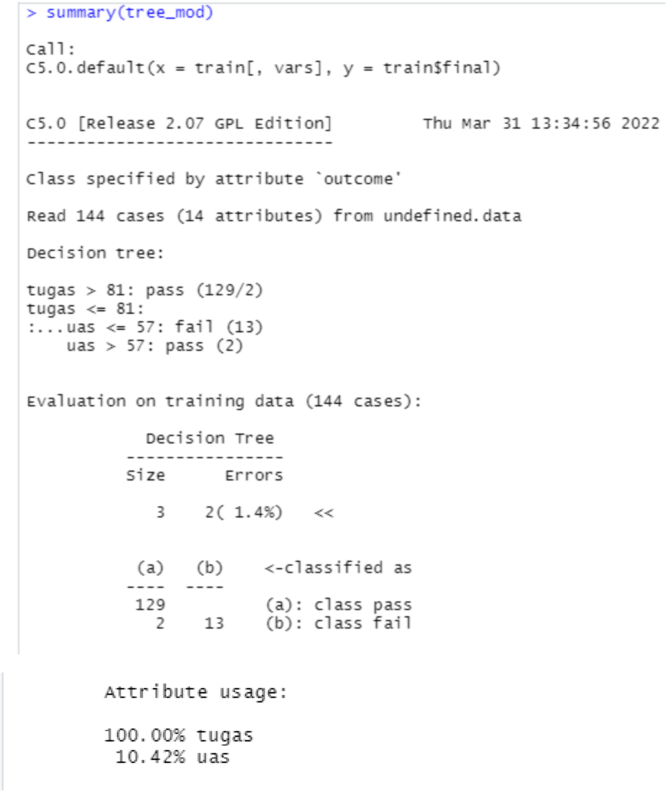
This stage is the process of applying certain methods to the prepared dataset. the method used is the classification of training data using the C5.0 algorithm, data processing is carried out with the help of the Rstudio application where the initial stage is to classify the dataset into pass and repeat classes. Input from modeling with C5.0 can be seen in Figure 4.



```
#modelling data
vars <- c("guardian", "famsize", "Pstatus", "internet", "higher", "activity",
          "Medu", "Fedu", "Mjob", "Fjob", "tugas", "uts", "uas")
str(kalk[, c(vars, "final")])
tree_mod <- C5.0(x = train[, vars], y = train$final)
tree_mod
summary(tree_mod)
plot(tree_mod)
```

Fig. 4. Modelling Dataset

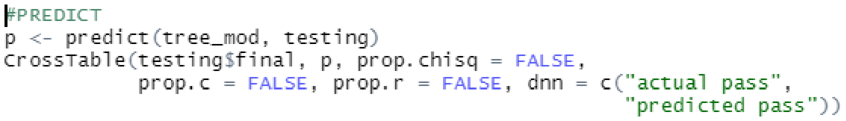
From Figure 4 it can be seen that all attributes are used and the dataset used is training data. Modeling was carried out 3 times using training data with different amounts according to the number of datasets that had been built in the data preparation process. The output generated from the modeling process is shown in Figure 5.



```
> summary(tree_mod)

Call:
C5.0.default(x = train[, vars], y = train$final)


C5.0 [Release 2.07 GPL Edition]         Thu Mar 31 13:34:56 2022
-------------------------------

Class specified by attribute `outcome'

Read 144 cases (14 attributes) from undefined.data

Decision tree:

tugas > 81: pass (129/2)
tugas <= 81:
:...uas <= 57: fail (13)
    uas > 57: pass (2)


Evaluation on training data (144 cases):

            Decision Tree
          ----------------
          Size      Errors

             3    2( 1.4%)   <<


           (a)   (b)    <-classified as
          ----  ----
           129          (a): class pass
             2    13    (b): class fail


        Attribute usage:

        100.00% tugas
         10.42% uas
```

Fig. 5. Output Modelling Dataset

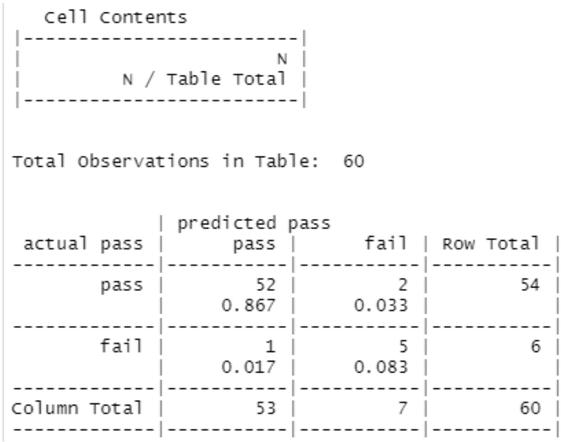
In Figure 5, 144 observations in the training data are applied to build the C5.0 model which produces 129 observations in the graduating class and 13 observations in the repeat class. At this stage, a prediction process is also carried out using the testing data that has been shared made in the data preparation process. Prediction input can be seen in Figure 6.



```
#PREDICT
p <- predict(tree_mod, testing)
CrossTable(testing$final, p, prop.chisq = FALSE,
           prop.c = FALSE, prop.r = FALSE, dnn = c("actual pass",
                                                     "predicted pass"))
```

Fig. 6. Prediciton input

Seen in Figure 6, the results of the modeling that have been carried out on the training data are applied to the input predictions by using data testing with the output in the form of a crosstable shown in Figure 7.



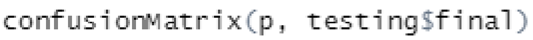
```
  Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  60


             | predicted pass
 actual pass |      pass |      fail | Row Total |
-------------|-----------|-----------|-----------|
        pass |        52 |         2 |        54 |
             |     0.867 |     0.033 |           |
-------------|-----------|-----------|-----------|
        fail |         1 |         5 |         6 |
             |     0.017 |     0.083 |           |
-------------|-----------|-----------|-----------|
Column Total |        53 |         7 |        60 |
-------------|-----------|-----------|-----------|
```

Fig. 7. Output of predictions on testing data

In Figure 7 on the testing data, as many as 60 observations on the testing data are applied in predicting the results of learning calculus. A total of 54 observations were applied to the prediction of passing, 52 observations worth passing were successfully predicted and 2 failed to be predicted. A total of 5 observations worth repeating were correctly predicted and 1 observation failed to be predicted from a total of 6 observations.

3.5 Evaluation

Evaluation is the stage of evaluating the C5.0 model that has been built. The evaluation is carried out using a confusion matrix whose input can be seen in Figure 8.



```
confusionMatrix(p, testing$final)
```

Fig. 8. Input Confusion Matrix On Rstudio

The results of the confusion matrix calculation in Figure 8 are presented in tabular form which can be seen in the table 3.

Table 3 – Confusion Matrix

No	Confusion matrix	70 : 30
1.	Accuracy	95%
2.	Precision (positive predict value)	98.11%
3.	Sensitivity	96.3%
4.	Spesificity	83.3%

3.6 Deployment

At this stage, the knowledge or information that has been obtained is presented in a special form, one of which is a simple report that describes the final results of the entire data mining process that has been carried out so that it can be used by users. The presentation of knowledge can be seen in Figure 9.

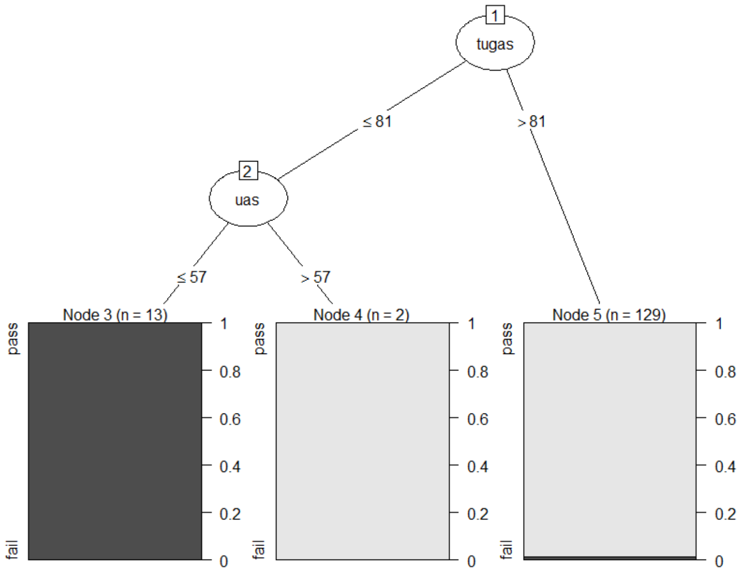
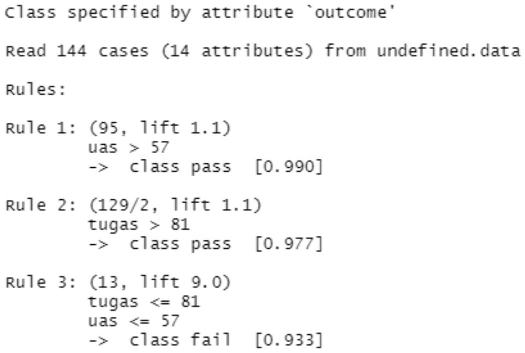


Fig. 9. Decision tree

Figure 9 is a decision tree of the model that has been built. Next, build a rule model to understand the decision tree that has been built. The input rule model can be seen in Figure 10.



```
Class specified by attribute `outcome'

Read 144 cases (14 attributes) from undefined.data

Rules:

Rule 1: (95, lift 1.1)
        uas > 57
        ->  class pass  [0.990]

Rule 2: (129/2, lift 1.1)
        tugas > 81
        ->  class pass  [0.977]

Rule 3: (13, lift 9.0)
        tugas <= 81
        uas <= 57
        ->  class fail  [0.933]
```

Fig. 10. Rule model C5.0

In Figure 10, there are 3 rules, namely when the Exam (UAS) value is > 57 then it is included in the pass class, when the assignment is > 81 then it is included in the graduating class, and when the assignment is <81 and UAS <57 then it is included in the repeat class (fail).

4. Conclusion

From the research that has been done, it can be concluded that the C5.0 algorithm can be used to predict calculus learning outcomes. The prediction process is carried out using a classification method with the C5.0 algorithm with the attributes of guardians, number of family members, residence status, internet, activity, desire to continue studies, parents' last education (father and mother), parents' occupations, assignment scores, UAS, and UTS. The final result of the C5.0 classification process forms a decision tree with 3 rules in it. The performance of the C5.0 algorithm gets an accuracy of 95%.

References

Aesyi, U. S., Lahitani, A. R., Diwangkara, T. W., & Kurniawan, R. T. (2021). Deteksi Dini Mahasiswa Drop Out Menggunakan C5. 0. *JISKA (Jurnal Informatika Sunan Kalijaga), 6*(2), 113-119.

Benediktus, N., & Oetama, R. S. (2020). The decision tree c5. 0 classification algorithm for predicting student academic performance. *Ultimatics: Jurnal Teknik Informatika*, *12*(1), 14-19.

Cherfi, A., Nouira, K., & Ferchichi, A. (2018). Very fast C4. 5 decision tree algorithm. *Applied Artificial Intelligence*, *32*(2), 119-137.

Damanik, I. S., Windarto, A. P., Wanto, A., Andani, S. R., & Saputra, W. (2019, August). Decision tree optimization in C4. 5 algorithm using genetic algorithm. In *Journal of Physics: Conference Series* (Vol. 1255, No. 1, p. 012012). IOP Publishing.

Febriantono, M. A., Pramono, S. H., Rahmadwati, R., & Naghdy, G. (2020). Classification of multiclass imbalanced data using cost-sensitive decision tree C5. 0. *IAES International Journal of Artificial Intelligence*, *9*(1), 65.

Fahmi, R. N. (2021). Implementasi Metode K-Means Clustering dalam Analisis Persebaran UMKM di Jawa Barat. *JOINS (Journal of Information System)*, *6*(2), 211-220.

Han, J., Kamber, M., & Mining, D. (2006). *Concepts and techniques. Morgan Kaufmann*, 340, 94104-3205.

Hanin, N. A. (2022). *Analisis Layanan Informasi Berbasis Chatbot Menggunakan Framework Rasa Open Source Pada Objek Wisata Candi Prambanan* (Doctoral dissertation, Institut Teknologi Telkom Purwokerto).

Rahman, K. (2018). Perkembangan Lembaga Pendidikan Islam di Indonesia. *Jurnal Tarbiyatuna: Kajian Pendidikan Islam*, *2*(1), 1-14.

Renaldi, D. (2020). Penerapan Association Rule Data Mining Untuk Rekomendasi Produk Kosmetik Pada Pt. Fabiando Sejahtera Menggunakan Algoritma Apriori. *Algor*, *2*(1), 1-11.

Sari, B. N. (2017). Prediksi Performa Akademik Siswa Pada Pelajaran Matematika Menggunakan Bayesian Networks dan Algoritma Klasifikasi Machine Learning. KNPMP II, Universitas Muhammadiyah Surakarta. ISSN, 2502-6526.

Sokkhey, P., Navy, S., Tong, L., & Okazaki, T. (2020). Multi-models of educational data mining for predicting student performance in mathematics: a case study on high schools in Cambodia. *IEIE Transactions on Smart Processing and Computing*, *9*(3), 217-229.

Zhang, X., Xue, R., Liu, B., Lu, W., & Zhang, Y. (2018, July). Grade prediction of student academic performance with multiple classification models. In *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery* (ICNC-FSKD) (pp. 1086-1090). IEEE.