Part I. Building Analytics Model

1. Introduction

ABC Company is a language and education center located in MacArthur Highway, Guiguinto, Bulacan, Philippines. The company was established in 2014 and has since been providing language and educational services to the local community.

ABC Company is owned and managed by a team of language and education experts who are passionate about providing high-quality language and education services to their clients. The center employs a team of dedicated and experienced language instructors who are committed to helping students achieve their language goals.

The center offers a wide range of language programs, including English as a Second Language (ESL), Mandarin, Japanese, and Korean. The center also offers test preparation courses for various language proficiency exams, including TOEIC, TOEFL, IELTS, and JLPT.

The mission of ABC Company is to provide quality assistance and training in the Philippines by developing individuals through Excellent Education, Value Transformation, for the enrichment and enhancement of lives. The center's vision is to be the premier language skills provider for Filipinos who wants to establish a career and sustainable living in Japan, through knowledge, Skills, Virtues and Professional competence.

The values of ABC Company include excellence, integrity, teamwork, innovation, and customer satisfaction. Their pledge is to prepare the learner to develop the knowledge, skills and attitudes of the learners in accordance with the standards on the Japanese language as well as to gain competency as they are to be exposed in work environment in Japan. The center is committed to providing its clients with the highest quality language and education services, using innovative and effective teaching methods, and maintaining a high level of integrity and professionalism in all its interactions with clients.

2. Business Process

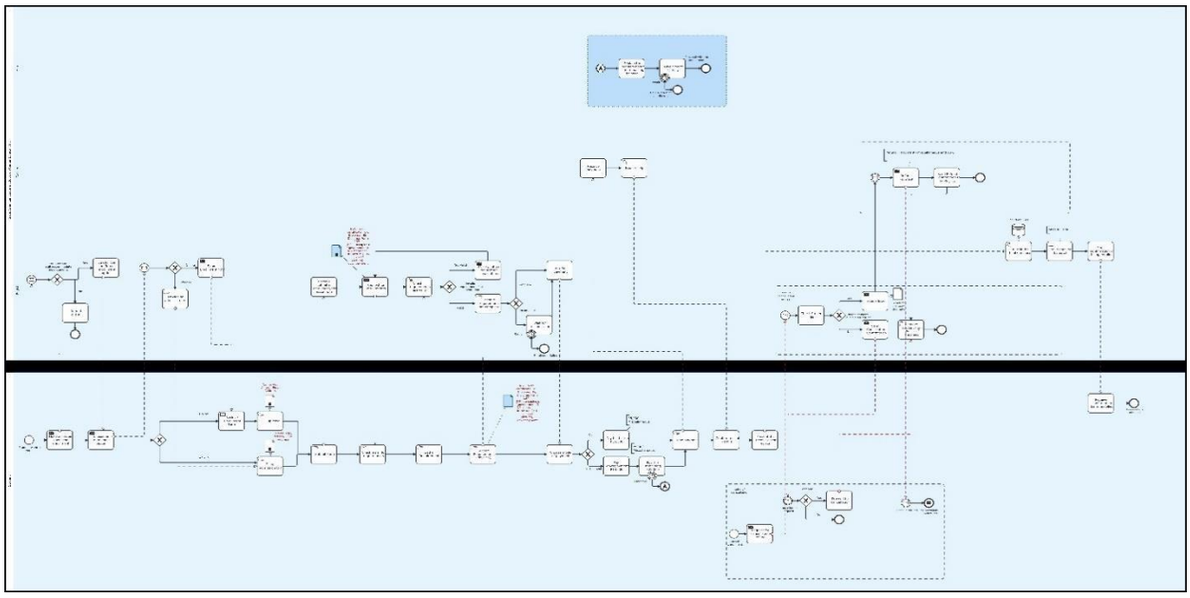


Figure 1

BPMN of ABC Company

The Enrollment process starts when a student wants to enroll in ABC Company They do have two options they could inquire through online, or they could go to the physical location. Ones the student wants to go through the enrollment process they will fill out a form with their basic information. Once the company receives the form, they will send a list of requirements that the client needs to gather. Once the client already gathered all their requirements and submitted it to the company. The company will then go through this thoroughly and check if there are any missing requirements, If there is, the company will let the client to resubmit the late requirements within five days if the client fails to submit it, the enrollment process will fail. If the client succeeded in submitting the late requirements the enrollment process will continue, and they will be required to pay for the tuition. The tuition costs around 35,000 php this already includes all the materials that will be used in the learning process. The client does have two options, they could pay the tuition in full, or they could put a downpayment first, the downpayment is a minimum of 13,000 php and within 3 months they must complete the remaining 22,000 php. After processing the payment, the company will then give a receipt to provide proof of the client's payment. After receiving the receipt, the client will present the proof of payment to the registrar so that he/she will be included in the final list of the students that will be group by batch. After this process the company will send an email confirmation stating their enrollment confirmation and class details. If ever the client's mind changes about the enrollment, he/she will have to request an enrollment cancellation request. Once the company receives it, they will go to the enrollees file and check if the student is eligible for a refund. If the student still falls under the cancellation term by the company, he/she will then receive a refund of the miscellaneous fees and the 10% of the payment that they made. If not, they won't receive a refund, but the cancellation of the enrollment will still go through. The company will send a confirmation email about the cancellation and the refund, if applies.

3. Business Problems

ABC Company is facing a problem with a decreasing completion rate of students. If this problem persists, it can lead to several negative consequences. Firstly, it can result in a decline in the institution's reputation, which may negatively impact future enrollments, leading to decreased revenue and financial instability for the language center. Secondly, students who do not complete their courses may not acquire the necessary language skills, leading to reduced job opportunities and lower earning potential. Finally, a low completion rate can also result in reduced satisfaction levels among students, further damaging the institution's reputation and affecting future enrollments.

4.Proposed Solution

Our proposed analytical solution is to develop a classification model that can predict whether a student is a completer or non-completer. This will involve identifying the significant attributes that contribute to student completion rates. The model will be built using machine learning algorithms and trained, we will also develop a system that would predict whether a student is completer or noncompleter.

Implementing the proposed analytical solution can provide a positive impact to the organization by addressing the decreasing student-retention rate. With a classification model that can predict whether a student is a completer or non-completer, the organization can proactively identify at-risk students and provide timely interventions to improve their chances of completing their program of study. This can lead to an increase in completion rates, which can improve the organization's reputation and competitiveness in the market. Additionally, the developed system can also provide insights into the factors that contribute to student completion rates, which can be used to further improve the organization's programs and services. Overall, the proposed solution can help the organization achieve its objective of increasing the completion rate of students and ultimately, contribute to its long-term success.

5. Data Understanding

The dataset comprises 3,000 instances and 11 attributes.

Additionally, from the article, the attribute of educational attainment, specifically the educational background of a student's, is relevant to predicting retention because it is a measure of socioeconomic status. First-generation college students, who are the first in their families to attend college, often come from families with lower levels of educational attainment and, therefore, lower socioeconomic status. As a result, they may have less financial and mentoring support from their families, which can contribute to a higher risk of dropping out. Therefore, educational attainment is an important variable to consider when predicting college retention, as it can help identify students who may be at a higher risk of dropping out and who may need additional support to succeed in college.

Commuters are also more likely to opt out of school. (Pascarella & Terenzini, 1998). This effect has been ascribed by scholars to commuters' lack of social interaction. (Thompson et al., 1993). Commuters likely face greater time constraints because they are more likely to be elderly, employed, and have a family. (Kuh et al.) Thus, the attributes of proximity and transportation are relevant because they are factors that affect a student's decision to commute. Commuting students often face more time pressures because they have to travel to and from campus, which can be time-consuming and exhausting. This can make it more difficult for them to balance their academic and personal responsibilities, leading to a higher risk of dropping out.

Moreover, commuting students may also experience a lack of social interaction compared to those who live on campus. Living on campus provides students with opportunities to engage in various activities, clubs, and events, which can help them build relationships with peers and create a sense of community. In contrast, commuting students may miss out on these opportunities and feel disconnected from the campus community, leading to feelings of isolation and potentially contributing to a higher risk of dropping out.

The mode of transportation used for commuting can also be a factor. Students who rely on public transportation may have to navigate complex schedules and routes, which can add to their time pressures and increase the likelihood of missing classes or other academic obligations. On the other hand, students who drive to campus may face challenges such as parking difficulties or traffic congestion, which can also cause stress and disrupt their schedules. Several studies also indicate that male students are less likely to persist than female students that's why the team believes that the gender attribute is relevant for prediction.

Table 1

Data Dictionary

Attribute Name	Description	Data Type / Measure	Format	Sample acceptable values	Maximum Field Size	Data Codes [For numeric that represents categories]
Gender	The gender identity of the student	String/Nominal	Not applicable	Male, Female,	6	Not applicable
Transportation	The mode of transportation used by the student to commute to the learning center	String/Nominal	Not applicable	Public Transportation, Private Transportation	20	Not applicable
Income	The annual household income of the student	Numeric	Not applicable	25000, 50000, 75000, 100000	10	Not applicable
Educational Attainment	The highest level of education completed by the student	String/Nominal	Not applicable	Junior High School Graduate Senior High School Ongoing, Junior High School Ongoing, Junior High School Undergraduate, Junior High	30	Not applicable

				School Graduate, Elementary Undergraduate, Elementary Graduate, Elementary Ongoing, Senior High School Graduate, College Ongoing, Senior High School Undergraduate, College Graduate, College Undergraduate		
Proximity	The distance between the student's residence and the learning center	Numeric Miles	Not applicable	1 - 15 km 31 + km 16 - 30 km	10	Not applicable
Learning Modality	The method of instruction used by the learning center	String/Nominal	Not applicable	Face to Face, Online Class	15	Not applicable
Employment History	The employment status of the student	String/Nominal	Not applicable	Employed, Unemployed	10	Not applicable
Purpose	The primary reason the student is attending the learning center	String/Nominal	Not applicable	Hobby, Requirements in School, For Employment Purposes	20	Not applicable
Payment	The method of payment used by the student to pay the tuition fee	String/Nominal	Not applicable	Monthly Installment, Cash Payment	20	Not applicable
Outcome	The final academic status of the student	String/Nominal	Not applicable	Completer, Non-Completer	13	Not applicable
Proficiency	Proficiency level of the student in the language being taught	String/Nominal	Not applicable	Beginner, Intermediate, Advanced	15	Not applicable

6. Preprocessing Techniques

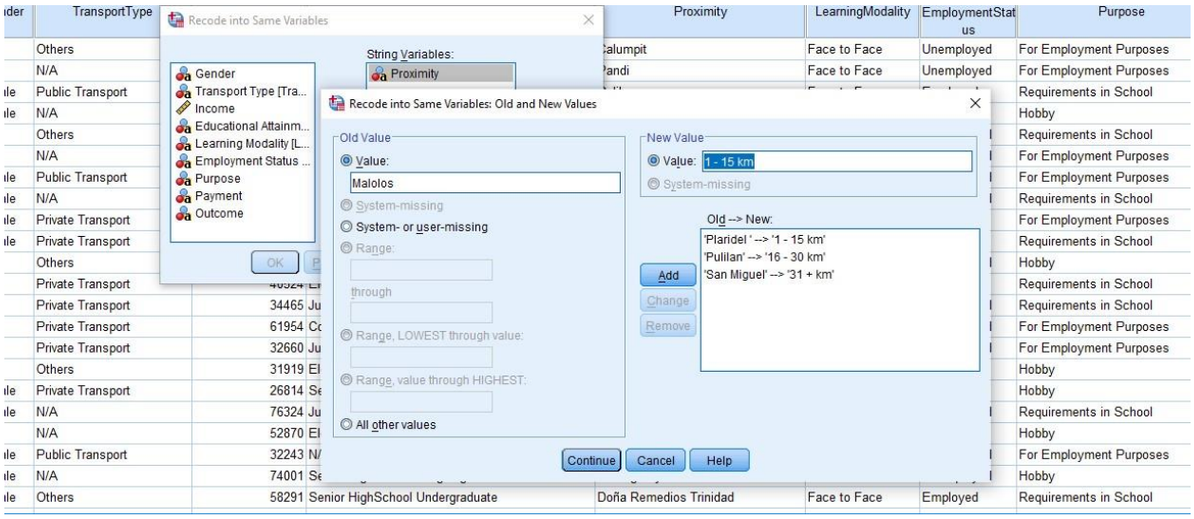


Figure 2

Preprocessing Techniques in SPSS

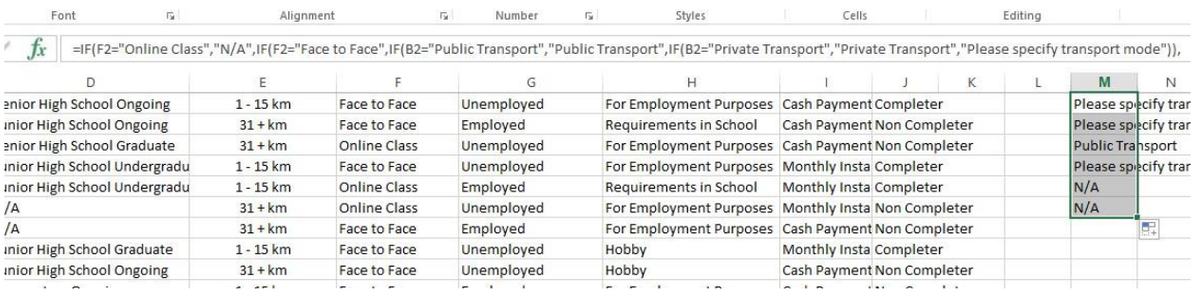
First, the researchers recode that proximity to the same variable using spss, converting the municipalities in Bulacan into 3 categories mainly 1 – 15 km, 16 – 30 km, and 31 + km.

	A	B	C	D	E	F	G	
1	Gender	Transport	Income	Educational Attainment	Proximity	Learning Modality	Employment Status	Purpose
2	Male	Others	39416	Junior High School Graduate	1 - 15 km	Face to Face	Unemployed	For Employ
3	Male	N/A	44004	Senior High School Ongoing	31 + km	Face to Face	Unemployed	For Employ
4	Female	Public Tra	98956	Junior High School Ongoing	16 - 30 km	Face to Face	Employed	Requireme
5	Female	N/A	57875	Elementary Undergraduate	31 + km	Face to Face	Employed	Hobby

Figure 3

Preprocessing Techniques in Excel

After that the researchers change the learning modality using this excel formula “=IF(E2="31 + km","Online Class","Face to Face") This formula uses the IF function to check if the cell E2 contains the text "31 + km". If it does, the formula returns "Online Class". If it doesn't, the formula returns "Face to Face".



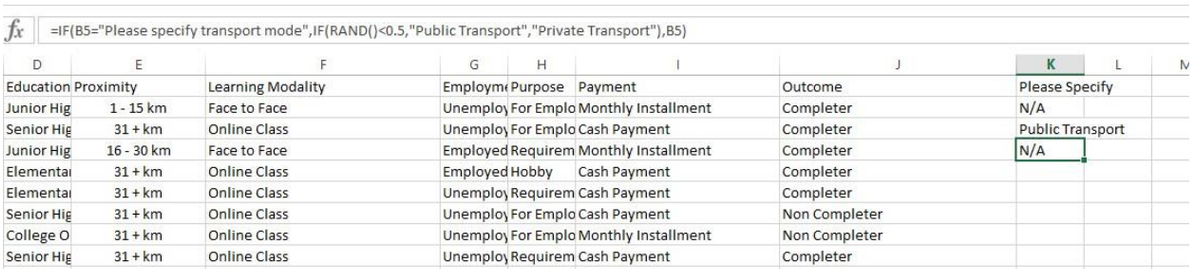
D	E	F	G	H	I	J	K	L	M	N
enior High School Ongoing	1 - 15 km	Face to Face	Unemployed	For Employment Purposes	Cash Payment	Completer			Please specify trar	
nior High School Ongoing	31 + km	Face to Face	Employed	Requirements in School	Cash Payment	Non Completer			Please specify trar	
enior High School Graduate	31 + km	Online Class	Unemployed	For Employment Purposes	Cash Payment	Non Completer			Public Transport	
nior High School Undergradu	1 - 15 km	Face to Face	Unemployed	For Employment Purposes	Monthly Insta	Completer			Please specify trar	
nior High School Undergradu	1 - 15 km	Online Class	Employed	Requirements in School	Monthly Insta	Completer			N/A	
/A	31 + km	Online Class	Unemployed	For Employment Purposes	Monthly Insta	Non Completer			N/A	
/A	31 + km	Face to Face	Employed	For Employment Purposes	Cash Payment	Non Completer				
nior High School Graduate	1 - 15 km	Face to Face	Unemployed	Hobby	Monthly Insta	Completer				
nior High School Ongoing	31 + km	Face to Face	Unemployed	Hobby	Cash Payment	Non Completer				

Figure 3.1

Preprocessing Techniques in Excel

And after recoding the proximity and Learning modality we then begin to pre-process the Transport Type, using this Excel formula “=IF(F2="Online Class","N/A",IF(F2="Face to Face",IF(G2="Public Transport","Public Transport",IF(G2="Private Transport","Private Transport","Please specify transport mode")), "Invalid value"))”. This formula uses the IF function with nested IF functions to check the value of cell F2 and G2. If F2 contains "Online Class", the formula returns "N/A". If it contains "Face to Face", the

formula checks the value of G2. If G2 contains "Public Transport", the formula returns "Public Transport". If G2 contains "Private Transport", the formula returns "Private Transport". If G2 is blank, the formula returns "Please specify transport mode". If G2 contains any other value, the formula returns "Invalid value". It doesn't end here because it generates "Please specify transport mode" && "Invalid value" which is not included in the three choices namely ""Private Transport " "Public Transport" for Face to Face, and N/A for online class type.



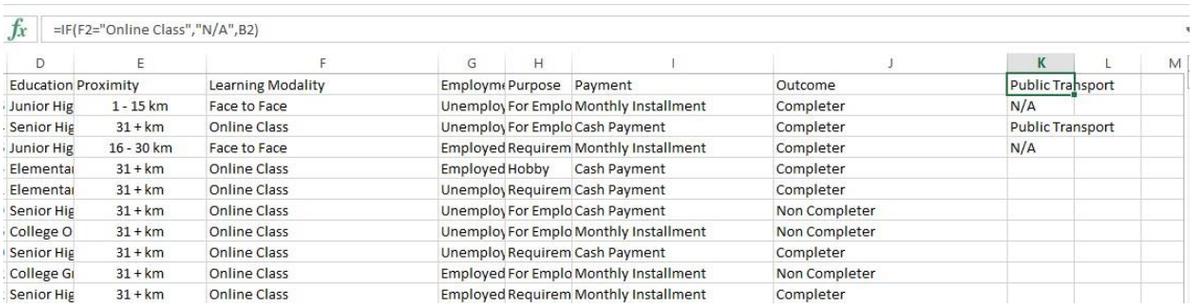
fx =IF(B5="Please specify transport mode",IF(RAND()<0.5,"Public Transport","Private Transport"),B5)

D	E	F	G	H	I	J	K	L	N
Education	Proximity	Learning Modality	Employme	Purpose	Payment	Outcome	Please Specify		
Junior Hig	1 - 15 km	Face to Face	Unemploy	For Emplo	Monthly Installment	Completer	N/A		
Senior Hig	31 + km	Online Class	Unemploy	For Emplo	Cash Payment	Completer	Public Transport		
Junior Hig	16 - 30 km	Face to Face	Employed	Requirem	Monthly Installment	Completer	N/A		
Elementai	31 + km	Online Class	Employed	Hobby	Cash Payment	Completer			
Elementai	31 + km	Online Class	Unemploy	Requirem	Cash Payment	Completer			
Senior Hig	31 + km	Online Class	Unemploy	For Emplo	Cash Payment	Non Completer			
College O	31 + km	Online Class	Unemploy	For Emplo	Monthly Installment	Non Completer			
Senior Hig	31 + km	Online Class	Unemploy	Requirem	Cash Payment	Completer			

Figure 3.2

Preprocessing Techniques in Excel

The researchers then use this formula "=IF(B2="Please specify transport mode",IF(RAND()<0.5,"Public Transport","Private Transport"),B2)" This formula uses the IF function to check if B2 contains "Please specify transport mode". If it does, the formula randomly chooses between "Public Transport" or "Private Transport" using the RAND() function. If the RAND() value is less than 0.5, the formula returns "Public Transport", otherwise it returns "Private Transport". If B2 does not contain "Please specify transport mode", the formula returns the original content of B2.This changes the "Please specify transport mode" to either "Public Transport" or "Private Transport".



fx =IF(F2="Online Class","N/A",B2)

D	E	F	G	H	I	J	K	L	M
Education	Proximity	Learning Modality	Employme	Purpose	Payment	Outcome	Public Transport		
Junior Hig	1 - 15 km	Face to Face	Unemploy	For Emplo	Monthly Installment	Completer	N/A		
Senior Hig	31 + km	Online Class	Unemploy	For Emplo	Cash Payment	Completer	Public Transport		
Junior Hig	16 - 30 km	Face to Face	Employed	Requirem	Monthly Installment	Completer	N/A		
Elementai	31 + km	Online Class	Employed	Hobby	Cash Payment	Completer			
Elementai	31 + km	Online Class	Unemploy	Requirem	Cash Payment	Completer			
Senior Hig	31 + km	Online Class	Unemploy	For Emplo	Cash Payment	Non Completer			
College O	31 + km	Online Class	Unemploy	For Emplo	Monthly Installment	Non Completer			
Senior Hig	31 + km	Online Class	Unemploy	Requirem	Cash Payment	Completer			
College Gi	31 + km	Online Class	Employed	For Emplo	Monthly Installment	Non Completer			
Senior Hig	31 + km	Online Class	Employed	Requirem	Monthly Installment	Completer			

Figure 3.2

Preprocessing Techniques in Excel

Lastly the researchers use this formula "=IF(F2="Online Class","N/A",F2)" This formula uses the IF function to check if F2 contains "Online Class". If it does, the formula returns "N/A". If it does not, the formula returns the original content of F2. This removes the other value aside from N/A for the Online Class type.

7. Attribute Selection

Table 2

Attribute Selection Evaluator

Evaluator	Search Method	Significant Attributes (Arrange based on significance)
InfoGainAttributeEval	Ranker	Proficiency, Educational Attainment, Proximity, Purpose, Transport Type, Learning Modality, Employment Status ,Gender, Payment, Income
ClassifierAttributeEval	Ranker	Proficiency, Income, TransportType, Payment,EducationalAttainment,Proximity,LearningModality,EmploymentStatus,Purpose,Gender
GainRatioAttributeEval	Ranker	Proficiency, Proximity, EducationalAttainment, LearningModality, TransportType,Purpose,EmploymentStatus,Gender,Payment,Income

Table 3

Final List of Selected Attributes

Attributes	Data type / Measure
Proficiency	String/Nominal
EducationalAttainment	String/Nominal
Proximity	String/Nominal
Purpose	String/Nominal
TransportType	String/Nominal

8.Model Selection

Random Forest is supervised ensemble machine learning approach for classification, regression and other tasks that operates by constructing a number of decision trees during training and producing as its output the class that is mode of the classes of the individual trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

According to Abubakar, Y. S., & Ahmad, N. A. (2017). “Prediction of students’ performance in e-learning environment using random forest.” Performance of students may be influenced by several factors such as gender, age, parents’ socioeconomic situation, area of resident, nature of school being attended, school medium of teaching, number of study hours spent daily, and nature of accommodation which may be school own hostel or otherwise [12]. Several research about factors affecting students’ performance at different study levels have been conducted by many authors. Students’ performance prediction is one of the earlies and most valuable applications of Educational Data Mining (EDM) and its objective is to measure the hidden value of students’ performance, understanding or grade from the other information, attitude or behavior of those students.

This paper employs Random Forest to forecast students' success and retention, which is an effective approach to reducing the occurrence of academic failure. If the system identifies a student as being at risk of academic failure, the teaching staff can intervene to prevent it.

Figure 4

Random Forest Model in Training Set

```
=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.58 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.15 seconds
```

The Random Forest algorithm was used to train the model over 100 iterations, and the Random Tree algorithm served as the base learner. The model's building required 0.45 seconds.

Figure 5

Random Forest Validation Result in Training Set

```
=== Summary ===

Correctly Classified Instances        2306               76.8667 %
Incorrectly Classified Instances       694               23.1333 %
Kappa statistic                          0.4726
Mean absolute error                      0.2787
Root mean squared error                  0.3711
Relative absolute error                 62.8664 %
Root relative squared error             78.816  %
Total Number of Instances             3000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.837    0.370    0.820      0.837    0.829      0.473    0.864     0.937     Completer
                 0.630    0.163    0.658      0.630    0.644      0.473    0.864     0.733     Non Complete
Weighted Avg.    0.769    0.301    0.766      0.769    0.767      0.473    0.864     0.869
```

With a Kappa statistic of 0.4726, the classifier correctly identified 76.87% of the occurrences, showing an average degree of agreement. The classifier performed better on the Completer class (precision=0.820, recall=0.837, F-Measure=0.829) than the non-Complete class (precision=0.658, recall=0.630, FMeasure=0.644), according to the detailed accuracy by class.

Figure 6

Random Forest Confusion Matrix in Training Set

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1679  326 |    a = Completer
  368  627 |    b = Non Complete
```

This confusion matrix displays how well the Random Forest classifier performed using the training set of data. While classifying 368 non-Complete instances as Completer and 326 Completer instances as non-Completer, the classifier successfully identified 1679 Completer instances and 627 NonCompleter instances.

Figure 7

Random Forest Model in Supplied Test Set

```
=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.42 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.03 seconds
```

Figure 8

Random Forest Validation Result in Supplied Test Set

```
=== Summary ===

Correctly Classified Instances         458               76.3333 %
Incorrectly Classified Instances       142               23.6667 %
Kappa statistic                          0.4714
Mean absolute error                      0.2835
Root mean squared error                  0.371
Relative absolute error                 63.552  %
Root relative squared error             78.3067 %
Total Number of Instances              600

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.823     0.353     0.819      0.823    0.821      0.471    0.867     0.937     Completer
               0.647     0.177     0.653      0.647    0.650      0.471    0.867     0.745     Non Complete
Weighted Avg.  0.763     0.293     0.763      0.763    0.763      0.471    0.867     0.872
```

On this provided test set, with a Kappa statistic of 0.4714, the classifier accurately identified 76.33% of the occurrences, showing a moderate level of agreement. The classifier performed better on the Completer class (precision=0.819, recall=0.823, F-Measure=0.821) than the non-Complete class (precision=0.653, recall=0.647, F-Measure=0.650), according to the detailed accuracy by class.

Figure 9

Random Forest Confusion Matrix in Supplied Test Set

```
=== Confusion Matrix ===

   a   b   <-- classified as
 326  70 |   a = Completer
  72 132 |   b = Non Complete
```

With a Kappa statistic of 0.4714, the classifier accurately identified 76.33% of the occurrences, showing an average degree of agreement. The classifier performed better on the Completer class (precision=0.819, recall=0.823, F-Measure=0.821) than the non-Complete class (precision=0.653, recall=0.647, F-Measure=0.650), according to the detailed accuracy by class. The confusion matrix indicates that while the classifier accurately identified 132 examples as non-Complete and 326 instances as Completer, it incorrectly identified 70 instances as Non-Completer when they were actually Completer and 72 instances as Completer when they were obviously Non-Completer.

Figure 10

Random Forest Model in 10-Fold Cross Validation

```
=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.31 seconds
```

Figure 11

Random Forest Validation Result in 10-Fold Cross Validation

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

Correctly Classified Instances        1995               66.5    %
Incorrectly Classified Instances      1005               33.5    %
Kappa statistic                          0.2442
Mean absolute error                      0.3358
Root mean squared error                  0.4461
Relative absolute error                 75.7347 %
Root relative squared error             94.7474 %
Total Number of Instances             3000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.750    0.506    0.749      0.750    0.749      0.244    0.744     0.883     Completer
                 0.494    0.250    0.495      0.494    0.495      0.244    0.744     0.488     Non Complete
Weighted Avg.    0.665    0.421    0.665      0.665    0.665      0.244    0.744     0.752
```

The model appears to be performing alright but not overly well. About two thirds of the cases are classified correctly according to accuracy, which is roughly 66.5%. Additionally, the projected and actual class labels do not agree any more than would be anticipated by chance, according to the Kappa statistic, which is quite low at 0.2442. When examining the detailed accuracy per class, it appears that the model

does a better job of classifying the "Completer" class than the "Non-Completer" class (precision and recall are both around 0.75 for the Completer class). This could mean that the model is more inclined to foresee members of the "Completer" class.

Figure 12

Random Forest Confusion Matrix in 10-Fold Cross Validation

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1503  502 |    a = Completer
  503  492 |    b = Non Complete
```

With 72 occurrences of "Non-Completer" being categorized as "Completer" (false positives) and 70 instances of "Completer" being classed as "Non Completer" (false negatives) on the test set, the confusion matrix demonstrates that the model tends to classify more instances as "Completer" than "Non Completer". 502 instances of "Non-Completer" were categorized as "Completer" on the 10-fold crossvalidation, resulting in false positives, and 503 instances of "Completer" were identified as "Non Completer," resulting in false negatives.

Figure 13

Random Forest Model in 10% Split

```
=== Classifier model (full training set) ===

RandomForest

Bagging with 100 iterations and base learner

weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities

Time taken to build model: 0.23 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0.1 seconds
```

Figure 14

Random Forest Validation Result in 10% Split

```
=== Summary ===

Correctly Classified Instances        1801               66.7037 %
Incorrectly Classified Instances       899               33.2963 %
Kappa statistic                          0.2447
Mean absolute error                      0.3426
Root mean squared error                  0.4674
Relative absolute error                 75.5619 %
Root relative squared error             99.2557 %
Total Number of Instances             2700

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.752     0.508     0.752      0.752     0.752      0.245    0.737     0.870     Completer
               0.492     0.248     0.493      0.492     0.492      0.245    0.737     0.484     Non Complete
Weighted Avg.  0.667     0.422     0.667      0.667     0.667      0.245    0.737     0.743
```

The model successfully classified 66.70% of the cases in the test set, according to its accuracy on the test set, which was 66.70%. The match between the model's predictions and the actual outcomes was somewhat better than chance, according to the Kappa statistic of 0.2447.
With a true positive rate of 0.752 and a precision of 0.752, the model did better at predicting "Completer" instances, according to the detailed accuracy by class. The model had trouble correctly identifying "NonCompleter" cases, as evidenced by the fact that the actual positive rate for these instances was only 0.492.

Figure 15

Random Forest Confusion Matrix in 10% Split

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1365  449 |    a = Completer
  450  436 |    b = Non Complete
```

1365 instances were successfully identified as Completer (true positives) and are in fact Completer. False negatives were used to label 449 instances that are truly Completer as non-Completer. 450 occurrences that should have been classed as non-Completer instead were given the false positive label of Completer. True negatives: 436 cases that should have been classed as non-Completer were in fact classified as Non-Completer. In general, out of 2700 examples, the model accurately identified 1801 instances.

Naïve Bayes classifiers are considered as simple probabilistic classifiers that apply Bayes' theorem. This theorem is based on the probability of a hypothesis, given the data and some prior knowledge. The naive Bayes classifier assumes that all features in the input data are independent of each other, which is often not true in real-world scenarios. However, despite this simplifying assumption, the naive Bayes classifier is widely used because of its efficiency and good performance in many real-world applications.

According to Y Divya Bharathi et al. (2018). A Framework for Student Academic Performance Using Naive Bayes Classification Technique, the student performance is successfully predicted using Naive Bayes classification technique. It helps the management take timely action to improve the student performance through extra coaching and counselling. Student failures as a phenomenon have been extensively studied and modelled. To identify the possible causes to seek strategies to prevent it.

The naïve Bayes algorithm is selected as the best algorithm for prediction based on performance detail. The model can be depended on by both students and academic staff to decide the questions/answers that will enhance academic performance and improve institutional success

Our study is to focus on student completion rate in a specific subject based on their performance of test result components during the performance by applying the Naive Bayes Classification algorithm.

Figure 16

Naïve Bayes Model in Training Set

```
=== Classifier model (full training set) ===

Naive Bayes Classifier

                                                Class
Attribute                                 Completer Non Complete
                                             (0.67)       (0.33)
=============================================================
TransportType
  Private Transport                           428.0        192.0
  N/A                                        1186.0        612.0
  Public Transport                            394.0        194.0
  [total]                                    2008.0        998.0

EducationalAttainment
  Junior High School Graduate                 145.0         80.0
  Senior High School Ongoing                  145.0         81.0
  Junior High School Ongoing                  144.0         78.0
  Elementary Undergraduate                    165.0         77.0
  Elementary Ongoing                          148.0         80.0
  Senior High School Graduate                 157.0         72.0
  College Ongoing                             176.0         86.0
  Senior HighSchool Undergraduate             157.0         86.0
  College Graduate                            169.0         71.0
  Elementary Graduate                         136.0         85.0
  N/A                                         134.0         79.0
  College Undergraduate                       171.0         67.0
  Junior High School Undergraduate            171.0         66.0
  [total]                                    2018.0       1008.0

Proximity
  1 - 15 km                                   571.0        237.0
  31 + km                                    1186.0        612.0
  16 - 30 km                                  251.0        149.0
  [total]                                    2008.0        998.0

Purpose
  For Employment Purposes                     647.0        339.0
  Requirements in School                      670.0        342.0
  Hobby                                       691.0        317.0
  [total]                                    2008.0        998.0

Profeciency
  Advanced                                   1005.0          1.0
  Beginner                                    496.0        481.0
  Intermediate                                507.0        516.0
  [total]                                    2008.0        998.0


Time taken to build model: 0 seconds
```

This output displays the findings from analyzing the training dataset using the Naive Bayes classifier model. Using the outcomes of the five attributes "TransportType", "EducationalAttainment", "Proximity", "Purpose", and "Proficiency", the model is trained to predict whether a data sample falls to the "Complete" or "Non Completer" class. The table lists the total number of samples in each class as well as the number of samples in each class for each attribute. The proportion of correctly identified samples in the "Complete" class shows that the model was able to attain an accuracy of 67% on the training set.

Figure 17

Naïve Bayes Validation in Training Set

```
=== Summary ===

Correctly Classified Instances        2097               69.9    %
Incorrectly Classified Instances       903               30.1    %
Kappa statistic                          0.3294
Mean absolute error                      0.3296
Root mean squared error                  0.4061
Relative absolute error                 74.3494 %
Root relative squared error             86.2504 %
Total Number of Instances             3000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.763    0.429    0.782      0.763    0.772      0.330    0.776     0.899     Completer
                 0.571    0.237    0.544      0.571    0.557      0.330    0.776     0.538     Non Complete
Weighted Avg.    0.699    0.366    0.703      0.699    0.701      0.330    0.776     0.779
```

The Naive Bayes classifier model scored 69.9% on the training set. 2097 of 3000 training set occurrences were successfully classified, whereas 903 were misclassified. Kappa was 0.3294, suggesting reasonable agreement between anticipated and actual classes. The model's predictions averaged 0.33 units off due to the mean absolute error of 0.3296 and the root mean squared error of 0.4061. The model's relative absolute error and root relative squared error were 74.3494% and 86.2504%, respectively, which were excessive given the target attribute's domain.

The model scored best in predicting the Completer class with a true positive rate of 0.763, precision of 0.782, and recall of 0.763. The model had 0.429 false positives for the Completer class. The model's Non-Completer class performance was poorer, with a true positive rate of 0.571, precision of 0.544, and recall of 0.571. Non-Completer had 0.237 false positives. The model's performance was balanced across the two classes because the weighted average of the true positive rate, precision, recall, and F-measure was 0.699. The ROC and precision-recall curve areas were 0.776 and 0.779, respectively.

Figure 18

Naïve Bayes Confusion Matrix in Training Set

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1529  476 |    a = Completer
  427  568 |    b = Non Complete
```

The confusion matrix shows that the model correctly identified 1529 occurrences as Completer and misclassified 476 as Non-Completer. It properly categorized 568 occurrences as non-completer and misidentified 427 as complete. The model has a greater true positive rate and precision for categorizing occurrences as Completer than Non-Completer.

Figure 19

Naïve Bayes Model in Supplied Test Set

```
=== Classifier model (full training set) ===

Naive Bayes Classifier

                                          Class
Attribute                           Completer Non Complete
                                      (0.67)       (0.33)
=================================================================
TransportType
  Private Transport                    428.0        192.0
  N/A                                 1186.0        612.0
  Public Transport                     394.0        194.0
  [total]                             2008.0        998.0

EducationalAttainment
  Junior High School Graduate          145.0         80.0
  Senior High School Ongoing           145.0         81.0
  Junior High School Ongoing           144.0         78.0
  Elementary Undergraduate             165.0         77.0
  Elementary Ongoing                   148.0         80.0
  Senior High School Graduate          157.0         72.0
  College Ongoing                      176.0         86.0
  Senior HighSchool Undergraduate      157.0         86.0
  College Graduate                     169.0         71.0
  Elementary Graduate                  136.0         85.0
  N/A                                  134.0         79.0
  College Undergraduate                171.0         67.0
  Junior High School Undergraduate     171.0         66.0
  [total]                             2018.0       1008.0

Proximity
  1 - 15 km                            571.0        237.0
  31 + km                             1186.0        612.0
  16 - 30 km                           251.0        149.0
  [total]                             2008.0        998.0

Purpose
  For Employment Purposes              647.0        339.0
  Requirements in School               670.0        342.0
  Hobby                                691.0        317.0
  [total]                             2008.0        998.0

Profeciency
  Advanced                            1005.0          1.0
  Beginner                             496.0        481.0
  Intermediate                         507.0        516.0
  [total]                             2008.0        998.0


Time taken to build model: 0 seconds
```

Naive Bayes model trained on 3000 occurrences and 6 attributes. TransportType, EducationalAttainment, Proximity, Purpose, Proficiency, and Outcome. The classifier model is trained on the whole training set and assessed on a user-supplied test set of undetermined size.

Figure 20

Naïve Bayes Validation Result in Supplied Test Set

```
=== Summary ===

Correctly Classified Instances         421               70.1667 %
Incorrectly Classified Instances       179               29.8333 %
Kappa statistic                          0.3529
Mean absolute error                      0.3355
Root mean squared error                  0.4086
Relative absolute error                 75.1951 %
Root relative squared error             86.2479 %
Total Number of Instances              600

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.745    0.382    0.791      0.745    0.767      0.354    0.781     0.897     Completer
               0.618    0.255    0.555      0.618    0.585      0.354    0.781     0.568     Non Complete
Weighted Avg.  0.702    0.339    0.711      0.702    0.705      0.354    0.781     0.785

=== Confusion Matrix ===
```

Figure 21

Naïve Bayes Confusion Matrix in Supplied Test Set

```
=== Confusion Matrix ===

   a   b   <-- classified as
 295 101 |   a = Completer
  78 126 |   b = Non Complete
```

The confusion matrix and evaluation measures gave the Naive Bayes classifier 70.1667% accuracy. 421 cases were correctly identified (70.1667%) and 179 wrongly classified (29.8333%). Predicted and actual results agree moderately, according to Kappa 0.3529.

Comparing class-specific accuracy, the Completer class has a higher true positive rate (0.745) than the Non Completer class (0.618). The classifier correctly identifies Completers better than Non Completers. The classifier is more precise when it identifies Completer (0.791) than Non
Completer (0.555). The classifier performs similarly for both classes because its ROC and PRC regions are 0.781. The modest MCC (Matthews correlation coefficient) of 0.354 shows a favorable connection between predicted and actual classes.

Figure 22

Naïve Bayes Model in 10 Fold Cross Validation

```
=== Classifier model (full training set) ===

Naive Bayes Classifier

                                        Class
Attribute                       Completer Non Complete
                                   (0.67)       (0.33)
=================================================================
TransportType
  Private Transport                 428.0        192.0
  N/A                              1186.0        612.0
  Public Transport                  394.0        194.0
  [total]                          2008.0        998.0

EducationalAttainment
  Junior High School Graduate       145.0         80.0
  Senior High School Ongoing        145.0         81.0
  Junior High School Ongoing        144.0         78.0
  Elementary Undergraduate          165.0         77.0
  Elementary Ongoing                148.0         80.0
  Senior High School Graduate       157.0         72.0
  College Ongoing                   176.0         86.0
  Senior HighSchool Undergraduate   157.0         86.0
  College Graduate                  169.0         71.0
  Elementary Graduate               136.0         85.0
  N/A                               134.0         79.0
  College Undergraduate             171.0         67.0
  Junior High School Undergraduate  171.0         66.0
  [total]                          2018.0       1008.0
```

Proximity		
1 - 15 km	571.0	237.0
31 + km	1186.0	612.0
16 - 30 km	251.0	149.0
[total]	2008.0	998.0
Purpose		
For Employment Purposes	647.0	339.0
Requirements in School	670.0	342.0
Hobby	691.0	317.0
[total]	2008.0	998.0
Profeciency		
Advanced	1005.0	1.0
Beginner	496.0	481.0
Intermediate	507.0	516.0
[total]	2008.0	998.0

The model forecasts that 428 students who used private transport graduated, whereas 192 did not. The model forecasts College Graduates (169), Junior High School Graduates (145), and Senior High School Graduates (157). Junior High Undergraduates (171) completed the fewest.

The model predicts that students who reside 31+ kilometres from the university had the most non-completers (612) and the most completers (571).

The model anticipates that the most completers utilized transportation for hobby purposes (691), followed by school requirements (670) and employment (647).

Moreover, the model predicts that the majority of completers (1005) had advanced competence, while the majority of non-completers (481) had beginner skill.

Figure 23

Naïve Bayes Validation Result in 10-Fold Cross Validation

```
=== Summary ===

Correctly Classified Instances        2047               68.2333 %
Incorrectly Classified Instances       953               31.7667 %
Kappa statistic                          0.288
Mean absolute error                      0.332
Root mean squared error                  0.409
Relative absolute error                 74.8728 %
Root relative squared error             86.8816 %
Total Number of Instances             3000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.756    0.466    0.766      0.756    0.761      0.288    0.758     0.892     Completer
                 0.534    0.244    0.521      0.534    0.527      0.288    0.758     0.500     Non Complete
Weighted Avg.    0.682    0.393    0.684      0.682    0.683      0.288    0.758     0.762
```

10-fold cross-validation results in a model accuracy of 68.2333%. Kappa is 0.288, which shows that the projected and actual classes correspond fairly well. The mean absolute and root mean squared errors are 0.332 and 0.409, respectively. The relative absolute and root relative squared errors are 74.8728% and 86.8816%, correspondingly.

The non-completer class has a lower true positive rate (TPR) of 0.534 whereas the completer class has a higher TPR of 0.756. False positive rates (FPR) were higher for non-completers (0.244) than for completers (0.466). Instances that the model predicts as Completer are more likely to be accurate (0.766 vs. 0.521). The recall is the TPR for each class. Completers (0.761) had a higher F-measure than Non Completers (0.527), which is a weighted harmonic mean of precision and recall. A respectable correlation between expected and actual classes is shown by the MCC, which is 0.288. In the ROC and PRC curves of the model, the Completer class performs better than the Non Completer class. Precision, recall, and Fmeasure exhibited higher weighted averages for completers (0.684) than for non-completers (0.683).

Figure 24

Naïve Bayes Confusion Matrix in 10-Fold Cross Validation

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1516  489 |    a = Completer
  464  531 |    b = Non Complete
```

1516 instances were accurately identified as completers by the model, while 531 examples were labeled as incomplete. However, 489 occurrences were wrongly labeled as Non-Completer when they were Completer, while 464 instances were mislabeled as Completer when they weren't.

Figure 25

Naïve Bayes Model in 10% Split

```
=== Classifier model (full training set) ===

Naive Bayes Classifier

                                                Class
Attribute                                 Completer Non Complete
                                             (0.67)       (0.33)
=============================================================================
TransportType
  Private Transport                           428.0        192.0
  N/A                                        1186.0        612.0
  Public Transport                            394.0        194.0
  [total]                                    2008.0        998.0

EducationalAttainment
  Junior High School Graduate                 145.0         80.0
  Senior High School Ongoing                  145.0         81.0
  Junior High School Ongoing                  144.0         78.0
  Elementary Undergraduate                    165.0         77.0
  Elementary Ongoing                          148.0         80.0
  Senior High School Graduate                 157.0         72.0
  College Ongoing                             176.0         86.0
  Senior HighSchool Undergraduate             157.0         86.0
  College Graduate                            169.0         71.0
  Elementary Graduate                         136.0         85.0
  N/A                                         134.0         79.0
  College Undergraduate                       171.0         67.0
  Junior High School Undergraduate            171.0         66.0
  [total]                                    2018.0       1008.0

Proximity
  1 - 15 km                                   571.0        237.0
  31 + km                                    1186.0        612.0
  16 - 30 km                                  251.0        149.0
  [total]                                    2008.0        998.0

Purpose
  For Employment Purposes                     647.0        339.0
  Requirements in School                      670.0        342.0
  Hobby                                       691.0        317.0
  [total]                                    2008.0        998.0

Profeciency
  Advanced                                   1005.0          1.0
  Beginner                                    496.0        481.0
  Intermediate                                507.0        516.0
  [total]                                    2008.0        998.0


Time taken to build model: 0 seconds
```

As an illustration, the TransportType attribute has values of Private Transport (428 Completer instances and 192 Non-Completer instances), N/A (1186 Completer instances and 612 Non Completer instances), and Public Transport (394 Completer instances and 194 Non Completer instances). 2008 total

Completer instances with any value for TransportType, and 998 total Non-Completer instances with any value for TransportType. The probability is 0.67 and 0.33 for the Completer and Non-Completer classes.

Figure 26

Naïve Bayes Validation Result in 10% Split

```
=== Summary ===

Correctly Classified Instances        1811               67.0741 %
Incorrectly Classified Instances       889               32.9259 %
Kappa statistic                          0.2695
Mean absolute error                      0.3332
Root mean squared error                  0.4164
Relative absolute error                 73.4977 %
Root relative squared error             88.4173 %
Total Number of Instances             2700

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.734     0.458     0.766      0.734     0.750      0.270    0.757     0.892     Completer
               0.542     0.266     0.498      0.542     0.519      0.270    0.757     0.506     Non Complete
Weighted Avg.  0.671     0.395     0.678      0.671     0.674      0.270    0.757     0.765
```

90% of the dataset was used to train the model, while the remaining 10% was used to test it. The summary indicates that 67.0741% of the occurrences were correctly classified by the model, whereas 32.9259% of the instances were wrongly classified. The classifier's predictions and the actual labels have a reasonable amount of agreement, according to the Kappa statistic of 0.2695.
According to the detailed accuracy by class, the true positive rate (TPR) and false positive rate (FPR) for the Completer class are 0.734 and 0.458, respectively. The Completer class has a precision of 0.766 and a recall of 0.734. The Matthews correlation coefficient (MCC) is 0.270, while the F-measure is 0.750. Precision-Recall Curve (PRC) area is 0.892, and Receiver Operating Characteristic (ROC) area is 0.757. The TPR and FPR for the Non Completer class are 0.542 and 0.266, respectively. The Non Completer class has a precision of 0.498 and a recall of 0.542. The MCC is 0.270, while the F-measure is 0.519. The areas of the PRC are 0.506 and the ROC are 0.757.

It also includes the weighted average of the TPR, FPR, precision, recall, F-measure, MCC, ROC area, and PRC area. Precision, recall, and F-measure have weighted averages of 0.678, 0.671, and 0.674,

Figure 27

Naïve Bayes Confusion Matrix in 10% Split

```
=== Confusion Matrix ===

   a    b   <-- classified as
1331  483 |    a = Completer
 406  480 |    b = Non Complete
```

According to the confusion matrix, there were 1331 Completer instances that were correctly identified as such (true positives), 483 Completer instances that were misclassified as Non-Completer (false negatives), 406 Non Completer instances that were misclassified as Completer (false positives), and 480 Non Completer instances that were identified as such (true negatives).

Moving on, J48 algorithm is one of the most widely used machine learning algorithms to examine the data categorically and continuously. The C4.5 algorithm (J48) is mostly used among many fields for classifying data. It produces decision trees based on information theory. It is an extension of Ross Quinlan's

earlier ID3 algorithm also known in Weka as J48, J standing for Java. The decision trees generated by C4.5 are used for classification, and for this reason, C4.5 is often referred to as a statistical classifier.

According to Tun, M. M., & Htay, Y. Y. (2020). “Predict Students’ Performance by Using J48 Algorithm”. The critical issue to the community is to monitor the progress of students’ performance. We can use data mining techniques for this purpose. J48 algorithm is one of the famous classification algorithms present today to generate decision trees in data mining technique. Weka machine learning tool is applied to make classification. In this paper, we tested result classification accuracy was computed. This J48 classification algorithm give accuracy with 78.2%. The prediction of students’ performance in institution has become one of the most important needs of that institute to improve the quality of the teaching process of that institution. In this process, we get to know the needs of the students and hence we can fulfil those needs to get better results. Students who need special attention from the teachers can also be identified from this process.

Our study aims to create a similar predictive model using J48 with ABC Company. Classifying students based on pre-enrollment information using J48 and the rules presented for each node would allow the administrative and academic staff to identify students who would be “at-risk” of dropping the course even before they start with their study. Then the student support systems, such as orientation, advising, and mentoring programs, could be used to positively impact the academic successes of such students.

Figure 28

J48 Model in Training Set

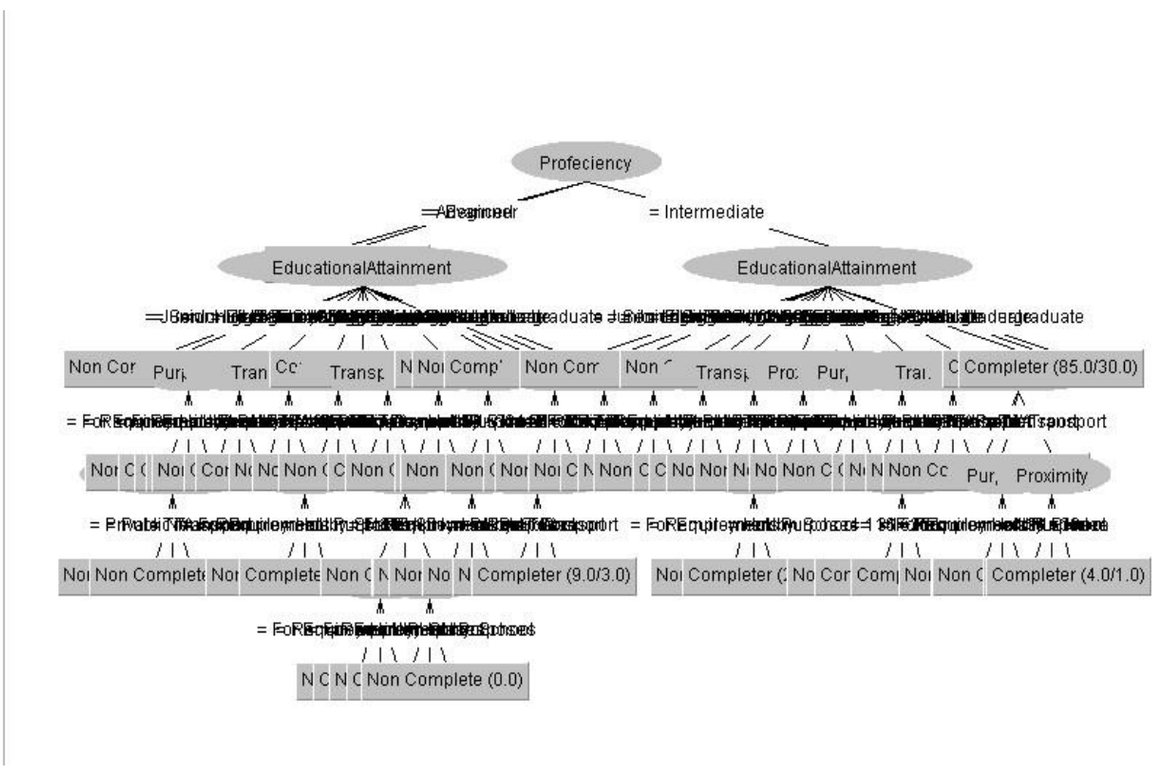


Figure 28.1

J48 Model in Training Set

```
Profeciency = Advanced: Completer (1004.0)
Profeciency = Beginner
|   EducationalAttainment = Junior High School Graduate: Non Complete (79.0/34.0)
|   EducationalAttainment = Senior High School Ongoing
|   |   Purpose = For Employment Purposes
|   |   |   TransportType = Private Transport: Non Complete (5.0/1.0)
|   |   |   TransportType = N/A: Completer (9.0/3.0)
|   |   |   TransportType = Public Transport: Non Complete (7.0/3.0)
|   |   Purpose = Requirements in School: Non Complete (19.0/4.0)
|   |   Purpose = Hobby: Completer (18.0/6.0)
|   EducationalAttainment = Junior High School Ongoing
|   |   Purpose = For Employment Purposes: Completer (25.0/9.0)
|   |   Purpose = Requirements in School: Completer (29.0/12.0)
|   |   Purpose = Hobby: Non Complete (22.0/6.0)
|   EducationalAttainment = Elementary Undergraduate
|   |   TransportType = Private Transport: Completer (22.0/8.0)
|   |   TransportType = N/A: Completer (46.0/16.0)
|   |   TransportType = Public Transport
|   |   |   Purpose = For Employment Purposes: Completer (3.0/1.0)
|   |   |   Purpose = Requirements in School: Non Complete (8.0/1.0)
|   |   |   Purpose = Hobby: Completer (4.0/1.0)
|   EducationalAttainment = Elementary Ongoing
|   |   Proximity = 1 - 15 km: Completer (22.0/8.0)
|   |   Proximity = 31 + km: Non Complete (35.0/14.0)
|   |   Proximity = 16 - 30 km: Non Complete (12.0/4.0)
|   EducationalAttainment = Senior High School Graduate: Completer (69.0/27.0)
|   EducationalAttainment = College Ongoing
|   |   TransportType = Private Transport: Completer (9.0/3.0)
|   |   TransportType = N/A: Non Complete (59.0/25.0)
|   |   TransportType = Public Transport
|   |   |   Proximity = 1 - 15 km
|   |   |   |   Purpose = For Employment Purposes: Completer (3.0)
```

Figure 28.2

J48 Model in Training Set

```
|   |   |   |   Purpose = Requirements in School: Non Complete (2.0)
|   |   |   |   Purpose = Hobby: Completer (2.0/1.0)
|   |   |   Proximity = 31 + km: Non Complete (0.0)
|   |   |   Proximity = 16 - 30 km
|   |   |   |   Purpose = For Employment Purposes: Non Complete (3.0)
|   |   |   |   Purpose = Requirements in School: Completer (3.0/1.0)
|   |   |   |   Purpose = Hobby: Non Complete (0.0)
|   EducationalAttainment = Senior HighSchool Undergraduate
|   |   Purpose = For Employment Purposes: Completer (27.0/10.0)
|   |   Purpose = Requirements in School: Completer (27.0/12.0)
|   |   Purpose = Hobby: Non Complete (34.0/12.0)
|   EducationalAttainment = College Graduate
|   |   Proximity = 1 - 15 km
|   |   |   Purpose = For Employment Purposes: Completer (9.0/1.0)
|   |   |   Purpose = Requirements in School: Non Complete (6.0/2.0)
|   |   |   Purpose = Hobby: Non Complete (6.0/1.0)
|   |   Proximity = 31 + km: Completer (50.0/19.0)
|   |   Proximity = 16 - 30 km: Non Complete (10.0/4.0)
|   EducationalAttainment = Elementary Graduate: Non Complete (74.0/35.0)
|   EducationalAttainment = N/A: Non Complete (77.0/35.0)
|   EducationalAttainment = College Undergraduate: Completer (69.0/30.0)
|   EducationalAttainment = Junior High School Undergraduate
|   |   Proximity = 1 - 15 km
|   |   |   TransportType = Private Transport: Non Complete (8.0/2.0)
|   |   |   TransportType = N/A: Non Complete (0.0)
|   |   |   TransportType = Public Transport: Completer (9.0/3.0)
|   |   Proximity = 31 + km: Non Complete (40.0/19.0)
|   |   Proximity = 16 - 30 km: Completer (14.0/5.0)
Profeciency = Intermediate
|   EducationalAttainment = Junior High School Graduate
|   |   TransportType = Private Transport: Completer (18.0/8.0)
|   |   TransportType = N/A: Non Complete (38.0/15.0)
|   |   TransportType = Public Transport: Completer (11.0/3.0)
|   EducationalAttainment = Senior High School Ongoing: Non Complete (74.0/26.0)
|   EducationalAttainment = Junior High School Ongoing
|   |   Purpose = For Employment Purposes: Non Complete (29.0/12.0)
|   |   Purpose = Requirements in School: Completer (29.0/12.0)
|   |   Purpose = Hobby: Completer (22.0/11.0)
|   EducationalAttainment = Elementary Undergraduate
|   |   Purpose = For Employment Purposes: Non Complete (30.0/14.0)
|   |   Purpose = Requirements in School: Non Complete (31.0/12.0)
|   |   Purpose = Hobby: Completer (25.0/8.0)
|   EducationalAttainment = Elementary Ongoing: Non Complete (79.0/37.0)
|   EducationalAttainment = Senior High School Graduate
|   |   TransportType = Private Transport: Completer (16.0/7.0)
|   |   TransportType = N/A
|   |   |   Purpose = For Employment Purposes: Completer (18.0/8.0)
|   |   |   Purpose = Requirements in School: Non Complete (15.0/6.0)
|   |   |   Purpose = Hobby: Completer (20.0/8.0)
|   |   TransportType = Public Transport: Non Complete (19.0/7.0)
|   EducationalAttainment = College Ongoing
```

Figure 28.3

J48 Model in Training Set

```
|   |   Proximity = 1 - 15 km: Completer (22.0/10.0)
|   |   Proximity = 31 + km: Non Complete (48.0/20.0)
|   |   Proximity = 16 - 30 km: Completer (12.0/3.0)
|   EducationalAttainment = Senior HighSchool Undergraduate
|   |   Purpose = For Employment Purposes: Non Complete (32.0/14.0)
|   |   Purpose = Requirements in School: Non Complete (25.0/8.0)
|   |   Purpose = Hobby: Completer (19.0/6.0)
|   EducationalAttainment = College Graduate
|   |   Purpose = For Employment Purposes: Non Complete (27.0/10.0)
|   |   Purpose = Requirements in School
|   |   |   Proximity = 1 - 15 km: Non Complete (6.0/1.0)
|   |   |   Proximity = 31 + km: Completer (17.0/7.0)
|   |   |   Proximity = 16 - 30 km: Completer (5.0)
|   |   Purpose = Hobby: Completer (17.0/6.0)
|   EducationalAttainment = Elementary Graduate
|   |   TransportType = Private Transport: Completer (14.0/5.0)
|   |   TransportType = N/A: Non Complete (51.0/20.0)
|   |   TransportType = Public Transport: Non Complete (12.0/3.0)
|   EducationalAttainment = N/A
|   |   TransportType = Private Transport: Non Complete (12.0/5.0)
|   |   TransportType = N/A
|   |   |   Purpose = For Employment Purposes: Completer (14.0/6.0)
|   |   |   Purpose = Requirements in School: Non Complete (15.0/6.0)
|   |   |   Purpose = Hobby: Completer (16.0/6.0)
|   |   TransportType = Public Transport
|   |   |   Proximity = 1 - 15 km: Non Complete (12.0/5.0)
|   |   |   Proximity = 31 + km: Completer (0.0)
|   |   |   Proximity = 16 - 30 km: Completer (4.0/1.0)
|   EducationalAttainment = College Undergraduate: Completer (82.0/36.0)
|   EducationalAttainment = Junior High School Undergraduate: Completer (85.0/30.0)

Number of Leaves  :     83

Size of the tree :      114
```

This is a trained J48 decision tree model that classifies new instances based on their attribute values. The tree is constructed by iteratively dividing the dataset into ever-smaller subsets, based on the attribute values, until each subset contains instances of the same class. Following a route from the tree's root to a leaf node that corresponds to the predicted class, the tree can then be utilized to classify the latest instances.

Every internal node of the tree reflects a divide on an attribute, and the branches coming from each node serve any potential outcomes for that attribute. The leaf nodes represent the class prediction for instances that end up in that leaf. The numbers in parentheses next to each leaf node indicate the number of instances in the training set that correspond to that leaf and the quantity of instances that were correctly classified.

The J48 pruned tree algorithm was applied to the dataset to produce the outcome seen above. Based on factors including educational attainment, purpose, mode of transportation, and proximity to a learning institution.

The proficiency level is the root of the tree, from which it sprouts the intermediate/beginner and advanced branches. If the proficiency level is advanced, it is expected that the student will finish the program. The tree further divides based on level of schooling if the competency level is intermediate/beginner.

The tree divides once more according to other elements including proximity to a learning institution, purpose, and mode of transportation for each category of educational achievement. The resulting tree leaves show the expected proficiency level and whether or not the student is likely to finish the course.

As a result, the J48 pruned tree method generates a decision tree that estimates the likelihood that a course will be completed based on a variety of variables, including educational attainment, proficiency level, proximity to the learning institution, purpose, and mode of transportation.

Consider, for instance, the first line of the tree: Proficiency = Advanced: Completer (1004.0) This signifies that if the "Proficiency" attribute of an instance has the value "Advanced", it will be classified as "Completer". There were 1004 instances in the training set that belonged to this leaf node, and they were all correctly classified.

The tree also contains divisions on other attributes, such as "EducationalAttainment" and "TransportType", and the combinations of attribute values that lead to each leaf node can be determined by tracing the path from the tree's root to each leaf. For instance, if the instance has the value "Elementary Undergraduate" for "EducationalAttainment" and "Private Transport" for "TransportType", it will end up in the leaf node with the label "Completer" and the count (22.0/8.0). This indicates that there were 22 instances in the training set with these attribute values, and 8 of them were incorrectly classified.

Figure 29

J48 Validation Result in Training Set

```
=== Summary ===

Correctly Classified Instances        2220               74      %
Incorrectly Classified Instances       780               26      %
Kappa statistic                          0.4232
Mean absolute error                      0.3101
Root mean squared error                  0.3937
Relative absolute error                 69.9366 %
Root relative squared error             83.6316 %
Total Number of Instances             3000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.789    0.359    0.816      0.789    0.802      0.424    0.823     0.916     Completer
                 0.641    0.211    0.601      0.641    0.621      0.424    0.823     0.632     Non Complete
Weighted Avg.    0.740    0.310    0.745      0.740    0.742      0.424    0.823     0.822
```

The success rate of a machine learning model on a binary classification task with two classes—Completer and Non-Completer—is displayed in the assessment report. The model had an overall accuracy of 74% after being trained on a dataset of 3000 examples, and it was able to accurately classify 2220 instances.

This demonstrates that 789 out of 1000 Completer instances were properly detected by the model, yielding a True Positive Rate (Recall) of 0.789. It also misclassified 359 out of 1000 occurrences of Non-Completer as Completer, yielding a False Positive Rate of 0.359. Since the model's classification accuracy for Completer instances was 0.816, 81.6% of all Completer instances were truly Completer. Similar to that, the Precision for instances classified as Non-Completer was 0.601, meaning that 60.1% of all instances classified as Non Completer were in fact Non-Completer.

The model's Kappa statistic, which measures the degree of agreement between the model's predictions and the actual class labels, was 0.4232. The model's mean absolute error was 0.3101, which indicates that the model generally underestimated the probability of the positive class by 0.31. The model exhibited a greater error rate for cases that were further away from the decision boundary, as indicated by the Root Mean Squared Error of 0.3937.

Overall, the model worked well, with acceptable precision and accuracy for both classes.

Figure 30

J48 Confusion Matrix in Training Set

```
=== Confusion Matrix ===

   a    b   <-- classified as
1582  423 |    a = Completer
 357  638 |    b = Non Complete
```

In the confusion matrix, the quantity of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for each class serves as a summary of how well the binary classification model performed on the test data. In this case, the model identified 1582 instances as Completers (class a), of which 1439 were accurately identified as such (TP), while 423 were incorrectly identified as NonCompleter (FP). The model correctly recognized 638 instances as Non-Completer (class b), while 357 instances were incorrectly classified as Completers (FN).

The confusion matrix reveals that the model had an overall accuracy of 74% and performed better in categorizing Completers (class a) than Non Completers (class b). In further detail, the model correctly identified 79% of Completers (true positive rate or recall) but only 64% of Non-Completers (recall), demonstrating that the model is more effective at identifying people who will complete the program than those who won't. The model's accuracy, which gauges the percentage of accurate positive predictions, was higher for Completers (81%) than for Non-Completers (60%).

Figure 31

J48 Model in Supplied Test Set

```
|   |   [illegible]
|   EducationalAttainment = Elementary Graduate: Non Complete (74.0/35.0)
|   EducationalAttainment = N/A: Non Complete (77.0/35.0)
|   EducationalAttainment = College Undergraduate: Completer (69.0/30.0)
|   EducationalAttainment = Junior High School Undergraduate
|   |   Proximity = 1 - 15 km
|   |   |   TransportType = Private Transport: Non Complete (8.0/2.0)
|   |   |   TransportType = N/A: Non Complete (0.0)
|   |   |   TransportType = Public Transport: Completer (9.0/3.0)
|   |   Proximity = 31 + km: Non Complete (40.0/19.0)
|   |   Proximity = 16 - 30 km: Completer (14.0/5.0)
Profeciency = Intermediate
|   EducationalAttainment = Junior High School Graduate
|   |   TransportType = Private Transport: Completer (18.0/8.0)
|   |   TransportType = N/A: Non Complete (38.0/15.0)
|   |   TransportType = Public Transport: Completer (11.0/3.0)
|   EducationalAttainment = Senior High School Ongoing: Non Complete (74.0/26.0)
|   EducationalAttainment = Junior High School Ongoing
|   |   Purpose = For Employment Purposes: Non Complete (29.0/12.0)
|   |   Purpose = Requirements in School: Completer (29.0/12.0)
|   |   Purpose = Hobby: Completer (22.0/11.0)
|   EducationalAttainment = Elementary Undergraduate
|   |   Purpose = For Employment Purposes: Non Complete (30.0/14.0)
|   |   Purpose = Requirements in School: Non Complete (31.0/12.0)
|   |   Purpose = Hobby: Completer (25.0/8.0)
|   EducationalAttainment = Elementary Ongoing: Non Complete (79.0/37.0)
|   EducationalAttainment = Senior High School Graduate
|   |   TransportType = Private Transport: Completer (16.0/7.0)
|   |   TransportType = N/A
|   |   |   Purpose = For Employment Purposes: Completer (18.0/8.0)
|   |   |   Purpose = Requirements in School: Non Complete (15.0/6.0)
|   |   |   Purpose = Hobby: Completer (20.0/8.0)
|   |   TransportType = Public Transport: Non Complete (19.0/7.0)
|   EducationalAttainment = College Ongoing
|   |   Proximity = 1 - 15 km: Completer (22.0/10.0)
|   |   Proximity = 31 + km: Non Complete (48.0/20.0)
|   |   Proximity = 16 - 30 km: Completer (12.0/3.0)
|   EducationalAttainment = Senior HighSchool Undergraduate
|   |   Purpose = For Employment Purposes: Non Complete (32.0/14.0)
|   |   Purpose = Requirements in School: Non Complete (25.0/8.0)
|   |   Purpose = Hobby: Completer (19.0/6.0)
|   EducationalAttainment = College Graduate
|   |   Purpose = For Employment Purposes: Non Complete (27.0/10.0)
|   |   Purpose = Requirements in School
|   |   |   Proximity = 1 - 15 km: Non Complete (6.0/1.0)
|   |   |   Proximity = 31 + km: Completer (17.0/7.0)
|   |   |   Proximity = 16 - 30 km: Completer (5.0)
|   |   Purpose = Hobby: Completer (17.0/6.0)
|   EducationalAttainment = Elementary Graduate
|   |   TransportType = Private Transport: Completer (14.0/5.0)
|   |   TransportType = N/A: Non Complete (51.0/20.0)
|   |   TransportType = Public Transport: Non Complete (12.0/3.0)
```

Figure 31.1

J48 Model in Supplied Test Set

```
|   |   TransportType = Public Transport: Non Complete (12.0/5.0)
|   EducationalAttainment = N/A
|   |   TransportType = Private Transport: Non Complete (12.0/5.0)
|   |   TransportType = N/A
|   |   |   Purpose = For Employment Purposes: Completer (14.0/6.0)
|   |   |   Purpose = Requirements in School: Non Complete (15.0/6.0)
|   |   |   Purpose = Hobby: Completer (16.0/6.0)
|   |   TransportType = Public Transport
|   |   |   Proximity = 1 - 15 km: Non Complete (12.0/5.0)
|   |   |   Proximity = 31 + km: Completer (0.0)
|   |   |   Proximity = 16 - 30 km: Completer (4.0/1.0)
|   EducationalAttainment = College Undergraduate: Completer (82.0/36.0)
|   EducationalAttainment = Junior High School Undergraduate: Completer (85.0/30.0)

Number of Leaves  :     83

Size of the tree :      114
```

The tree is divided into three categories: beginner, intermediate, and advanced. The choice is provided for novices based on their level of education. If they are a senior high school student, they can be neither ongoing or a graduate, and a choice is made depending on their purpose of studying, proximity to the school, and transport type. If they are a junior high school graduate, the predicted completeness of their transport mode is non-completer. The decision is made for advanced level users based on their educational level, the purpose of their studying, the proximity to their school, and the mode of transportation.

Figure 32

J48 Validation Result in Supplied Test Set

```
=== Summary ===

Correctly Classified Instances         449               74.8333 %
Incorrectly Classified Instances       151               25.1667 %
Kappa statistic                          0.4567
Mean absolute error                      0.3182
Root mean squared error                  0.3966
Relative absolute error                 71.3157 %
Root relative squared error             83.7037 %
Total Number of Instances              600

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.775     0.304     0.832      0.775     0.803      0.459     0.822     0.913     Completer
               0.696     0.225     0.615      0.696     0.653      0.459     0.822     0.641     Non Complete
Weighted Avg.  0.748     0.277     0.758      0.748     0.752      0.459     0.822     0.821
```

This is a summary of the model's evaluation using the supplied test data. A total of 449 out of 600 cases were properly identified by the model, yielding a 74.83% overall precision. The match between the anticipated and actual class labels was moderate, as indicated by the kappa statistic, which measures it, of 0.4567.

In comparison to the Non Completer class, which had a precision of 0.615 and recall of 0.696, the model did better in categorizing the Completer class, with a precision of 0.832 and recall of 0.775. The model did fairly well on both classes, as evidenced by the weighted average of precision, recall, and Fmeasure, which was 0.758.

The root mean squared error was 0.3966, while the mean absolute error was 0.3182. The relative absolute error was 71.3157%, and the root relative squared error was 83.7037%. Lower values represent better performance. These metrics provide the average absolute and comparative difference between the anticipated and actual class frequencies.

Figure 33

J48 Confusion Matrix in Supplied Test Set

```
=== Confusion Matrix ===

   a   b   <-- classified as
 307  89 |   a = Completer
  62 142 |   b = Non Complete
```

307 Completer and 142 Non Completer confusion matrix instances were properly identified by the model. However, 89 Completers were Non Completers, and 62 Non Completers were Completers. Although there is place for improvement in terms of lowering inaccurate results and false positives, the model does a better job of predicting Completers than Non Completers.

Figure 34

J48 Model in 10 Fold Cross Validation

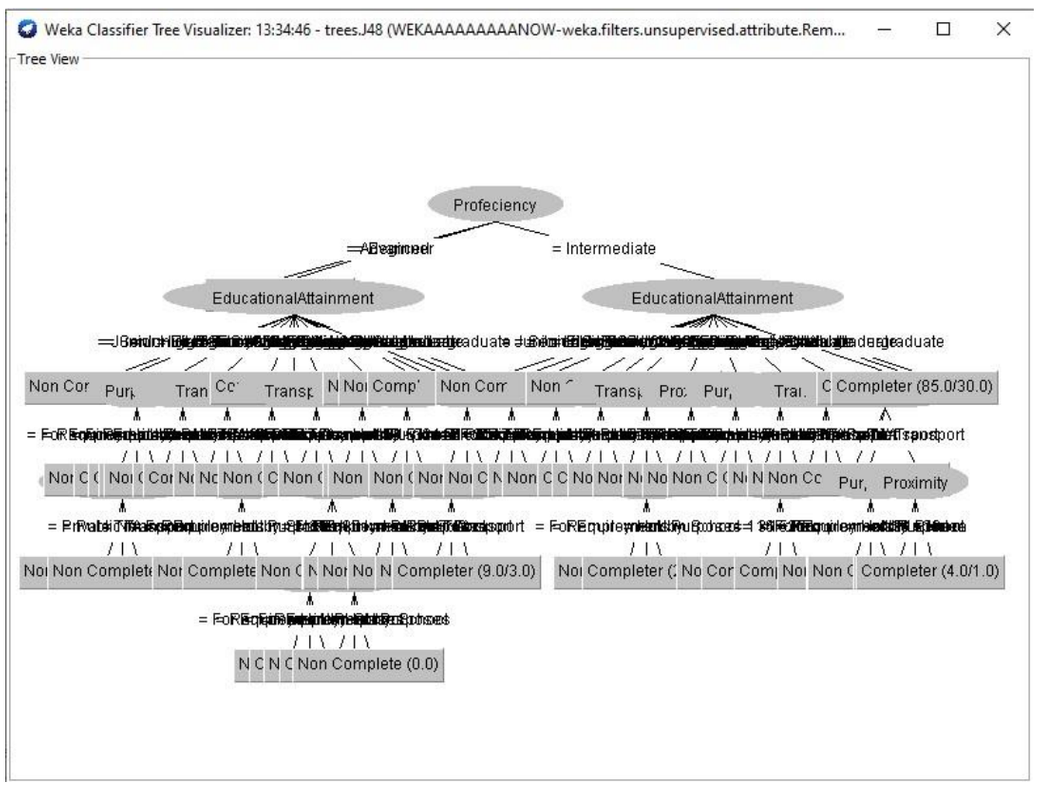


Figure 34.1

J48 Model in 10 Fold Cross Validation

```
Profeciency = Advanced: Completer (1004.0)
Profeciency = Beginner
|   EducationalAttainment = Junior High School Graduate: Non Complete (79.0/34.(
|   EducationalAttainment = Senior High School Ongoing
|   |   Purpose = For Employment Purposes
|   |   |   TransportType = Private Transport: Non Complete (5.0/1.0)
|   |   |   TransportType = N/A: Completer (9.0/3.0)
|   |   |   TransportType = Public Transport: Non Complete (7.0/3.0)
|   |   Purpose = Requirements in School: Non Complete (19.0/4.0)
|   |   Purpose = Hobby: Completer (18.0/6.0)
|   EducationalAttainment = Junior High School Ongoing
|   |   Purpose = For Employment Purposes: Completer (25.0/9.0)
|   |   Purpose = Requirements in School: Completer (29.0/12.0)
|   |   Purpose = Hobby: Non Complete (22.0/6.0)
|   EducationalAttainment = Elementary Undergraduate
|   |   TransportType = Private Transport: Completer (22.0/8.0)
|   |   TransportType = N/A: Completer (46.0/16.0)
|   |   TransportType = Public Transport
|   |   |   Purpose = For Employment Purposes: Completer (3.0/1.0)
|   |   |   Purpose = Requirements in School: Non Complete (8.0/1.0)
|   |   |   Purpose = Hobby: Completer (4.0/1.0)
|   EducationalAttainment = Elementary Ongoing
|   |   Proximity = 1 - 15 km: Completer (22.0/8.0)
|   |   Proximity = 31 + km: Non Complete (35.0/14.0)
|   |   Proximity = 16 - 30 km: Non Complete (12.0/4.0)
|   EducationalAttainment = Senior High School Graduate: Completer (69.0/27.0)
|   EducationalAttainment = College Ongoing
|   |   TransportType = Private Transport: Completer (9.0/3.0)
|   |   TransportType = N/A: Non Complete (59.0/25.0)
|   |   TransportType = Public Transport
|   |   |   Proximity = 1 - 15 km
|   |   |   |   Purpose = For Employment Purposes: Completer (3.0)
|   |   |   |   Purpose = Requirements in School: Non Complete (2.0)
|   |   |   |   Purpose = Hobby: Completer (2.0/1.0)
|   |   |   Proximity = 31 + km: Non Complete (0.0)
|   |   |   Proximity = 16 - 30 km
|   |   |   |   Purpose = For Employment Purposes: Non Complete (3.0)
|   |   |   |   Purpose = Requirements in School: Completer (3.0/1.0)
|   |   |   |   Purpose = Hobby: Non Complete (0.0)
|   EducationalAttainment = Senior HighSchool Undergraduate
|   |   Purpose = For Employment Purposes: Completer (27.0/10.0)
|   |   Purpose = Requirements in School: Completer (27.0/12.0)
|   |   Purpose = Hobby: Non Complete (34.0/12.0)
|   EducationalAttainment = College Graduate
|   |   Proximity = 1 - 15 km
|   |   |   Purpose = For Employment Purposes: Completer (9.0/1.0)
|   |   |   Purpose = Requirements in School: Non Complete (6.0/2.0)
|   |   |   Purpose = Hobby: Non Complete (6.0/1.0)
|   |   Proximity = 31 + km: Completer (50.0/19.0)
|   |   Proximity = 16 - 30 km: Non Complete (10.0/4.0)
```

Figure 34.2

J48 Model in 10 Fold Cross Validation

```
|   EducationalAttainment = Elementary Graduate: Non Complete (74.0/35.0)
|   EducationalAttainment = N/A: Non Complete (77.0/35.0)
|   EducationalAttainment = College Undergraduate: Completer (69.0/30.0)
|   EducationalAttainment = Junior High School Undergraduate
|   |   Proximity = 1 - 15 km
|   |   |   TransportType = Private Transport: Non Complete (8.0/2.0)
|   |   |   TransportType = N/A: Non Complete (0.0)
|   |   |   TransportType = Public Transport: Completer (9.0/3.0)
|   |   Proximity = 31 + km: Non Complete (40.0/19.0)
|   |   Proximity = 16 - 30 km: Completer (14.0/5.0)
Profeciency = Intermediate
|   EducationalAttainment = Junior High School Graduate
|   |   TransportType = Private Transport: Completer (18.0/8.0)
|   |   TransportType = N/A: Non Complete (38.0/15.0)
|   |   TransportType = Public Transport: Completer (11.0/3.0)
|   EducationalAttainment = Senior High School Ongoing: Non Complete (74.0/26.0)
|   EducationalAttainment = Junior High School Ongoing
|   |   Purpose = For Employment Purposes: Non Complete (29.0/12.0)
|   |   Purpose = Requirements in School: Completer (29.0/12.0)
|   |   Purpose = Hobby: Completer (22.0/11.0)
|   EducationalAttainment = Elementary Undergraduate
|   |   Purpose = For Employment Purposes: Non Complete (30.0/14.0)
|   |   Purpose = Requirements in School: Non Complete (31.0/12.0)
|   |   Purpose = Hobby: Completer (25.0/8.0)
|   EducationalAttainment = Elementary Ongoing: Non Complete (79.0/37.0)
|   EducationalAttainment = Senior High School Graduate
|   |   TransportType = Private Transport: Completer (16.0/7.0)
|   |   TransportType = N/A
|   |   |   Purpose = For Employment Purposes: Completer (18.0/8.0)
|   |   |   Purpose = Requirements in School: Non Complete (15.0/6.0)
|   |   |   Purpose = Hobby: Completer (20.0/8.0)
|   |   TransportType = Public Transport: Non Complete (19.0/7.0)
|   EducationalAttainment = College Ongoing
|   |   Proximity = 1 - 15 km: Completer (22.0/10.0)
|   |   Proximity = 31 + km: Non Complete (48.0/20.0)
|   |   Proximity = 16 - 30 km: Completer (12.0/3.0)
|   EducationalAttainment = Senior HighSchool Undergraduate
|   |   Purpose = For Employment Purposes: Non Complete (32.0/14.0)
|   |   Purpose = Requirements in School: Non Complete (25.0/8.0)
|   |   Purpose = Hobby: Completer (19.0/6.0)
|   EducationalAttainment = College Graduate
|   |   Purpose = For Employment Purposes: Non Complete (27.0/10.0)
|   |   Purpose = Requirements in School
|   |   |   Proximity = 1 - 15 km: Non Complete (6.0/1.0)
|   |   |   Proximity = 31 + km: Completer (17.0/7.0)
|   |   |   Proximity = 16 - 30 km: Completer (5.0)
|   |   Purpose = Hobby: Completer (17.0/6.0)
|   EducationalAttainment = Elementary Graduate
|   |   TransportType = Private Transport: Completer (14.0/5.0)
|   |   TransportType = N/A: Non Complete (51.0/20.0)
|   |   TransportType = Public Transport: Non Complete (12.0/3.0)
```

This was created using 10-Fold Cross Validation and trained on a dataset containing data about language learners' educational attainment, proficiency level, proximity to the school, transport type, and purpose of learning a language.

Depending on the language learners' level of skill, the tree is divided into two primary branches. For learners who are classified as beginners, the first branch on the left, and for learners who are classified as intermediate or advanced, the second branch on the right.

The first parameter used by the tree to further categorize novices is their level of education. The learners who have completed junior high school are categorized as non-completers, and those who have completed elementary school or are currently enrolled in senior high school are further split based on their motivation for learning a language and their way of transportation. Learners who don't have a means of transportation or who utilize private transportation are categorized as completers, while those who do so are categorized as non-completers. Language learners who study a language to fulfill academic or recreational needs are also categorized as non-completers, as opposed to language learners who study a language to advance their careers.

The tree also uses the learner's educational standing as the initial factor for intermediate or advanced learners. Non-completers are learners who have completed junior high school, graduated from elementary school, or are currently enrolled in college. Senior high school and college students are further divided according to their motivation for learning a language, proximity to the school, and mode of transportation. Learners who don't have a means of transportation or who utilize private transportation are categorized as completers, while those who do so are categorized as non-completers. Language learners who study a language to meet academic requirements are also categorized as non-completers, as opposed to language learners who study a language for joy, who are categorized as completers. College graduates are split according to how close they reside to the institution; those who live within 15 km are considered completers, while those who live farther away are considered non-completers. Those who use public transportation and live within 16 to 30 kilometers of the school are classified as completers, while those who live further away or use private transportation are classified as non-completers. This is applicable to students in junior high school.

Figure 35

J48 Validation Result in 10-Fold Cross Validation

```
=== Summary ===

Correctly Classified Instances        1998               66.6    %
Incorrectly Classified Instances      1002               33.4    %
Kappa statistic                          0.2542
Mean absolute error                      0.3345
Root mean squared error                  0.4281
Relative absolute error                 75.4502 %
Root relative squared error             90.9223 %
Total Number of Instances             3000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.740    0.483    0.755      0.740    0.748      0.254    0.744     0.879     Completer
                 0.517    0.260    0.497      0.517    0.506      0.254    0.744     0.493     Non Complete
Weighted Avg.    0.666    0.409    0.669      0.666    0.668      0.254    0.744     0.751
```

The model classified 66.6% of 3000 cases correctly and 33.4% incorrectly using cross-validation. Kappa is 0.2542. Root mean squared error is 0.4281, mean absolute error 0.3345. The model was assessed for "Completer" and "Non-Completer" accuracy.

The model accurately classified 74% of "Completer" cases. The false positive rate (FP Rate) is 0.483, meaning 48.3% of non-Completer cases were wrongly labeled as "Completer". The model's "Completer" class prediction accuracy is 75.5%. "Completer" has 0.74 recall, or sensitivity. F-measure, a weighted average of precision and recall, is 0.748. Matthew’s correlation coefficient (MCC) is 0.254. The model can discriminate between classes because its ROC Area is 0.744. Precision-recall curve area is 0.879.

The model successfully identified 51.7% of "non-Completer" occurrences, with a TP Rate of 0.517.

The FP Rate is 0.26, meaning 26% of instances in this class were misclassified as "Completer". The "NonCompleter" class has a precision of 0.497, meaning the model predicts "Non-Completer" instances 49.7% of the time. "Non-Completer" has 0.517 recall, or sensitivity. F-measure: 0.506. 0.254 MCC. ROC Area is 0.744, PRC Area 0.493.

The weighted average of TP Rate, FP Rate, precision, recall, F-measure, MCC, ROC Area, and PRC Area is also supplied. Weighted average TP Rate: 0.666, FP Rate: 0.409, Precision: 0.669, Recall: 0.666, Fmeasure: 0.668, MCC: 0.254. Weighted average ROC Area is 0.744 and PRC Area is 0.751.

Figure 36

J48 Confusion Matrix in 10-Fold Cross Validation

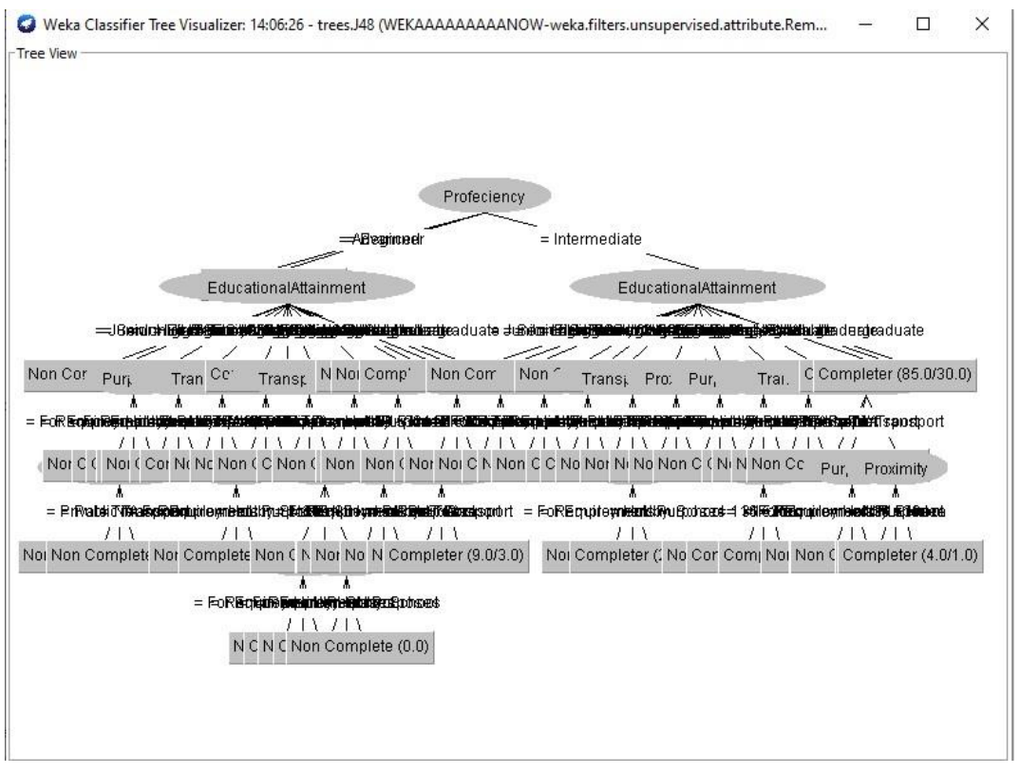
```
=== Confusion Matrix ===

    a    b   <-- classified as
 1484  521 |    a = Completer
  481  514 |    b = Non Complete
```

In contrast to the 481 false negatives (misclassified as Non-Completer) and the 521 false positives (misclassified as Completer), there were 1484 true positives (properly classified as Completer) and 514 true negatives (correctly classified as Non Completer).

Figure 37

J48 Model in 10 % Split



Figure

J48 Model in 10 % Split

37.1

Enrollment Process .

Figure

J48 Model in 10 % Split

```
Profeciency = Advanced: Completer (1004.0)
Profeciency = Beginner
|   EducationalAttainment = Junior High School Graduate: Non Complete (79.0/34.0)
|   EducationalAttainment = Senior High School Ongoing
|   |   Purpose = For Employment Purposes
|   |   |   TransportType = Private Transport: Non Complete (5.0/1.0)
|   |   |   TransportType = N/A: Completer (9.0/3.0)
|   |   |   TransportType = Public Transport: Non Complete (7.0/3.0)
|   |   Purpose = Requirements in School: Non Complete (19.0/4.0)
|   |   Purpose = Hobby: Completer (18.0/6.0)
|   EducationalAttainment = Junior High School Ongoing
|   |   Purpose = For Employment Purposes: Completer (25.0/9.0)
|   |   Purpose = Requirements in School: Completer (29.0/12.0)
|   |   Purpose = Hobby: Non Complete (22.0/6.0)
|   EducationalAttainment = Elementary Undergraduate
|   |   TransportType = Private Transport: Completer (22.0/8.0)
|   |   TransportType = N/A: Completer (46.0/16.0)
|   |   TransportType = Public Transport
|   |   |   Purpose = For Employment Purposes: Completer (3.0/1.0)
|   |   |   Purpose = Requirements in School: Non Complete (8.0/1.0)
|   |   |   Purpose = Hobby: Completer (4.0/1.0)
|   EducationalAttainment = Elementary Ongoing
|   |   Proximity = 1 - 15 km: Completer (22.0/8.0)
|   |   Proximity = 31 + km: Non Complete (35.0/14.0)
|   |   Proximity = 16 - 30 km: Non Complete (12.0/4.0)
|   EducationalAttainment = Senior High School Graduate: Completer (69.0/27.0)
|   EducationalAttainment = College Ongoing
|   |   TransportType = Private Transport: Completer (9.0/3.0)
|   |   TransportType = N/A: Non Complete (59.0/25.0)
|   |   TransportType = Public Transport
|   |   |   Proximity = 1 - 15 km
|   |   |   |   Purpose = For Employment Purposes: Completer (3.0)
|   |   |   |   Purpose = Requirements in School: Non Complete (2.0)
|   |   |   |   Purpose = Hobby: Completer (2.0/1.0)
|   |   |   Proximity = 31 + km: Non Complete (0.0)
|   |   |   Proximity = 16 - 30 km
|   |   |   |   Purpose = For Employment Purposes: Non Complete (3.0)
|   |   |   |   Purpose = Requirements in School: Completer (3.0/1.0)
|   |   |   |   Purpose = Hobby: Non Complete (0.0)
|   EducationalAttainment = Senior HighSchool Undergraduate
|   |   Purpose = For Employment Purposes: Completer (27.0/10.0)
|   |   Purpose = Requirements in School: Completer (27.0/12.0)
|   |   Purpose = Hobby: Non Complete (34.0/12.0)
|   EducationalAttainment = College Graduate
|   |   Proximity = 1 - 15 km
|   |   |   Purpose = For Employment Purposes: Completer (9.0/1.0)
|   |   |   Purpose = Requirements in School: Non Complete (6.0/2.0)
|   |   |   Purpose = Hobby: Non Complete (6.0/1.0)
|   |   Proximity = 31 + km: Completer (50.0/19.0)
|   |   Proximity = 16 - 30 km: Non Complete (10.0/4.0)
```

Enrollment Process .

Figure

J48 Model in 10 % Split

37.2

Enrollment Process .

Figure

J48 Model in 10 % Split

```
|   EducationalAttainment = Elementary Graduate: Non Complete (74.0/35.0)
|   EducationalAttainment = N/A: Non Complete (77.0/35.0)
|   EducationalAttainment = College Undergraduate: Completer (69.0/30.0)
|   EducationalAttainment = Junior High School Undergraduate
|   |   Proximity = 1 - 15 km
|   |   |   TransportType = Private Transport: Non Complete (8.0/2.0)
|   |   |   TransportType = N/A: Non Complete (0.0)
|   |   |   TransportType = Public Transport: Completer (9.0/3.0)
|   |   Proximity = 31 + km: Non Complete (40.0/19.0)
|   |   Proximity = 16 - 30 km: Completer (14.0/5.0)
Profeciency = Intermediate
|   EducationalAttainment = Junior High School Graduate
|   |   TransportType = Private Transport: Completer (18.0/8.0)
|   |   TransportType = N/A: Non Complete (38.0/15.0)
|   |   TransportType = Public Transport: Completer (11.0/3.0)
|   EducationalAttainment = Senior High School Ongoing: Non Complete (74.0/26.0)
|   EducationalAttainment = Junior High School Ongoing
|   |   Purpose = For Employment Purposes: Non Complete (29.0/12.0)
|   |   Purpose = Requirements in School: Completer (29.0/12.0)
|   |   Purpose = Hobby: Completer (22.0/11.0)
|   EducationalAttainment = Elementary Undergraduate
|   |   Purpose = For Employment Purposes: Non Complete (30.0/14.0)
|   |   Purpose = Requirements in School: Non Complete (31.0/12.0)
|   |   Purpose = Hobby: Completer (25.0/8.0)
|   EducationalAttainment = Elementary Ongoing: Non Complete (79.0/37.0)
|   EducationalAttainment = Senior High School Graduate
|   |   TransportType = Private Transport: Completer (16.0/7.0)
|   |   TransportType = N/A
|   |   |   Purpose = For Employment Purposes: Completer (18.0/8.0)
|   |   |   Purpose = Requirements in School: Non Complete (15.0/6.0)
|   |   |   Purpose = Hobby: Completer (20.0/8.0)
|   |   TransportType = Public Transport: Non Complete (19.0/7.0)
|   EducationalAttainment = College Ongoing
|   |   Proximity = 1 - 15 km: Completer (22.0/10.0)
|   |   Proximity = 31 + km: Non Complete (48.0/20.0)
|   |   Proximity = 16 - 30 km: Completer (12.0/3.0)
|   EducationalAttainment = Senior HighSchool Undergraduate
|   |   Purpose = For Employment Purposes: Non Complete (32.0/14.0)
|   |   Purpose = Requirements in School: Non Complete (25.0/8.0)
|   |   Purpose = Hobby: Completer (19.0/6.0)
|   EducationalAttainment = College Graduate
|   |   Purpose = For Employment Purposes: Non Complete (27.0/10.0)
|   |   Purpose = Requirements in School
|   |   |   Proximity = 1 - 15 km: Non Complete (6.0/1.0)
|   |   |   Proximity = 31 + km: Completer (17.0/7.0)
|   |   |   Proximity = 16 - 30 km: Completer (5.0)
|   |   Purpose = Hobby: Completer (17.0/6.0)
|   EducationalAttainment = Elementary Graduate
|   |   TransportType = Private Transport: Completer (14.0/5.0)
|   |   TransportType = N/A: Non Complete (51.0/20.0)
|   |   TransportType = Public Transport: Non Complete (12.0/3.0)
```

Figure

J48 Model in 10 % Split

37.3

```
|   EducationalAttainment = N/A
|   |   TransportType = Private Transport: Non Complete (12.0/5.0)
|   |   TransportType = N/A
|   |   |   Purpose = For Employment Purposes: Completer (14.0/6.0)
|   |   |   Purpose = Requirements in School: Non Complete (15.0/6.0)
|   |   |   Purpose = Hobby: Completer (16.0/6.0)
|   |   TransportType = Public Transport
|   |   |   Proximity = 1 - 15 km: Non Complete (12.0/5.0)
|   |   |   Proximity = 31 + km: Completer (0.0)
|   |   |   Proximity = 16 - 30 km: Completer (4.0/1.0)
|   EducationalAttainment = College Undergraduate: Completer (82.0/36.0)
|   EducationalAttainment = Junior High School Undergraduate: Completer (85.0/30.0)

Number of Leaves  :     83

Size of the tree :      114
```

On a dataset with a 10% percentage split, this model was trained. The data are initially divided by proficiency level by the model. It then further divides each proficiency level depending on the additional factors. For instance, if the proficiency level is "Beginner," it divides according to amount of schooling. A further division is made based on the reason for learning English if the educational attainment is "Junior High School Graduate," and so forth.

The numbers in the tree show how many instances there are for every possible pairing of attribute values and skill level. For instance, "Non-Completer (79.0/34.0)" denotes 79 cases where the educational attainment is "Junior High School Graduate" and the proficiency level is "Beginner," and 34 of those instances have not finished their program.

Figure 38

J48 Validation Result in 10 % Split

```
=== Summary ===

Correctly Classified Instances        1813               67.1481 %
Incorrectly Classified Instances       887               32.8519 %
Kappa statistic                          0.2628
Mean absolute error                      0.3309
Root mean squared error                  0.4079
Relative absolute error                 72.9966 %
Root relative squared error             86.6098 %
Total Number of Instances             2700

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.745    0.480    0.761      0.745    0.753      0.263    0.754     0.859     Completer
                 0.520    0.255    0.499      0.520    0.510      0.263    0.754     0.498     Non Complete
Weighted Avg.    0.671    0.406    0.675      0.671    0.673      0.263    0.754     0.740
```

Enrollment Process .

Figure

J48 Model in 10 % Split

In the test split, the model correctly identified 67% of occurrences, but misclassified 33%. The kappa statistic, which evaluates the degree of agreement between actual and anticipated classes, is 0.26, suggesting fair agreement. The root mean squared error and mean absolute error both point to moderate to high errors in prediction.

39

J48 Confusion Matrix in 10 % Split

```
=== Confusion Matrix ===

    a    b   <-- classified as
 1352  462 |    a = Completer
  425  461 |    b = Non Complete
```

The model's predictions are contrasted with the actual outcomes in the confusion matrix. The projected classes are shown by the columns, whereas the actual classes are represented by the rows. The model successfully identified 1352 completer instances, but incorrectly identified 462 noncompleter instances. It also accurately identified 461 noncompleter instances, but incorrectly identified 425 completer ones.

Table 4

Summary of Algorithm's Results

	Accuracy	Kappa Value	Mean Absolute error	TP Rate	FP Rate	Precision	Recall	F-measure	Time to build
Random Forest - Use Training Set	76.8%	0.4726	0.2787	0.837	0.370	0.820	0.837	0.829	0.15
Random Forest - Supplied Test Set	76.3 %	0.4714	0.2835	0.823	0.353	0.819	0.823	0.821	0.03
Random Forest - 10 Fold Cross-validation	66.5 %	0.2442	0.3358	0.750	0.506	0.749	0.750	0.749	0.31
Random Forest - 90% Percentage Split	66.7%	0.2447	0.3426	0.752	0.508	0.752	0.752	0.752	0.1
NaiveBayes- Use Training Set	69.9 %	0.3294	0.3296	0.763	0.429	0.782	0.763	0.772	0
NaiveBayes- Supplied Test Set	70.1%	0.3529	0.3355	0.745	0.382	0.791	0.745	0.767	0
NaiveBayes- 10 Fold Cross-validation	68.2%	0.288	0.332	0.756	0.466	0.766	0.756	0.761	0
NaiveBayes- 90% Percentage Split	67.0%	0.2695	0.3332	0.734	0.458	0.766	0.734	0.750	0
J48 - Use Training Set	74 %	0.4232	0.3101	0.789	0.359	0.816	0.789	0.802	0.01
J48 - Supplied Test Set	74.8%	0.4567	0.3182	0.775	0.304	0.832	0.775	0.803	0.01

J48 - 10 Fold Cross-validation	66.6 %	0.2542	0.3345	0.740	0.483	0.755	0.740	0.748	0.01
J48 - 90% Percentage Split	67.1%	0.2628	0.3309	0.745	0.480	0.761	0.745	0.753	0.02

Because of its interpretability and its ability to provide light on the connections between input attributes and output variable, we think the J48 decision tree algorithm is an excellent fit for this classification problem.

After looking at how well each model performed, we concluded that Random Forest had the best accuracy, kappa, and f measures. Still, the closest value was obtained by J48. With an accuracy of 74.8%, J48 ranks just below the Random Forest's 76.8%. While the other two models have high accuracy and kappa values, the Naive Bayes model has the lowest of the three.

The time needed to construct the random forest model was significantly longer than that of the J48. Random Forest took 0.15 seconds more than J48 did on the training set. We conclude that this is because models that are easy to construct are likely to be simplistic and so a better fit for the data. Since J48's build time is the shortest of all the models, it can rapidly produce the decision tree model with high accuracy and low error rate.

And since it yields a model that is easy to understand, the J48 decision tree algorithm is a frequently adopted machine learning technique for classification tasks. It's a tree-based algorithm that divides the data recursively along the most important properties, making for an understandable tree structure. The resulting decision tree is clear and straightforward, and the model is fast to train even on huge datasets and resource-efficient. In addition to being resistant to the effects of noisy data, decision trees are also insensitive to outliers and missing values

Figure 40

Chosen ruleset to implement

```
Profeciency = Advanced: Completer (1004.0)
Profeciency = Beginner
|   EducationalAttainment = Junior High School Graduate: Non Complete (79.0/34.0)
|   EducationalAttainment = Senior High School Ongoing
|   |   Purpose = For Employment Purposes
|   |   |   TransportType = Private Transport: Non Complete (5.0/1.0)
|   |   |   TransportType = N/A: Completer (9.0/3.0)
|   |   |   TransportType = Public Transport: Non Complete (7.0/3.0)
|   |   Purpose = Requirements in School: Non Complete (19.0/4.0)
|   |   Purpose = Hobby: Completer (18.0/6.0)
|   EducationalAttainment = Junior High School Ongoing
|   |   Purpose = For Employment Purposes: Completer (25.0/9.0)
|   |   Purpose = Requirements in School: Completer (29.0/12.0)
|   |   Purpose = Hobby: Non Complete (22.0/6.0)
|   EducationalAttainment = Elementary Undergraduate
|   |   TransportType = Private Transport: Completer (22.0/8.0)
|   |   TransportType = N/A: Completer (46.0/16.0)
|   |   TransportType = Public Transport
|   |   |   Purpose = For Employment Purposes: Completer (3.0/1.0)
|   |   |   Purpose = Requirements in School: Non Complete (8.0/1.0)
|   |   |   Purpose = Hobby: Completer (4.0/1.0)
|   EducationalAttainment = Elementary Ongoing
|   |   Proximity = 1 - 15 km: Completer (22.0/8.0)
|   |   Proximity = 31 + km: Non Complete (35.0/14.0)
|   |   Proximity = 16 - 30 km: Non Complete (12.0/4.0)
|   EducationalAttainment = Senior High School Graduate: Completer (69.0/27.0)
|   EducationalAttainment = College Ongoing
|   |   TransportType = Private Transport: Completer (9.0/3.0)
|   |   TransportType = N/A: Non Complete (59.0/25.0)
|   |   TransportType = Public Transport
|   |   |   Proximity = 1 - 15 km
|   |   |   |   Purpose = For Employment Purposes: Completer (3.0)
|   |   |   |   Purpose = Requirements in School: Non Complete (2.0)
|   |   |   |   Purpose = Hobby: Completer (2.0/1.0)
|   |   |   Proximity = 31 + km: Non Complete (0.0)
|   |   |   Proximity = 16 - 30 km
|   |   |   |   Purpose = For Employment Purposes: Non Complete (3.0)
|   |   |   |   Purpose = Requirements in School: Completer (3.0/1.0)
|   |   |   |   Purpose = Hobby: Non Complete (0.0)
|   EducationalAttainment = Senior HighSchool Undergraduate
|   |   Purpose = For Employment Purposes: Completer (27.0/10.0)
|   |   Purpose = Requirements in School: Completer (27.0/12.0)
|   |   Purpose = Hobby: Non Complete (34.0/12.0)
|   EducationalAttainment = College Graduate
```

Figure 40.1

Chosen ruleset to implement

```
|   |   Proximity = 1 - 15 km
|   |   |   Purpose = For Employment Purposes: Completer (9.0/1.0)
|   |   |   Purpose = Requirements in School: Non Complete (6.0/2.0)
|   |   |   Purpose = Hobby: Non Complete (6.0/1.0)
|   |   Proximity = 31 + km: Completer (50.0/19.0)
|   |   Proximity = 16 - 30 km: Non Complete (10.0/4.0)
|   EducationalAttainment = Elementary Graduate: Non Complete (74.0/35.0)
|   EducationalAttainment = N/A: Non Complete (77.0/35.0)
|   EducationalAttainment = College Undergraduate: Completer (69.0/30.0)
|   EducationalAttainment = Junior High School Undergraduate
|   |   Proximity = 1 - 15 km
|   |   |   TransportType = Private Transport: Non Complete (8.0/2.0)
|   |   |   TransportType = N/A: Non Complete (0.0)
|   |   |   TransportType = Public Transport: Completer (9.0/3.0)
|   |   Proximity = 31 + km: Non Complete (40.0/19.0)
|   |   Proximity = 16 - 30 km: Completer (14.0/5.0)
Profeciency = Intermediate
|   EducationalAttainment = Junior High School Graduate
|   |   TransportType = Private Transport: Completer (18.0/8.0)
|   |   TransportType = N/A: Non Complete (38.0/15.0)
|   |   TransportType = Public Transport: Completer (11.0/3.0)
|   EducationalAttainment = Senior High School Ongoing: Non Complete (74.0/26.0)
|   EducationalAttainment = Junior High School Ongoing
|   |   Purpose = For Employment Purposes: Non Complete (29.0/12.0)
|   |   Purpose = Requirements in School: Completer (29.0/12.0)
|   |   Purpose = Hobby: Completer (22.0/11.0)
|   EducationalAttainment = Elementary Undergraduate
|   |   Purpose = For Employment Purposes: Non Complete (30.0/14.0)
|   |   Purpose = Requirements in School: Non Complete (31.0/12.0)
|   |   Purpose = Hobby: Completer (25.0/8.0)
|   EducationalAttainment = Elementary Ongoing: Non Complete (79.0/37.0)
|   EducationalAttainment = Senior High School Graduate
|   |   TransportType = Private Transport: Completer (16.0/7.0)
|   |   TransportType = N/A
|   |   |   Purpose = For Employment Purposes: Completer (18.0/8.0)
|   |   |   Purpose = Requirements in School: Non Complete (15.0/6.0)
|   |   |   Purpose = Hobby: Completer (20.0/8.0)
|   |   TransportType = Public Transport: Non Complete (19.0/7.0)
|   EducationalAttainment = College Ongoing
|   |   Proximity = 1 - 15 km: Completer (22.0/10.0)
|   |   Proximity = 31 + km: Non Complete (48.0/20.0)
|   |   Proximity = 16 - 30 km: Completer (12.0/3.0)
|   EducationalAttainment = Senior HighSchool Undergraduate
|   |   Purpose = For Employment Purposes: Non Complete (32.0/14.0)
|   |   Purpose = Requirements in School: Non Complete (25.0/8.0)
|   |   Purpose = Hobby: Completer (19.0/6.0)
|   EducationalAttainment = College Graduate
|   |   Purpose = For Employment Purposes: Non Complete (27.0/10.0)
|   |   Purpose = Requirements in School
|   |   |   Proximity = 1 - 15 km: Non Complete (6.0/1.0)
|   |   |   Proximity = 31 + km: Completer (17.0/7.0)
```

Figure 40.2

Chosen ruleset to implement

```
|   |   |   Proximity = 16 - 30 km: Completer (5.0)
|   |   Purpose = Hobby: Completer (17.0/6.0)
|   EducationalAttainment = Elementary Graduate
|   |   TransportType = Private Transport: Completer (14.0/5.0)
|   |   TransportType = N/A: Non Complete (51.0/20.0)
|   |   TransportType = Public Transport: Non Complete (12.0/3.0)
|   EducationalAttainment = N/A
|   |   TransportType = Private Transport: Non Complete (12.0/5.0)
|   |   TransportType = N/A
|   |   |   Purpose = For Employment Purposes: Completer (14.0/6.0)
|   |   |   Purpose = Requirements in School: Non Complete (15.0/6.0)
|   |   |   Purpose = Hobby: Completer (16.0/6.0)
|   |   TransportType = Public Transport
|   |   |   Proximity = 1 - 15 km: Non Complete (12.0/5.0)
|   |   |   Proximity = 31 + km: Completer (0.0)
|   |   |   Proximity = 16 - 30 km: Completer (4.0/1.0)
|   EducationalAttainment = College Undergraduate: Completer (82.0/36.0)
|   EducationalAttainment = Junior High School Undergraduate: Completer (85.0/30.0)
```

As far as we can tell, this decision tree classifier employs a set of guidelines to forecast whether a user or student would finish a certain study in light of their characteristics. The most crucial property (Proficiency) is checked first, then other attributes are checked based on how well the previous attribute performed. The leaf node that is reached by traveling down the tree and matching the user's attributes serves as the basis for the final forecast.

The individual's proficiency level is where the tree begins. The tree indicates that they will successfully perform their transportation assignment with 100% certainty if their competence level is "Advanced". The tree makes a prediction if the user's skill level is "Beginner" by looking at their other features. For instance, if the person is a "College Ongoing" student, it is anticipated that they would finish their study if they use "Private Transport" or if they are within "1-15 km" and utilizing "Public Transport" for "Employment Purposes" or "Hobby." If they are utilizing "N/A" or if they are within "16-30 km" and using "Public Transport" for "Requirements in School," it is expected that they won't finish their study. They are projected to fail to finish their study if they are within "31+ km" of each other.

The study is expected to be completed by the person if they are utilizing transportation for "Employment Purposes" or "Requirements in School" and they are a "Senior High School Undergraduate" student. It is expected that they won't finish their study if they are using transportation as a "Hobby".
If the person is a "College Graduate" student and within "1-15 km" and uses "Public Transport" for "Employment Purposes" or "Hobby," they are expected to finish it. If they are within "31+ km" distance or within "16-30 km" distance and using "Public Transport" for "Requirements in School," it is expected that they won't finish their study.

We eventually come to the conclusion that the model's prediction is roughly right but there are still room for improvement.

9. References

Abubakar, Y. S., & Ahmad, N. A. (2017). Prediction of students' performance in e-learning environment using random forest. International Journal of Innovative Computing, 7(2). https://doi.org/10.11113/ijic.v7n2.143

Astin, A. (1975). Preventing Students from Dropping Out. San Francisco: Jossey-Bass.

Fahd, K., Miah, S. J., & Ahmed, K. (2021). Predicting student performance in a blended learning environment using learning management system interaction data. Applied Computing and Informatics. https://doi.org/10.1108/aci-06-2021-0150

Hamoud, A. K., Humadi, A. M., Awadh, W. A., & Hashim, A. S. (2017). Students' Success Prediction based on Bayes Algorithms. International Journal of Computer Applications. https://doi.org/10.5120/ijca2017915506

Kuh, G., Gonyea, R., & Palmer, M. (n.d.). The Disengaged Commuter Student: Fact or Fiction? Indiana University Center for Postsecondary Research and Planning. http://nsse.iub.edu/pdf/commuter.pdf

Pascarella, E., & Terenzini, P. (1998). Studying College Students in the 21st Century: Meeting New Challenges. The Review of Higher Education, 21, 151-165. https://doi.org/10.1353/rhe.1998.0013

Peltier, G., Laden, R., & Matranga, M. (1999). Student Persistence in College: A Review of Research. Journal of College Student Retention, 1, 357-376.

Thompson, J., Samiratedu, V., & Rafter, J. (1993). The Effects of On-Campus Residence on First-Time College Students. NASPA Journal, 31, 41-47.

Tun, M. M., & Htay, Y. Y. (2020). Predict Students' Performance by Using J48 Algorithm. International Journal of Scientific Research in Science, Engineering and Technology, 578–582. https://doi.org/10.32628/ijsrset2073124

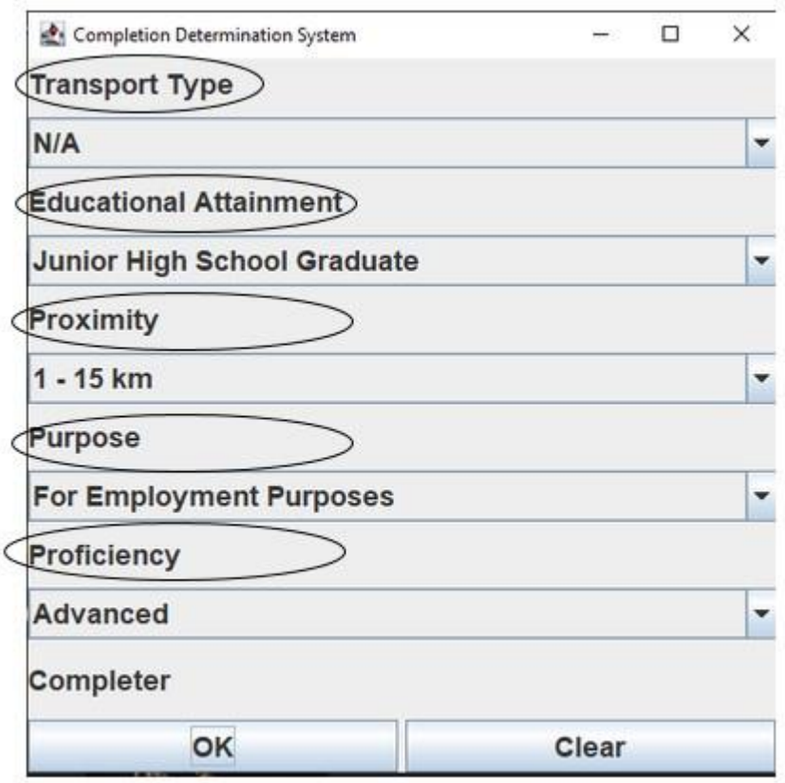
Y Divyabharathi et al. (2018). A Framework for Student Academic Performance Using Naive Bayes Classification Technique. Zenodo. https://doi.org/10.5281/zenodo.1277183

Yehuala, M. A. (2015). Application Of Data Mining Techniques For Student Success And Failure Prediction (The Case Of Debre_Markos University). International Journal of Scientific & Technology Research, 4(4), 91–94. https://www.ijstr.org/final-print/apr2015/Application-Of-Data-Mining-Techniques-For-Student-Success-And-Failure-Prediction-the-Case-Of-Debremarkos-University.p

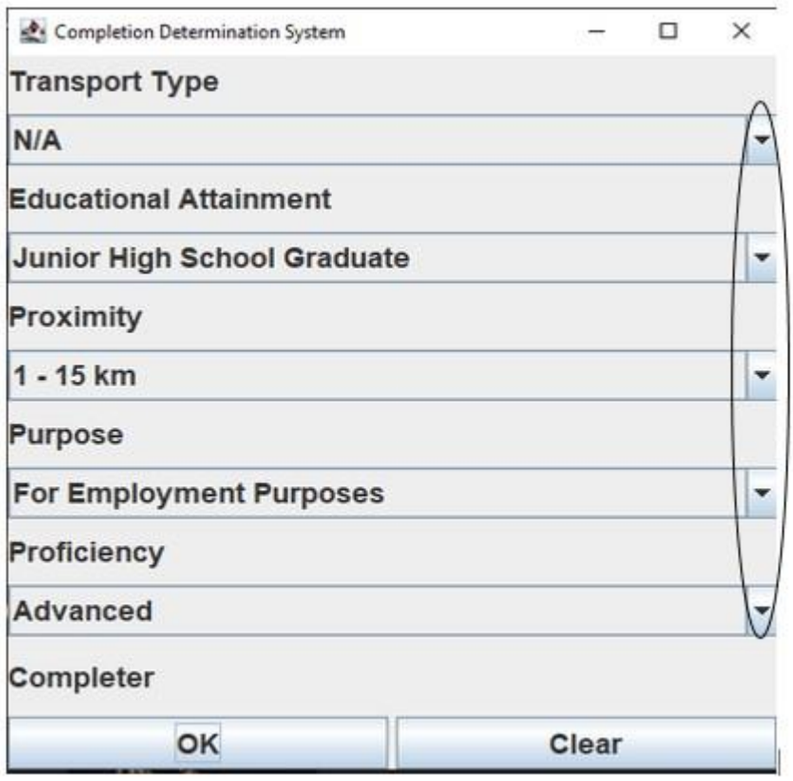
Part II. System Development

1. System

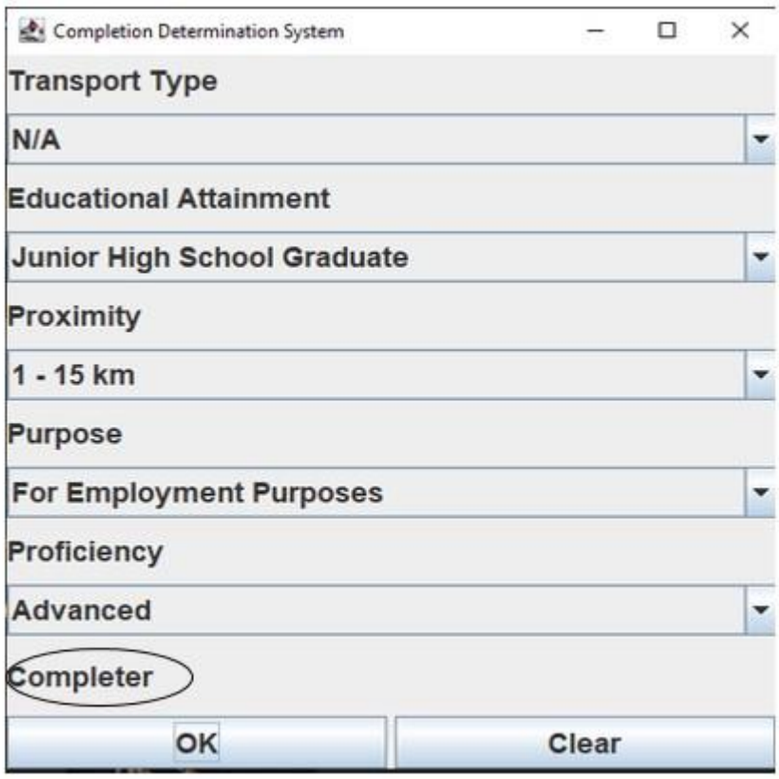
Here are the visuals for our system together with its significant features.

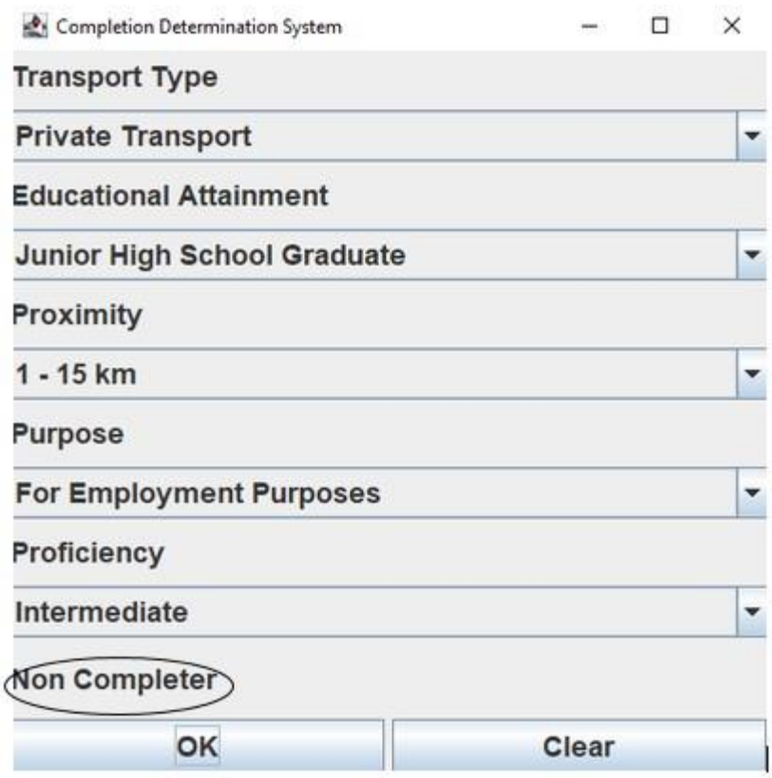


Predictors: The system has five main predictors, including educational attainment, transport type, proximity, purpose, and proficiency.



User-input fields: The system allows users to input their values for each predictor to generate the output.

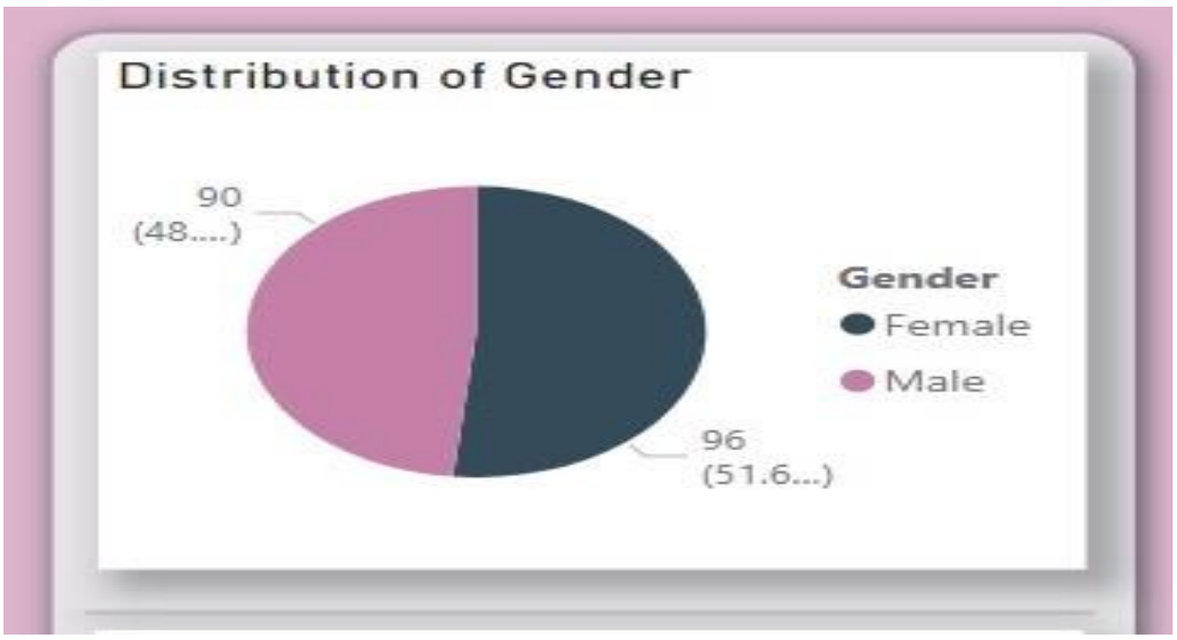




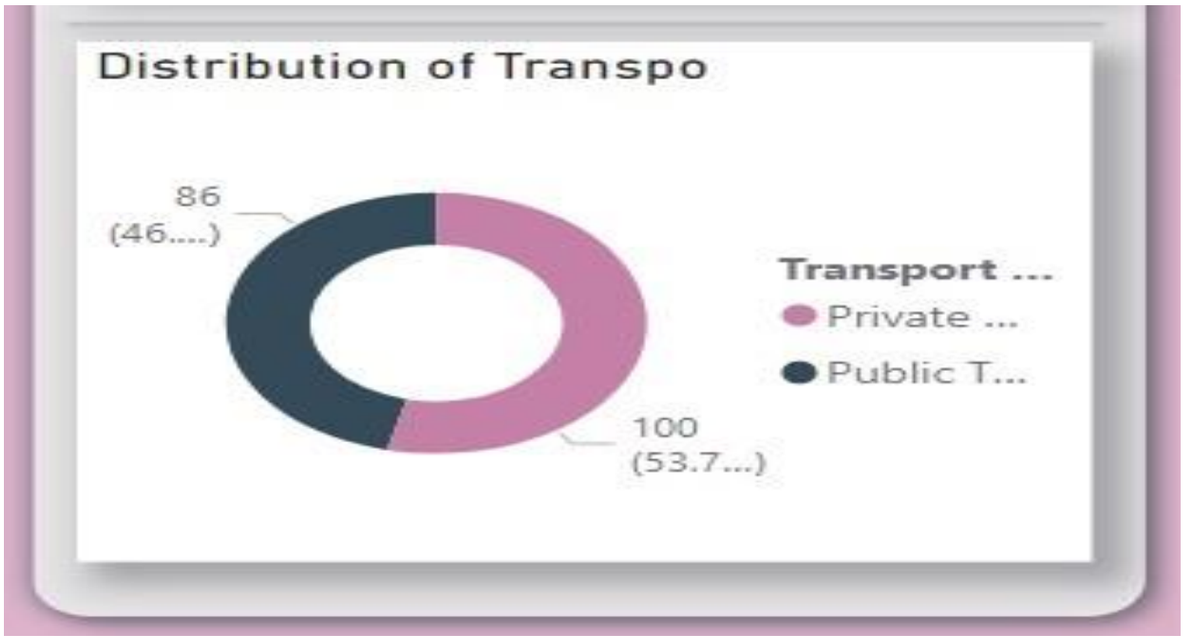
Output: The system generates an output of either "Completer" or "Non-completer" based on the user inputs.

2. Visual Analytics

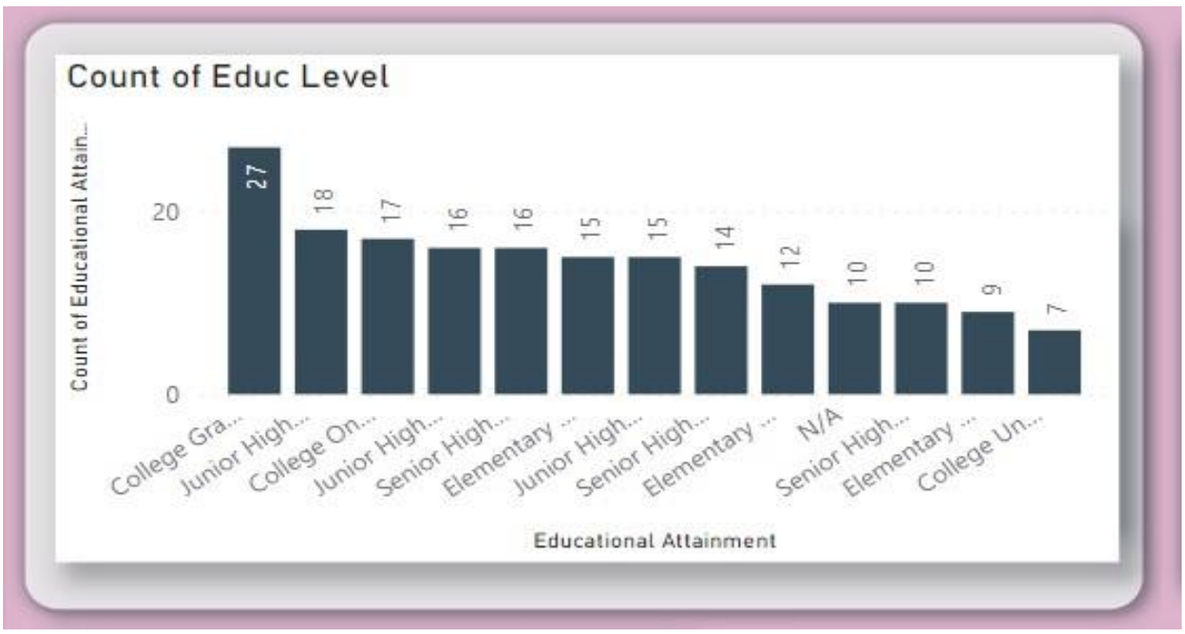
This is the full appearance of our visualization. In here, we have three sclicers and six charts/graphs.



In here, we used pie chart to showcase the distribution of gender within the areas of students. For instance, when the proximity 1-15 km is clicked and the slicer completer, those number of students who are female and male will be counted. It can be helpful to gain valuable insights into potential reasons that could be affecting completion rates by examining the gender distribution of completers and non-completers. For instance, if there are disproportionately more female non-completers than completers within a given proximity range, this may suggest that there are obstacles or difficulties that are unique to female students in that area and are impeding their ability to finish their studies. To address these issues and raise overall completion rates, this data can be utilized to guide targeted support and intervention services.



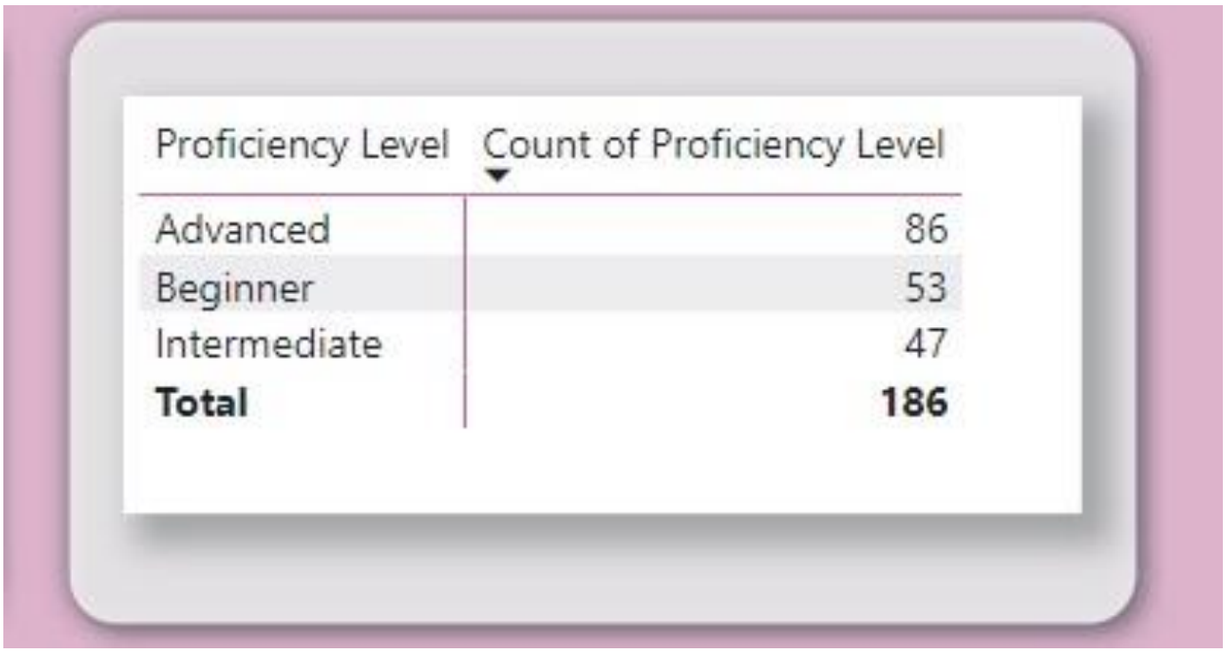
This is an essential insight since it offers details on the students' transportation choices, which may have an impact on their capacity to complete the program. Students who depend on public transit, for instance, can find it more difficult to attend in-person classes or access resources, which could affect their completion rate. The dashboard can spot any trends or patterns that could be handled to raise completion rates, such adding more online classes, by examining the distribution of mobility alternatives among completers and non-completers.



This is a significant finding because academic success is strongly predicted by educational attainment. We are able to spot any patterns or trends that might be connected to completion rates by examining the distribution of educational attainment between those who complete and those who do not. For instance, if we discover that many graduates have a college degree whereas many dropouts simply have a high school diploma, this may indicate that academic preparation and skills are important determinants of completion rates. This data can be used to create programs that better assist students in completing their studies and to establish targeted support and assistance services for students with a lower degree of educational attainment.

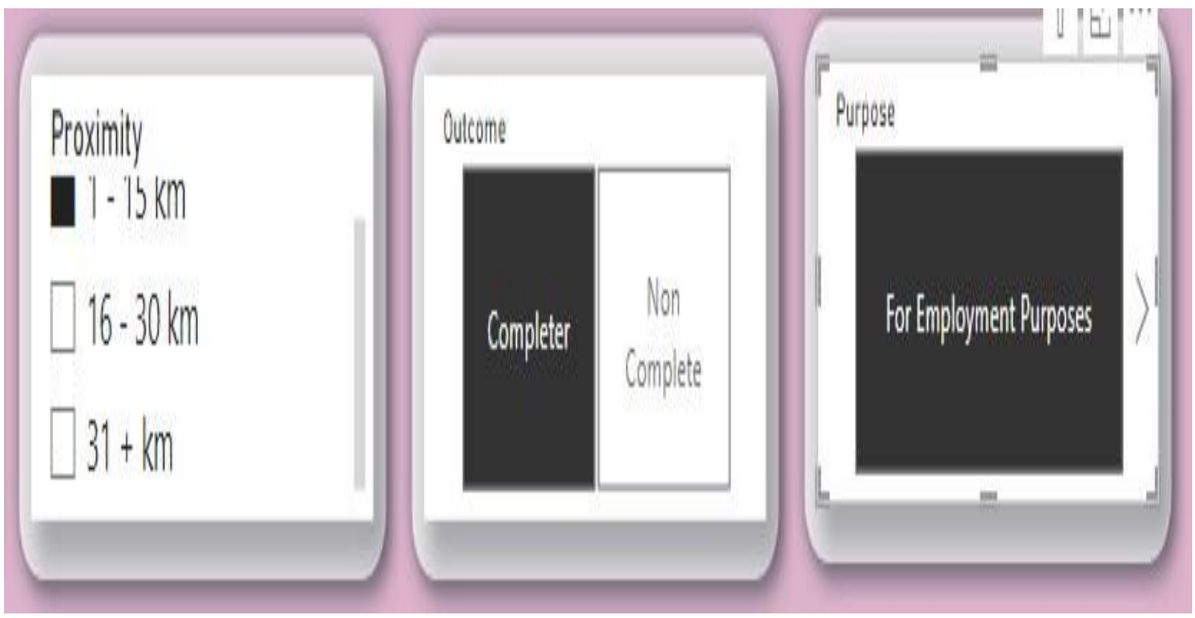
The total number of students is a crucial fact since it lets one know how big the population they're studying is. When making decisions and reaching conclusions based on the evidence, this is crucial. The total number of students in each category (completer and non-completer) also enables comparisons and the detection of any patterns or differences that might exist between the two groups.

This is an essential insight because it enables us to comprehend the work situation of the students who completed their studies or did not. It may indicate that there are underlying problems keeping students from finishing their studies, such as financial limitations or a lack of prospects, if we can discover that a sizable portion of non-completers are unemployed. On the other hand, if the majority of non-completers are employed, it may suggest that there are additional factors, such as a lack of interest or engagement with the program, that contributed to their failure to complete their studies. For the purpose of creating effective interventions and support systems to assist students in completing their education, it is crucial to comprehend the employment status of such individuals.



Proficiency Level	Count of Proficiency Level
Advanced	86
Beginner	53
Intermediate	47
Total	**186**

It's critical to understand learner's competency levels in order to identify their strengths as well as their weaknesses. With the help of this data, ABC Company may be customized to meet their unique needs, extra assistance or resources can be given.



Here are the slicer for our visualization. Proximity have three choices which are 1-15 km, 16-30 km, and 31+ km. Outcome are those students who have the potential to be completer and non-completer. Lastly is the purpose whether the students are studying for employment purpose, hobby, or for complying with their school requirements.