Chapter in Book

Determining Eligibility Criteria for Hostel Placement at National Defense University Malaysia using the Naïve Bayes Algorithm

Nurul Huda Jayu Alis Ali[1,*], Suhaila Ismail[2], Syahaneim Marzukhi[3], and Nurhafizah Moziyana Mohd Yusop[4]

1 National Defense University Malaysia; jayualisali25@gmail.com; 0009-0007-5681-3093
2 National Defense University Malaysia; suhaila@upnm.edu.my; 0000-0003-4236-2758
3 National Defense University Malaysia; syahaneim@upnm.edu.my; 0000-0001-9674-7384
4 National Defense University Malaysia; moziyana@upnm.edu.my 0000-0003-0394-7710
* Correspondence: jayualisali25@gmail.com

Abstract: *The study of Determining Eligibility Criteria for Hostel Placement at the National Defense University Malaysia (NDUM) using the Naïve Bayes algorithm began when the student residence management called Pusat Pengurusan Kediaman Pelajar (PPKP) introduced a new policy where civilian students in year two and above are required to stay outside of the NDUM campus. Following the meeting when the new policy was announced, the study revealed several problems and issues that this new policy had brought into. One of them is the evaluation of student placement qualifications based only on the quantity of student involvement in the university. Thus, to overcome this problem, a study was made to compare several factors or criteria selected using the Naïve Bayes algorithm. To produce more accurate and high-quality outcomes for student placement qualifications, an interview with stakeholders were conducted. The outcome of this study will assist PPKP to identify without heed to bias which students qualified for hostel placement. In this study, a more thorough discussion will be had on the use of the Naïve Bayes algorithm, as well as comparisons of other algorithms like the Decision Tree algorithm. The methodology of this study uses Cross-Industry Standard Process for Data Mining (CRISP-DM). The success of this study is anticipated to help PPKP and other NDUM residents in the future.*

Keywords: Algorithm; Naïve Bayes; Decision Tree; Modelling; Analysis

1. INTRODUCTION

Pusat Pengurusan Kediaman Pelajar (PPKP) is a department that is responsible for providing accommodation facilities for NDUM students. This study was conducted to help facilitate PPKP in conducting analysis and inspection for each student who is eligible for accommodation at NDUM's residential college. In May 2022, PPKP introduced a new policy where second-year civilian students and above are required to stay outside the campus at their own expense. This new policy amendment has three categories: 1) the mandatory category, 2) the accommodation priority category, and 3) the needy student category. The needy students consist of second-year civilian students and above in Bachelor and Diploma programs.

Therefore, a study is needed to filter the students, especially the second-year and above, to determine whether they are eligible to reside on the NDUM campus. This hostel placement study requires several important criteria in determining the eligibility for student placement on the NDUM campus. This is to produce accurate and quality student qualification results. 6 important criteria have been selected in this study, namely, the level of student involvement, family income, year of study, exam results, number of siblings, and place of residence. These 6 selected criteria are based on a questionnaire made through a google form to students to identify the satisfaction of the assessment method during the PPKP in assessing the qualifications of students based on one criterion only, which is the 'Level of Student Activity Involvement'. Based on these criteria, needy students also have the opportunity to get an exception and can stay in the NDUM hostel. The student must meet the minimum requirements (threshold), to enable them to qualify for a place in the NDUM hostel. The outcomes of the student placement data will be shown visually, particularly, through the use of implemented graphs and charts that may be seen on the created dashboard.

2. METHOD & MATERIAL

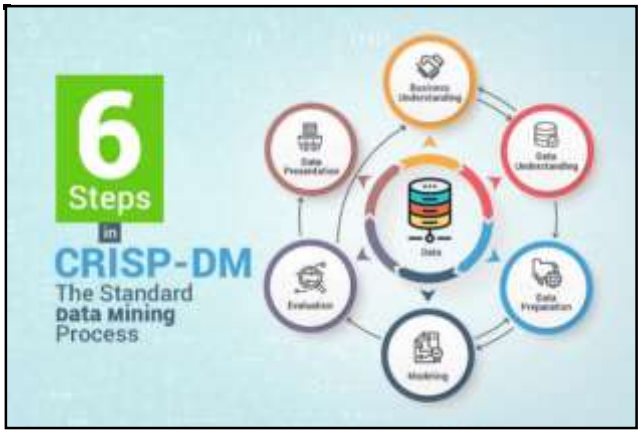

Figure 1. Data Mining Process Diagram (Source: Proglobal Business Solutions, 2022)

CRISP-DM (Cross Industry Standard Process for Data Mining) is a widely used data mining process model that outlines the steps involved in a data mining project as shown in Figure 1. The steps involved in CRISP-DM are:

i. Business Understanding: In this stage, the problem statement is defined, and the goals and objectives of the project are determined. The main objective of determining student placement eligibility at NDUM is to produce student placement qualification results to Pass and Fail using the Naïve Bayes algorithm based on the data and criteria that have been set through the analysis process using RapidMiner software. It was made based on the main problem that has been identified, which is the use of the current system used by the PPKP where the evaluation of student placement qualifications only depends on one criterion. also, students need to make an application to get placement on the UPNM campus. This makes it difficult for students in many ways where students have to apply for placement manually and have to check with the date that will be opened to make a placement application. Therefore, it will cause other problems and interfere with the PPKP process in making a screening assessment of student placement qualifications at UPNM. An alternative and analytical study is made to achieve the objective to produce the best model results.

ii. Data Understanding: In this stage, the data that will be used for the project is collected and analyzed. This includes identifying data sources, understanding data formats, and exploring data quality issues. An interview session was held with the Assistant Hostel Manager, which aims to obtain information related to the topic of Determining Eligibility Criteria for Student Placement at UPNM. From the results of this interview, she has been permitted to use the entire data of students residing inside and outside the NDUM campus for this study only. However, the existing data is not sufficient, therefore other alternatives are taken to meet the variable criteria that will influence the final results of the study. Another alternative is to make an appointment with the Senior Director of Student and Alumni Affairs which to obtain a way of calculating and evaluating student involvement in the university throughout one semester. To obtain JSK and FSTP student engagement data, the google form will be distributed and filled in by the students. The collected engagement data will be combined with the existing data and then use for pre-processing.

iii. Data Preparation: In this stage, the data is prepared for modelling. This involves cleaning the data, transforming the data, and creating new features from the data. After the data selection is made, data content that has problems such as missing data, null data, noisy data, trash data and many errors and inconsistencies will undergo pre-data processing.

iv. Modelling: In this stage, a variety of modelling techniques are applied to the data to create predictive models. The modelling techniques can include statistical modelling, machine learning, and other techniques. The data that has been collected is fed into the RapidMiner analysis tool, and the results from the built model will begin to explain what arose during the Business Understanding phase. Modelling is usually carried out in various iterations. Typically, data miners run several models using parameters or attributes and then refine the parameters. To ensure this phase runs, the data must be in good condition. The selection of an appropriate modelling technique is the first step in this phase. The algorithm model that will be used for this project is the Naïve Bayes algorithm. The Naïve Bayes Algorithm is used to identify and understand the relationship between the attributes and the variables selected to make the qualification prediction using Naïve Bayes.

v. Evaluation: In this stage, the models are evaluated based on their performance on test data. The evaluation criteria are defined in the business understanding stage. When the model and data have been built and prepared, it is important to re-evaluate the model by referring back to the stages of data preparation and modelling. It aims to assess if there are any improvements or details that need to be added at both levels to produce good results. To evaluate the performance of the model using the Naïve Bayes algorithm, a test of the existing sample data will be carried out with the own data sample to perform a prediction accuracy check.

vi. Deployment: In this stage, the model is deployed into the business process. This can involve integrating the model into a software system or creating a report that includes the model's predictions. The model built or known as the Naïve Bayes classifier will be used for predictive studies to determine eligibility criteria for student hostel placement within the NDUM campus. The user interface will also be built at this stage. This allows PPKP to see and interact with the display of the final results of the student placement analysis. In the end, the results will be visualized interactively in the form of graphics in the student placement dashboard.

3. FINDINGS

Determining Eligibility Criteria for Hostel Placement at the National Defense University of Malaysia using the Naïve Bayes Algorithm aims to analyse each criterion required and produce student placement qualification results to Pass and Fail based on the data and criteria that have been set through the analysis process using RapidMiner software. At the early stage, an interview with the NDUM assistant hostel manager and senior director of Student and Alumni Affairs (HEPA) was conducted to gather all important information that can be used for this study. Furthermore, a survey was designed and distributed to 1st, 2nd, and 3rd-year students from the Computer Science Department, Faculty of Defense Science & Technology (JSK, FSTP) NDUM students to fill in by JSK, FSTM students. The collected engagement data will be combined with the existing data and then the data will be used for data pre-processing.

3.1 Data collection

The total amount of data used as training and testing data is 147 students consisting of 1st, 2nd, and 3rd-year students from the Computer Science Department, Faculty of Defense Science & Technology (JSK, FSTP) NDUM.

3.2 Data Processing

Data that has been gathered is cleaned, and then the data is chosen as necessary. The level of student involvement, family income, year of study, exam results, number of siblings, and place of residence are the data utilised in the Naive-Bayes algorithm for the Determining Eligibility Criteria for Hostel Placement at NDUM. The data are collected from the NDUM assistant hostel manager, senior director of Student and Alumni Affairs (HEPA) and also survey findings. The data and information utilised in this project are shown in Figure 2.

Figure 2. Result of Preprocessing Data

The data that has been cleaned will be then used for modelling to get an accurate result to determine whether it can be used to determine the student's eligibility. Overall, preprocessing is a crucial step in data analysis and machine learning, as it helps to prepare the data for analysis, improve the accuracy and reliability of the results, and enhance the performance of the machine learning algorithm.

3.3 Survey Result

When realising that the data gathered from both the NDUM assistant hostel manager and the senior director of HEPA is not enough as there are one missing parameters that are crucial for this study which is student involvement data. Hence, a survey was conducted to collect it because there are more crucial parameters than student involvement that will make this study a success.

Figure 3 and Figure 4 show the pie chart and bar graph of the analysis of responses to the survey questions from students respectively. According to Figure 3, it can be concluded that students are not agreed with the action of PPKP which only look at one factor when determining the students who are eligible to get the hostels at NDUM. The majority of the students (90%) agreed that other factors must be taken into account as well other than the level of student involvement.

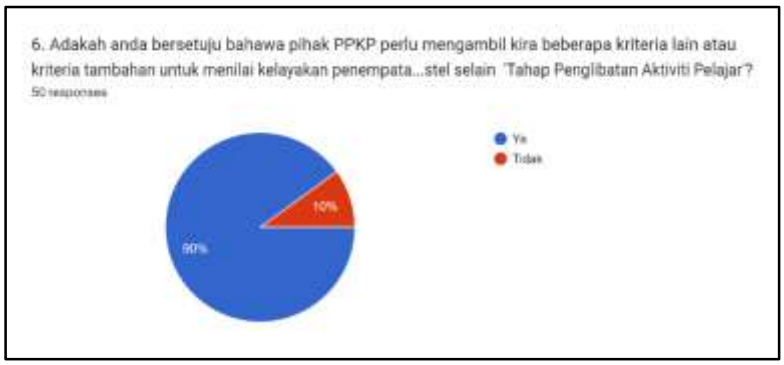

Figure 3. Considering additional factors

Hence, in the survey, we listed out other criteria that were found as important as student involvement and students are required to pick which criteria are important. They can also give opinions on other criteria that are not listed. Figure 4 shows the analysis response from the students.

Figure 4. Suggestions for Additional Criteria

Based on the survey, the majority of respondents (90.2%) concurred that PPKP should take the household income category into account in addition to the level of involvement in student activities. The distance from one's place of residence came in second (82.4%), indicating that the location of the students should also be taken into account while making decisions. Although less frequently mentioned as extra factors than household income and distance from the applicant's place of residence,

total dependents (64.7%) and academic year (56.1%) were. Less than half of the respondents (49%), who mentioned exam results, show that academic performance may not be as significant as other elements in the decision-making process. Lastly, only 8% of respondents offered additional criteria, showing that the survey's proposals were thorough and addressed the majority of the pertinent elements that should be taken into account.

Overall, the poll advises PPKP to take home income category, distance from dwelling, and amount of involvement in student activities into account when making selections. Although they were mentioned as prospective criteria, total dependents and academic year were not as well-liked as the previous two. Less than half of the respondents mentioned exam scores, suggesting that they might not be as significant as other elements in the decision-making process.

4. DISCUSSION

4.1 Calculation of Performance Results

The Naive Bayes Algorithm's calculation formula is shown in Figure 5, and the computation of the placement eligibility criterion using the first row of training data from the data set prior to prediction is shown in Figure 6.

$$P(H|X) = \frac{P(X|H)x\ P(H)}{P(X)}$$

Figure 5. Naïve Bayes Algorithm Formula

Reference:

X = Data with an unknown class

H = Hypothetical data X is a specific class

P(H|X) = Probabilistic-hypothesis H based on state x

P(H) = Probability-hypothesis H

P(X|H) = Probability-X based on the condition

P(X) = Probability-of X

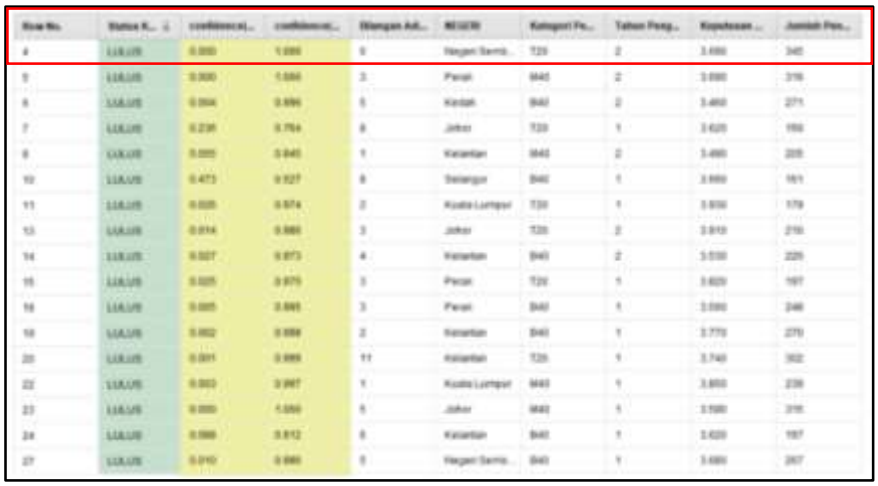

Figure 6. Data Set Before Prediction

Step 1: Count the number of classes and divide by the label

- P (Y = Pass) = 22/41
- P (y = Fail) = 19/41

Step 2: Count the number of total attributes that are the same as the class

- P (Number of Siblings = <=8 | Y Passed) = 8/22
- P (Number of Siblings = <=8 | Y Failed) = 7/19
- P (State = S | Y Passed) = 6/22
- P (State = S | Y Failed) = 3/19
- P (Income Category = T20 | Y Passed) = 7/22
- P (Income Category = T20 | Y Failed) = 3/19
- P (Year of Study = 2 | Y Passed) = 9/22
- P (Year of Study = 2 | Y Failed) = 5/19
- P (Exam Result = >=3.67 | Y Passed) = 14/22
- P (Exam Result = >=3.67 | Y Fail) = 10/19
- P (Total Income = >=300 | Y Passed) = 8/22
- P (Total Income = >=300 | Y Fail) = 0/19

Step 3: Total Attributes According to Each Pass and Fail Label

- P (P (Number of Siblings = <=8 | Y Passed) *
  P (State = S | Y Pass) *
  P (Income Category = T20 | Y Passed) *
  P (Year of Study = 2 | Y Passed) *
  P (Exam Result = >=3.67 | Y Passed) *
  P (Total Exams = >=300 | Y Passed)

=8/22 * 6/22 * 7/22 * 9/22 * 14/22 * 8/22 * 22/41
=0.001603

• P (P (Number of Siblings = <=8 | Y Failed) *
P (State = S | Y Failed) *
P (Income Category = T20 | Y Failed) *
P (Year of Study = 2 | Y Failed) *
P (Exam Result = >=3.67 | Y Failed) *
P (Total Exams = >=300 | Y Failed)
=7/19 * 3/19 * 3/19 * 5/19 * 10/19 * 0/19 * 19/41
=0

Step 4: Comparing the 'Pass and Fail' variables from the calculations made
=(P | Fail) < (P | Pass)
=0 < 0.001603
= Passed

Therefore, the result of the data prediction in Figure 6 first line is **Pass.**

Figure 7. Data Set After Prediction

The study found that the computation made manually and in RapidMiner software are the same. RapidMiner accurately processes the data.

4.2 Algorithm Performance Evaluation

At the level of algorithm performance evaluation, confusion matrix will be used. Confusion matrix can be thought of as a tool that evaluates if an algorithm is able to find tuples belonging to various classes. When the algorithm performs true value data classification, values from True-Positive and True-Negative provide information, whereas False-Positive and False-Negative provide information when the algorithm performs data classification incorrectly. Table of the confusion matrix is shown in Figure 8.

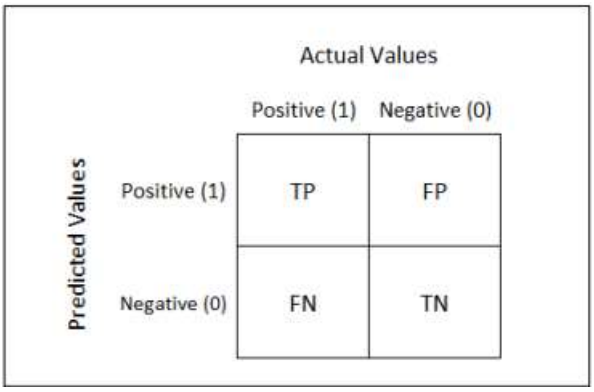

Figure 8. Confusion Matrix

Based on the results of an exercise comparing actual data and expectations, Table 5.1 serves as a test to establish eligibility requirements for dormitory placement.

Table 1. Confusion Matrix Determines Eligibility Criteria for Hostel Placement at NDUM

Predicted Value	**Actual Value**	
	Pass	**Fail**
Pass	22 (TP)	1 (FP)
Fail	0 (FN)	18 (TN)

- *Sensitivity/ Recall* = TP/ (TP + FN)

  = 22/ (22 + 0)

  = 1 * 100 = 100%
- *Specificity* = TN/ (TN + FP)

  = 18/ (18 + 1)

  = 0.9474 * 100 = 94.74%
- *Precision* = TP/ (TP + FP)

  = 22/ (22 + 1)

  = 0.9565 * 100 = 95.65%
- *Accuracy* = TP + TN/ (TP + TN + FP + FN)

  = 22 + 18/ (22 + 18 + 1 + 0)

  = 0.9756 * 100 = 97.56%

The accuracy value for determining the requirements for placement in the UPNM hostel is 97.56%. This can demonstrate that the Naive Bayes algorithm can generate results with a high degree of accuracy.

5. CONCLUSION

In conclusion, this study has identified and examined six crucial factors that will help determine whether students are qualified for housing at NDUM's residential college. The selection criteria are determined by factors including the degree of student involvement, family income, year of study, exam scores, sibling count, and residence. To ascertain if a student was eligible for placement in a hostel, predictive models were built using the Nave Bayes algorithm based on the chosen criteria. Through the use of implemented graphs and charts that can be seen on the developed dashboard, the results of the student placement data were visually presented. Pusat Pengurusan Kediaman Pelajar (PPKP) would be able to use this study to make screening determinations regarding student placement qualities at UPNM. It will also assist in determining the optimum model to generate Pass/Fail outcomes for student placement qualification. This study also shows that in order to guarantee the best model results, the data collection process should be carried out precisely and methodically. In the future, more sophisticated data mining methods might be used to increase the model's precision and potency in predicting student placement qualifications at NDUM.

Acknowledgements: We would like to acknowledge the Student Affairs Department and the Department of Computer Science, Faculty of Science and Defense Technology, National Defense University Malaysia for the financial support in this project.

References

Bhullar, Kaur, & Member IAENG. (2012). Use of data mining in education sector. *Proceedings of the World Congress on Engineering and Computer Science*, 1, 24–26.

Chan, A. S., Fachrizal, F., & Lubis, A. R. (2020). Outcome prediction using Naïve Bayes Algorithm in the selection of role hero mobile legend. *Journal of Physics: Conference Series*, 1566(1), 012041. https://doi.org/10.1088/1742-6596/1566/1/012041

ChartExpo. (2022, October 26). *How to use Excel Data Visualization Tool?* Retrieved November 18, 2022, from https://chartexpo.com/blog/excel-data-visualization-tool

Chauhan, G. (2018, October 8). *All about Naive Bayes - Towards data science.* Medium. Retrieved November 4, 2022, from https://towardsdatascience.com/all-about-naive-bayes-8e13cef044cf

CompareCamp. (n.d.). *RapidMiner Review*. Retrieved October 2, 2022, from https://comparecamp.com/rapidminer-review-pricing-pros-cons-features/

CompareHero.my Team. (2022, March 24). *T20, M40 and B40 Income classifications in Malaysia*. CompareHero. Retrieved October 24, 2022, from https://www.comparehero.my/budgets-tax/articles/t20-m40-b40-malaysia

Data prep and machine learning made fun, fast and simple. (2022, May 9). RapidMiner. Retrieved December 10, 2022, from https://rapidminer.com/blog/data-prep-machine-learning/

D*efinition of Predictive Modeling - Gartner Information Technology Glossary.* (n.d.). Gartner. Retrieved October 2, 2022, from https://www.gartner.com/en/information-technology/glossary/predictive-modeling

Gillis, A. S. (2022, April 15). *Data splitting.* Enterprise AI. Retrieved December 18, 2022, from https://www.techtarget.com/searchenterpriseai/definition/data-splitting

Guide to data cleaning: Definition, benefits, components, and how to clean your data. (n.d.). Tableau. Retrieved December 18, 2022, from https://www.tableau.com/learn/articles/what-is-data-cleaning

IBM. (2021, August 17). *Data preparation overview.* Retrieved September 28, 2022, from https://www.ibm.com/docs/en/spss-modeler/saas?topic=preparation-data-overview

J-PTIIK. (2022, December 13). *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer.* JPTIIK. Retrieved November 18, 2022, from https://j-ptiik.ub.ac.id/index.php/j-ptiik

k, T., & Wadhawa, M. (2016). Analysis and comparison study of data mining algorithms using Rapid Miner. *International Journal of Computer Science, Engineering and Applications*, 6(1), 9–21. https://doi.org/10.5121/ijcsea.2016.6102

Khasanah, A. U., & Harwati. (2017). A comparative study to predict student's performance using educational data mining techniques. *IOP Conference Series: Materials Science and Engineering*, 215, 012036. https://doi.org/10.1088/1757-899x/215/1/012036

Mulyani, Kurniadi, Nashrulloh, Julianto, & Regita. (2022). The prediction of PPA And Kip-Kuliah Scholarship recipients using Naive Bayes Algorithm. *Jurnal Teknik Informatika (JUTIF),* 3(4), 821–827. https://doi.org/10.20884/1.jutif.2022.3.4.297

Peling, I. B. A., Arnawan, I. N., Arthawan, I. P. A., & Janardana, I. G. N. (2017). Implementation of data mining to predict period of students study using Naive Bayes Algorithm. *International Journal of Engineering and Emerging Technology,* 2(1), 53. https://doi.org/10.24843/ijeet.2017.v02.i01.p11

Proglobalbusinesssolutions. (2022, June 13). Six steps in CRISP DM - *The standard data mining process.* PGBS. https://www.proglobalbusinesssolutions.com/six-steps-in-crisp-dm-the-standard-data-mining-process/

Rohani, T. (2022, February 26). Senarai Negeri di Malaysia (Mengikut Urutan dan Kawasan). *The Rohani*. https://therohani.com/senarai-negeri-di-malaysia/

Smart Vision Europe. (2020, June 17). Crisp DM methodology. Retrieved November 10, 2022, from https://www.sv-europe.com/crisp-dm-methodology/

Talend. (n.d.). *What is data preparation? Processes and example*. Talend - a Leader in Data Integration & Data Integrity. Retrieved September 25, 2022, from https://www.talend.com/resources/what-is-data-preparation/