Predictive Analysis of Dropouts in Information Technology Higher Education

U. G. N. Kumari
169317F

Dissertation submitted to the Faculty of Information Technology, University of Moratuwa, Sri Lanka for the partial fulfillment of the requirements of Degree of Master of Science in Information Technology

April 2019

Declaration

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student Signature of Student

U.G.N.Kumari ……………………….

Date:

Supervised by

Name of Supervisor: Signature of Supervisor:

Mr. S.Premaratne ……………………….

Date:

Acknowledgment

I special gratitude pass to Mr. S. Premaratne, Senior Lecturer, Faculty of Information Technology, University of Moratuwa for his help, direction, and supervision was given to me all through my undertaking, making it a triumph.

Very special gratitude goes out to all down at Advanced Technological Institute, Labudawa for helping and providing the enormous support for data collection.

My genuine appreciation to Mrs. S.C.P. De Silva, Department of IT, Faculty of IT, [1]University Of Moratuwa for the animating talks, occupied hours we were cooperating before due dates.

I‘m grateful to all the staff members of the Faculty of Information Technology, University of Moratuwa.

To wrap things up I might want to thank my family for supporting me profoundly throughout writing this thesis and my life in general.

Abstract

At this point attention on educational data mining methods have impact highly on predicting academic performance as the increased higher education dropout rates especially in information technology education has received huge attention in recent years due to the quality of higher education has been a topic of debate for many years. There is a huge necessity of mining educational data system and exact hidden knowledge to understand the factors affecting student dropouts and to understand the patterns that can lead to predict student performance at the entrance and improve and monitor performance of students enrolling in Information technology higher education by building early warning indicators based on factors affecting to dropouts and manage students drop out from the higher education. Data mining strategies have been utilized to effectively extract new, conceivably imperative Knowledge and diverse data mining techniques, for example, association, classification, clustering, prediction, sequential patterns, and decision trees are being utilized by numerous sorts of research. Identification of relevant attributes which affect to dropouts in ICT higher education is a leading concern in the field of education data mining as there are no significant studies that can be applied to understand the complex-inter correlated and distinct factors affecting to dropouts in ICT higher education Hence, the research has been conducted to identify the complex-inter correlated and distinct factors affecting to dropouts in ICT higher education. It is hypothesized that an experimental methodology can be adapted to generate a database that includes relevant information for extracting knowledge. The raw data will be preprocessed in terms of filling up missing values, transforming values in one form into another and relevant attribute/ variable selection. Thereby select student records, which can be used for classification prediction model construction. In constructing a classifier model different classifying algorithms can be applied and in this study evaluation of different classification algorithms is done to identify the most accurate algorithm. Finally, a predictive analyzing model will be building for student profile analyzing using the identified algorithm. Then this classification model will be used in developing an application to predict the students' dropouts. The overall research will be designed using the WEKA data mining tool and using java WEKA library for developing the application.

Table of Contents

List of Figures

List of Tables

Chapter 1

1 Introduction

1.1 Chapter Introduction

This chapter serves as an introduction to the study by detailing the background and motivation which lead to the development of this study. Structure of the report is also detailed at the end of this chapter.

1.2 Increased higher education dropout rate is a topic of debate for many years!

Students’ academic performance is basic for educational institutes in light of the fact that key intervention programs can be arranged in improving or keeping up student academic performance during their period of studies in the institute[1]. As of ongoing years various countries are focusing on elements which impact the probability of an student to drop out of university or college and elements affecting the low performance of academic work. The expansion in students' dropout rate in advanced education is one of the imperative issues in many universities[2].

Data mining separates the first and significant knowledge from a huge databases. Data mining can be executed in various regions, for example, Fraud identification, Medical, Education, Banking, Marketing, and Telecommunications. Feature Selection is a procedure to pick the most important factors from the large dataset as a subset that are indistinguishably appropriate for examination and for future forecast by evacuating the irrelevant or repetitive factors. A definitive goal of the feature selection procedure is to build the prescient exactness and diminishing the intricacy of student results. In the colleges or in scholastic organizations, it's hard to anticipate the frailer or dropout student in an early stage. Data assimilations are the primary procedure used to diminish student dropout rate and to expand the student enrolment rates in an institute. Dropout in residential higher education institute is brought about by scholastic, family and individual reasons, grounds condition and foundation of the college and shifts relying upon the instructive structure concurred by the college[3].

Data mining procedures have been utilized to effectively remove new, conceivably critical learning by many researchers and diverse data mining strategies and techniques were contrasted with assessing their appropriateness for anticipating dropout factors in higher education. There are diverse strategies accessible in Education Data Mining: Linear Regression, Clustering, Classification, Association Rule Mining that can be utilized to extract new knowledge from information accumulated in huge educational databases as data mining investigate the connection between factors stored in the educational databases to comprehend all the students more likely to be dropping out, and setting which they learn in.

While predictive analysis has been used in various organizations for quite a while, higher education is a for the most part late adopter of these procedures as a gadget to settle on a choice. Starting late, pros in the region of machine learning and information mining tried to address the student retention possibility. A couple of classification algorithms including Bayes classifier, Decision tree, Boosting strategies and Support Vector Machines have been made to foresee students wearing down with higher exactness contrasted with the conventional measurable techniques. [4].

Identification of relevant attributes which affect to dropouts in ICT higher education is a leading concern in the field of education data mining as there are no significant studies that can be applied to understand the complex-inter correlated and distinct factors affecting to dropouts in ICT higher education. One size fits all model cannot be effectively applied to all students, and the solution is to segment students and perform different approach for each group. Performing comprehensive researches on bigger data set, several years in duration, with different students of ICT study programs can improve the findings of studies.

In spite of the fact that Data mining systems have been utilized to effectively extricate new, conceivably vital information by researchers and different data mining strategies and techniques were contrasted with assess their reasonableness for predicting dropout factors in higher education, a Dropout Prediction Model that incorporates knowledge extract from understanding relation among affecting variables is required for the effective prediction of students' dropout with high precision as it is helpful for identifying students with high dropout possibility at the entrance and providing necessary warnings and indications to decrease the higher rate dropouts. Utilization of classification techniques maps information into predefined gatherings of classes and classes can be resolved before analyzing the data. Prediction models can work on all close to home, social, mental and other ecological factors that require the compelling expectation of the dropouts of students.

There are many studies that focused on using classification data mining techniques to predict the dropouts in the higher education sector, finding the most accurate algorithm among different classifying algorithms for a particular dataset is vital. It allows researchers to predict the unknown values of variables of interest given known values of other variables accurately. Also, the prediction of students' success at the entrance point is crucial and need more attention on more distinctive attributes to get more accurate results that can be utilized as a foundation for the development of decision support system at the student selection process. Modern researches' results demonstrate that mediation programs can effectively affect dropout, particularly for the first year. To viably use the restricted help assets for the intercession programs, it is attractive to distinguish ahead of time students who will in require the help to continue their studies.

Therefore this study proposes a model for identifying students with high dropout possibility at the entrance and analyzing student profiles that better suit for Information Technology education by identifying most accurate classifier algorithm for the dataset to build the classifier model and implementing an application based on the constructed classifier model. The study will be supported by information extracted from the Higher National Diploma in Information Technology course of Sri Lanka Institute of Advanced Technological Education.

1.3 Problem Definition

Thus, this study attempts to answer the question "Can classification algorithms used in data mining predictively analyze the dropouts in Information Technology higher education?" In achieving above, three contributions are identified,

1. Can correlation-based feature selection generate the most suitable factors affecting to dropouts
2. What are the most relevant factors to predictively analyze student dropouts

3. What is the most accurate classification algorithm to construct the classification model
4. Can the power of prediction tools such as decision tree algorithm are able to predict the student dropouts in information technology higher education accurately

1.4 Aim and Objectives

1.4.1 Aim

The aim of this study is to successfully predict the dropouts in Information Technology education based on the classification model constructed by WEKA data mining tool and develop student dropout prediction application using Java WEKA API.

1.4.2 Objectives

In meeting the aim above, below objectives were identified.

1. Identify most affecting factors for dropouts in information technology courses of higher education
2. Understand the complex-inter correlated and distinct factors affecting to dropouts in ICT higher education.
3. Construct a predictive model to determine the dropout features of a student using Data Mining Techniques.
4. Develop an application to predict the dropout status at the entrance to the ICT courses based on identified factors affecting to the dropouts
5. Use the developed application for the selection of a suitable candidate at the entrance to the information technology courses.

1.5 Structure of the report

The content of this report is divided into eight chapters viz., Introduction, Review of others work, Technology Adapted, Approach, Analysis and Design, Implementation, Results and Evaluation, and Conclusion and Further work. Chapter 1 provides an introduction to the study whereas Chapter 2 reflects the findings from the earlier studies. Chapter 3 details the rationale behind choosing the technology used. Chapter 4 and 5 depicts the approach, analysis, and design of the study respectively. Chapter 6 briefs the high-level implementation details whereas Chapter 7 analyses the results of the study. Conclusions and further work related to this study are available in Chapter 8.

1. 6 Chapter Summary

As briefed in this chapter, Data mining Techniques are being used in predicting students' dropouts status and classification method is becoming increasingly popular and has provided promising results. However, there is still space for innovation as most of the earlier studies suffer from limitations in its applicability. Chapter 2 details the findings from the earlier studies further highlighting the research problem.

Chapter 2

2 The emerging use of Data Mining Techniques in Education Sector

2.1 Chapter Introduction

This chapter looks back at the past studies which have used different factors affecting to students' dropouts related to personal, social, psychological and other environmental backgrounds which necessitate the effective prediction of the dropouts of students. The chapter also provides a detailed description of the data mining techniques used to predictively analyze the affecting factors. The chapter provides a detailed description of the background information briefed in the previous chapter, related studies from other disciplines, approach, findings and limitations of similar research, which lead to the motivation of this study.

2.2 Educational Data Mining and Increased Dropout Rates in Higher Education

Educational Data Mining studies take data from different schools, colleges, private higher educational organizations or through web-based learning courses and distance training courses. Research in EDM applies different data mining procedures, for example, classification and prediction to discover hidden patterns, distinguish connections among factors and discovery methods to investigate knowledge of identity contrasts in getting the hang of teaching-learning behavior. Data mining has pulled in a parcel of consideration in the exploration field and in the public arena as a whole in recent years because of tremendous accessibility of expansive measure of information and the requirement for transforming such information into valuable information and knowledge. Information mining, likewise called Knowledge Discovery in Databases (KDD), is the field of finding new and possibly helpful data from gigantic databases[5]

Ongoing years have also observed real changes in the types of EDM strategies that are utilized, with prediction and discovery with models expanding while relationship mining winds up rarer. It will intrigue perceive how these patterns move in the years to come, and what new sorts of research will rise up out of the expansion in discovery with models, a strategy noticeable in cognitive modeling and bioinformatics, yet up to this point uncommon in education research. Now, instructive information mining techniques have had some dimension of effect on education and related interdisciplinary fields, (for example, computerized reasoning in instruction, insightful coaching frameworks, and user modeling).

Despite the fact that the use of data mining strategies in higher education is still generally new, its utilization is proposed for the understanding and extraction of new and possibly significant knowledge from the educational data. Since one of the greatest difficulties that higher education faces are to improve student dropout rate, student dropout is a greater concern about student academic performances and enrolment management in higher education. It is accounted for that around one-fourth of students dropped out from institutes after their first year, particularly in Information Technology instruction. Student dropout has turned into a sign of resource management and enrolment the management programs. Ongoing study results demonstrate that intervention programs can effectively affect dropout, particularly for the first year. To successfully use the limited support assets for the mediation programs, it is alluring to distinguish in advance, students who will likely to be supported most from the intervention programs conducted by institutes. Different studies have been thinking about various parameters, their affiliation and their importance in predicting dropouts in higher education and have discovered that the student accomplishment in higher education is affected by different variables which influence the choice of dropping out.

2.2.1 Different Data Mining Techniques and Increased Dropout Rates in Higher Education

So as to distillate the holes in existing forecast strategies, factors that characterize the academic performance and the prediction strategies that can be utilized to decide the student academic performance should be considered. Many studies uncover that Neural Network, Decision Tree, SVM, K.NN, Naïve Bayes with Neural Network have the most accuracy and Classification strategy is the most frequenting EDM by surveying systems used in past studies[6].

Four important classification methodologies—Decision Tree calculations, Support Vector Machines, Artificial Neural Networks, and discriminant investigation — can be used to construct classifier models. Their performance and quality can be evaluated by considering their application on a dataset by using accuracy, precision, recall, and specificity performance measurements Although all the classifier models show proportionally high classification accuracy, improved decision tree classifier was perceived as the best concerning, exactness, accuracy and specificity. further, an investigation of the variable significance for each classifier model should be conceivable to perceive the most crucial attributes[7].

A decision tree is a standout amongst the newest and famously utilized data mining strategies. In a decision tree, there are three particular sorts of center points: root nodes, decision nodes, and leaf nodes. A Decision tree starts with a solitary hub, called the root hub. A Decision hub has something like two branches. A leaf hub speaks to order or choice. The utilization of this structure is to part the dataset at a node reliant on the significance of the information. The measures, for instance, entropy, Gini record, ID3 or C4.5 utilized for making choice trees.[1].

The feature selection techniques help specialists to determine and increase an important bit of knowledge of the correlation between influencing variables and the target class variable of dropping out and remain in the program. A further troublesome request remains: would one have the capacity to anticipate whether a student will proceed or graduate using these noteworthy factors? Supervised learning algorithms, for instance, classification can be used to address this issue. The classification results are intended to give experts early cautions of a student who most likely won't suffer or graduate soon, with the objective that they can realize early intercessions in an advantageous way [8].

A significantly exact prediction model can be created using a supervised classification approach in the machine learning field, they achieved 89% precision in dropout forecast errand with slope boosting decision tree model and can construct a framework in drop out prediction system, including data extraction from Edx arrange, data preprocessing, feature engineering and performance test on a few supervised classification models [9].

So as to classify the dropout student, the most well-known data mining methodologies can be founded on k-Nearest Neighbor (k-NN), Decision Tree (DT), Naive Bayes (NB) and Neural Networks (NN). These four diverse order algorithms can be trained and tested utilizing 10-crease cross-approval system. In spite of the fact that there was not a major contrast in sensitivities among these four classifier calculations, the 3-NN (87%) and DT (79.7%) classifiers are progressively delicate. Additionally, these touchy scores are promising outcomes for foreseeing dropout student in the higher education program dataset[10].

2.2.2. Different Factors Affecting Dropouts

Various studies have been considering numerous parameters, their association and their relevancy in predicting dropouts in higher education and have found that the student achievement in higher education is affected by different variables which influence the choice of dropping out

Prediction of student dropout reasons can be a troublesome undertaking because of different variables that can influence the choice. In addition, the total strategy is long and tedious. Data collected from different bases and need information preprocessing to be done first to clean and alter information so as to apply order algorithms. In preprocessing step selection algorithms are used to perceive features that will impact the prediction procedure the most. It might be seen various factors like statistical parameters, money related components, family parameters, etc. can be responsible for a student from dropping out from a higher education course. [2].

Many studies meant to find concealed knowledge from the EDM to examine factors influencing the dropouts to design and oversee higher education programs. These studies were able to provide

the required understanding to keep up the improved dropout rate from higher education courses [7].
It tends to be seen numerous components like statistic factors, financial elements, family factors can be in charge of understudies from dropping out from students from higher education courses. A study done for an investigation of factors affects to dropouts reveals that there were geographic contrasts among students who dropped out of school, remained selected, or effectively graduated inside six years of registration. Geographic separation did not have any kind of effect for dropouts or graduates. In any case, access to instructive offices is essential for them to be fruitful [8]

The significance of extracurricular exercises to foresee dropout appearing extracurricular exercises are amazing dropout indicators. Knowledge extracted from Educational data mining study recognized that components influence the better achievement of the students at higher education. Students' Grade Point Average was recognized as the key important factor to be considered, to offer proposals to an educational organization as a preventive activity of low performance of students and to structure an effective model to accomplish great academic achievement. The study was conveyed in Information Technology office and found that the components that affected the accomplishment of student's academic accomplishment are the average value of High School Diploma on National Exam (NEM), High School specialization, results of International Conference on Knowledge, Information and Creativity Support Systems (KICSS), placement test (Academic Potential Test), student attendance, and the teacher quality. The consequences of this investigation bolster a few components like inspiration, earned credit focuses, earlier examinations, desires that were observed in writing to be essential for keeping away from dropout however at times the aftereffects of this investigation were not the same as the previous studies while thinking about age, sexual orientation, working during studies, number of companions in the ICT field like elements.

The finding of one study related to dropout prediction propose that it could be of these elements are excessively vital in impacting the first year dropout in ICT higher education. It recommends that students' socioeconomics, student income, inspiration, performance in the college, student mental condition, institutional attributes, and year of studies are influencing to dropouts in ICT advanced education. Aside from the earlier scholastic variables, social and statistic factors some more student-focused components, institutional factors were considered in that study investigation.[11].

Another important study has observed students' qualities and clustered student groups dependent on their earlier academic progress and socio-statistic attributes. The exploration set apart out the measurably critical connection between the kind of secondary school and secondary school grades with institutes, Grade Point Average. Further demonstrated that the extent of the prediction was observed to be higher than the one with socio-statistic qualities and when the connection between scholarly achievement and normal for explicit subgroups was analyzed, secondary school program was observed to be a critical indicator of scholastic achievement. The greatness of the expectation was observed to be higher than the one with socio-statistic qualities

The results of the study conclude that "One size fits all" model cannot be effectively applied to all students and the solution is to segment students and perform different approach for each group[12].

Table 1 summarizes the proposed dropout build upon factors found in the literature. From this table, it is evident that most of the researchers found student income, demographics, Institutional Characteristics, Performance in the university, psychology, social integration as important attributes for predicting students' dropouts.

Student Income	Demographics	Institutional Characteristics	Performance in University	Psychology	Social Integration
*Household Income *Financial Aids *Working during studies	*Gender *Age *Marital Status *Parental Status *Place of residence	*Students Demographics Structure *Academic Support *Financial Support	*Earned Credit Points *Parent Education *Prior Study State *Mathematics Exam	*Satisfaction *Stress *Emotional Exhaustion *Perceived Quality of Education *Expectation	*Student Involvement *Relation with other students *Relation with Institute *Campus environment

Table 2.1: Factors that influence the dropouts in the higher education system

2.2.3 Current Status of Educational Data Mining

Presently EDM uses different techniques and methodologies to extract knowledge from data generated by the higher educational process and supports for predicting future and change the future. When summarizing all the finding of reviewed studies many investigators are in agreement with that there are different application areas where the EDM can be used to extract the hidden knowledge and recommend that following application region for EDM?

- Improve the current student models,
- Improve the ongoing domain models,
- Studying the academic arrangement stipulated by higher education programs, and
- Regular examination of students and their learning;

Further EDM is concerned with extracting knowledge from data gathered in the educational databases as data mining explore the relationship between variables stored in the educational databases to more readily comprehend learners , and setting which they learn [13].

Distinguishing proof of applicable credits which influence to dropouts in advanced education is the main worry in the field of data mining. There are distinctive techniques in EDM: Linear Regression, Clustering, Classification and Association Rule Mining. Utilization of Prediction and discovery strategies are getting to be popular and new investigations will rise up out of assessing classification models built from a few unique algorithms including Bayes classifier, decision tree, boosting techniques and support vector machines.

Dropouts in higher education rely upon sociodemographic, individual, scholarly and institutional attributes. A portion of the variables included found in this literature study is more student-focused and very little should be possible by others to impact student dropouts. The components that are hard to impact are socioeconomics; desires; earlier examinations; instructive arrangement; and year of studies. A few variables could be affected by higher education organizations so as to hold their student. Further examinations would be completed to plan explicit dropout expectation models for specific segments and rethink the entry requirements for higher education.

2. 3 Chapter Summary

As detailed in this chapter, the limitations of earlier studies set the base for the main contributions of this study. In efforts to answer the question whether the classification algorithms used in data mining predictively analyze the dropouts in Information Technology higher education from data collected from the Higher National Diploma in Information Technology course of Sri Lanka Institute of Advanced Technological Education do better, this study will explore the power of WEKA data mining tool to construct and evaluate classifier models using Decision Tree(J48), K-Nearest Neighbor (Lazy- IBK), Naïve Bayes and Rule-Based (ZeroR) algorithms. Then select the most accurate classifier model whose functionality is described in Chapter 3. Further Java WEKA API is used to develop the student dropout prediction application.

Chapter 3

3 Predictive Analysis through Classification using WEKA Data Mining Tool

3.1 Chapter Introduction

This chapter details the main technologies adopted reasons behind selecting the technologies mentioned and their applicability to this study.

3.2 Correlation-Based Feature Selection

Feature selection is utilized to choose a subset of input data most helpful for analysis and future prediction by removing features, which are immaterial of predicting information. It is utilized for expanding the accuracy of prediction and lessening the complexity of student results [12]. In the present study, Correlation-Based Feature Selection (CFS) was utilized to discover the attribute subsets that are exceptionally connected with the class however minimal correlation between features joined with search methodology best-first search (BFS). Best First Search technique begins with a vacant set of highlights and creates all conceivable single feature expansion. The subset with the highest assessment is picked and extended in a similar way by including a single feature. In the event that extending a subset results in no improvement, the search back to the following best-unexpanded subset and proceeds from that point. The search will end if five back to back completely extended subsets demonstrate no improvement over the present best subset. Decreased datasets might be passed to machine learning (ML) for building a classifier to predict the dropout student. The feature selection essentially influences the accuracy of the classifier [14].

3.3 Brief Introduction to Classification Techniques used in WEKA

Weka makes a substantial number of classification algorithms accessible. Countless machine learning algorithms available is one of the advantages of utilizing the Weka platform to work through machine learning issues.

Classification is a standout amongst the most well-known application areas of information mining. The fundamental undertaking in classification is appointing a class label among a lot of conceivable class esteems to an instance composed of a set of attributes. It is finished by utilizing a classifier model, which is worked by applying a learning algorithm on a training set made out of past instances having an indistinguishable variable set from the concealed instances.

The class mark of each instance in the training set is obviously known before training. After the learning stage, the classification performance of the classifier model constructed is assessed on a free test set before utilized. In classification, there is a wide range of methods and algorithms conceivable to use for building a classifier model. The absolute most mainstream ones can be considered decision tree algorithms, K-Nearest Neighbors, Bayesian belief networks systems, Support Vector Machines (SVM), Artificial Neural Networks(ANN), discriminant analysis(DA), Logistic regression, and Rule-based systems[2]. In this study, the first four of these are used.

3.3.1. Decision Tree Algorithms

Decision trees frequently emulate the human dimension considering so it is so easy to comprehend the data and settle on some great clarifications and Decision trees really make the client see the rationale for the information to interpret(not like discovery algorithms like SVM, K-NN, etc..) A decision tree can be communicated as a tree where each node represents a feature, each link (branch) represents a decision (rule), and leaf represents an outcome. The entire thought is to make a tree like this for the whole data and process a solitary result at each leaf (or limit the mistake in each leaf)

A decision tree algorithm expects to recursively split the observations into totally unrelated subgroups until there is no further split that has any kind of effect regarding factual or debasement measures. Among the impurity estimates that are utilized to find the homogeneity of instances in a node of the tree, Information Gain, Gain Ratio, and Gini Index are the most notable ones. More often than Gain is utilized in Iterative Dichotomiser (ID3), Gain Ratio in C4.5 and C5.0 (the successors of ID3) though Gini Index is utilized in Classification and Regression Trees (CART).

Decision Tree is a powerful and prominent prediction method, and it is the most well-known data mining method in the literature. There are a few prevalent decision tree algorithms, for example, ID3, C4.5 (J48), and C5. 0, and CART (characterization and relapse trees). Decision Tree is as a tree structure, where every node is either a leaf node (demonstrating the value of the target class of instances) or a decision node (determining a test to be completed on a single attribute value, with one branch and sub-tree for every conceivable result of the test).Decision Tree has numerous points of interest, for example, extremely quick classification of unknown records, simple understanding of little measured trees, strong structure to the outliers' effects, and an unmistakable sign of most imperative attribute for prediction, however, DTs are delicate to over-fitting especially in little information sets[10].

In this study, to generate a decision tree, the J48 algorithm was used, which is an extension of Quinlan's earlier ID3 algorithm. To construct the tree, the entropy measure was used in the determination of nodes. Since the attributes with higher the entropy cause more uncertainty in outcome, they were selected in order of increasing entropy.

3.3.2 Naive Bayes Algorithms

This is an unrealistic assumption because we expect the variables to interact and be dependent, although this assumption makes the probabilities fast and easy to calculate. Even under this unrealistic assumption, Naive Bayes has been shown to be a very effective classification algorithm.

Naïve Bayes is a classification method, and it accepts that the input values of attributes are nominal, in spite of the fact that it numerical sources of inputs are supported by assuming a distribution.

Naive Bayes utilizes a basic usage of Bayes Theorem (hence naive) where the prior probability for each class is determined from the training data and assumes to be autonomous of one another (actually called conditionally independent).

This is a farfetched assumption since we anticipate that the factors should connect and be needy, despite the fact that this assumption makes the probabilities quick and simple to ascertain. Indeed, even under this impossible assumption, Naive Bayes has been appeared to be an extremely compelling classification method.

Naïve Bayes ascertains the posterior probability for each class and makes a prediction for the class with the most elevated probability. Likewise, it supports both binary classification and multi-class classification issues.

Naturally, a Gaussian distribution is assumed for each numerical attributes.

You can change the algorithm to utilize a kernel estimator with the use Kernel Estimator contention that may better match the real circulation of the attributes in the dataset. On the other hand, you can naturally change over numerical attributes to nominal attributes with the utilization of Supervised Discretization parameter.

3.3.3 K-Nearest Neighbors Algorithms

The K-Nearest Neighbors algorithm supports both classification and regression. It is called KNN for short. KNN works by sorting the whole training dataset and querying it to find the k most comparable training patterns when making a prediction. Accordingly, there is no model other than the initial training dataset and the main calculation performed is the querying of the training dataset when a prediction is asked.

It is a straightforward algorithm, yet one that does not accept particularly about the issue other than that the separation between instances is significant in making predictions. When all measures are considered, it frequently accomplishes extremely great execution. When making predictions on classification issues, KNN will take the mode (most normal class) of the k most comparative examples in the training dataset. The measure of the area is constrained by the k parameter.

For instance, in the event that k is set to 1, at that point forecasts are made utilizing the single most comparative training instance to a given new pattern for which a prediction is asked. Basic values for k are 3, 7, 11 and 21, bigger for bigger dataset sizes. Weka can naturally find a good value for k utilizing cross-validation inside the algorithm by setting the cross-Validate parameter to true.

Another critical parameter is the distance measure utilized. This is designed in the Nearest-NeighbourSearchAlgorithm which controls the manner by which the training data is put away and searched.

The default is a linear search. Tapping the name of this search algorithm will give another arrangement window where there is a distance function parameter. As a matter of course, Euclidean distance is utilized to compute the distance between instances, which is useful for numerical data with a similar scale. Manhattan distance is great to utilize if attributes vary in measures or type.

3.3.4. Rule-based Algorithms

The term rule-based classification can be utilized to allude to any classification scheme plot that makes utilization of IF-THEN rules for class prediction. Rule based classification plots commonly comprise of the below mentioned segments:

Rule Induction Algorithm This alludes to the way toward extricating applicable IF-THEN rules from the data which should be possible straightforwardly utilizing sequential covering algorithms or in a indirectly from other data mining methods like decision tree building or association rule mining.

Rule Ranking Measures: This alludes to a few values that are utilized to quantify the helpfulness of a rule in giving accurate prediction. Rule ranking measures are regularly utilized in the rule induction algorithms to prune off superfluous rules and improve effectiveness. They are utilized in the class prediction algorithm to give a ranking to the rules which will be then be used to predict the class of new cases. ZeroR is the least difficult classification method which relies on the target and overlooks all predictors. ZeroR classifier essentially predicts the majority category (class). In spite of the fact that there is no predictability power in ZeroR, it is valuable for deciding a pattern performance as a benchmark for other classification techniques. This algorithm construct a frequency table for the target and select its most successive esteem.

3.4 Model Evaluation

There are some execution measures to assess classification models as far as the accuracy of the classification decision of the model. Accepting a binary classification task, the class variable values might be expected as Positive (P) and Negative (N). Actual l positives (P) that are effectively named as positives by the classifier are named as true positives (TP) while genuine positives inaccurately named as negatives by the classifier are considered as false negatives (FN). Along these lines, true negatives (N) that are effectively named as negatives are taken as true negatives (TN) while true negatives inaccurately named as positives are considered as false negatives (FP). These terms are given in the confusion matrix of Table: 3.1.

		Predicted Class		
Actual Class		Positive	Negative	Total
	Positive	TP	FN	P
	Negative	FP	TN	N
	Total	P’	N’	P+N

Table: 3.1 Confusion matrix for performance evaluation

The calculations of performance estimates, for example, accuracy (recognition rate), precision, and recall (affectability or true positive rate) are given in the accompanying formulas. Precision estimates the rate of correctness of the class predictions to all predictions. Precision estimates the accuracy rate of the class predictions done as positive by the classifier though recall estimates the rate of positives effectively predicted as positive by the classifier.

$$\textbf{Accuracy} = \frac{\textbf{TP TN}}{\textbf{TP TN FP FN}} \ \mathbf{*} \ \mathbf{100}$$

Precision: exactness – what % of examples that the classifier labeled as positive are actually positive

$$\textbf{Precision} = \frac{\textbf{TP}}{\textbf{TP + FP}} * 100$$

Recall: completeness – what % of positive examples did the classifier label as positive?

$$\textbf{Recall} = \frac{\textbf{TP}}{\textbf{TP+FN}} * 100$$

3.5 Java WEKA API

Utilizing the graphical tools, similar to the WEKA Explorer, or simply the command line is adequate for the typical user. In any case, WEKA's unmistakably defined API ("application programming interface") makes it extremely simple to "embed" it in other projects that need support for data mining tasks.

Weka is a standard Java tool for performing both machine-learning experiments and for installing trained models in Java applications. It tends to be utilized for supervised and unsupervised learning. There are three different ways to utilize WEK first utilizing the command line, second, utilizing Weka GUI, and third through its API with Java. Weka's library gives a vast gathering of machine learning algorithms, implemented in Java.

Weka's library in java gives a vast gathering of classes and methods to embed WEKA in java programs.

3.6 Chapter Summary

As explained in this chapter, classification algorithms can be effectively used to address the research problem of this study. Prediction of new student dropout can be supported from the application developed using Java WAKA API. High-level approach as to how the technology of classifying can be utilized is further detailed in the next chapter.

Chapter 4
4 J48 Decision Tree Algorithm to Predict Student Dropouts

4.1 Chapter Introduction

This chapter defines the research hypothesis and briefs the high-level approach to solving the problem definition identified in Chapter 2, by adopting the said technology. The rationale behind the approach is also provided in this chapter.

4.2 Research Questions

Thus, this study attempts to answer the question "Can classification algorithms used in data mining predictively analyze the dropouts in Information Technology higher education?" In achieving above, four contributions are identified,

1. Can correlation-based feature selection generate the dataset with most suitable factors affecting to dropouts
2. What are the most relevant factors to predictively analyze student dropouts
3. What is the most accurate classification algorithm to construct the classification model
4. Can the power of prediction tools such as decision tree algorithm are able to predict the student dropouts in information technology higher education accurately

4.3 Input

4.3.1 Selection of specific dataset

In predicting the student dropouts in Information Technology Higher Education information extracted from the Higher National Diploma in Information Technology course of Sri Lanka Institute of Advanced Technological Education was selected. This higher education course was selected since "Dropout" is a complex issue and even more critical in this institute and hence assume to collect sufficient and relevant dataset for investigation.

4.3.2 Dropout data

Accomplishing the aim of this examination requires a dependable, adaptable and an early appraisal of the dropout information. To build up a steady model, need to utilize prepared data, including the target class attribute. This is known as the training dataset and it is utilized to make a model.

Likewise, need to check the validity of the made model with another known dataset called the test dataset.

In this study, a dataset was accumulated by following 4000 students enlisted at Advanced Technological Institute Labuduwa for Higher National Diploma in Information Technology course considered from 2014 to 2017.

4.4 Output

As stated in Chapter 2, as output, this study makes four contributions in answering the main research problem “Can classification algorithms used in data mining predictively analyze the dropouts in Information Technology higher education?”

First, the study attempts to identify important factors affecting dropouts by reviewing the related literature. By applying correlation-based feature selection most important factors could be identified. Secondly, this study seeks to explore the most accurate classifier algorithm that can be used to construct the predictive analysis model by applying different classifying algorithms on the collected dataset and evaluating their performance. While evaluating the performance of the classifier models, it will also answer the question whether the power of prediction tool such as decision tree algorithm is able to predict the student dropouts in information technology higher education accurately by considering the percentage value of model accuracy.

4.5 Process

By looking into the related work factors influencing to the dropout, imperative attributes to be considered was resolved and dataset was gathered from the Institute records and standard questionnaire by focusing attention regarding influenced attributes. Dataset was entered to MS Excel worksheet.

Data preprocessing was connected to gauge the quality and reasonableness of data by stacking the dataset into WEKA data mining tool. For this, remove missing values; smoothing noisy data, determination of important attributes from the dataset or evacuating insignificant attributes, recognizing or removing outlier values from a dataset, and resolving inconsistencies of data was finished.

After preprocessing, Dataset was prepared to apply the data mining techniques. Used Feature selection to choose a subset of input data most helpful for analysis and prediction by taking out attributes, which are irrelevant for predicting. Do these means over and over until getting the great model.

Most exact classifier algorithms by reviewing techniques used in past studies [1], Decision Tree (J48), K-Nearest Neighbor (Lazy-IBK), Naïve Bayes and rule-based (ZeroR) were connected on the dataset and assessed by considering their Confusion metrics measures.

Classification algorithms build the classifier, and the classifier is worked from the training dataset instances and their associated class labels.

Here the test data is utilized to assess the exactness of classification rules. The classification rules can be connected to the new data tuples if the accuracy is viewed as satisfactory.

J48 Decision tree and IBK K-Nearest Neighbor calculation were recognized as the most accurate classification algorithms for this dataset and chose it to build the prescient examination display since it has 100% exactness of anticipating dropouts.

In view of the classification model, the implementation of dropout prediction application was finished utilizing Java WEKA API.

Finally, test the entire application utilizing new dataset got from Institute records.

4.6 Users

The success of this research will definitely be utilized as a foundation for the development of a decision support system at the student selection process. Administrators can consider the consequences of the study and can utilize the extracted knowledge to improve the structure of the mediation programs that have significant effect on dropout, particularly for the first year. To viably use the constrained help assets for the intercession programs, it will be attractive to recognize ahead of time, students who will in general need the help and this study support administrators in distinguishing such students.

Additionally, the prediction of students' success at the entrance point is critical and need more consideration on distinct attributes to get increasingly precise outcomes that can.

4.7 Chapter Summary

As briefed in this chapter, this study extracts the factors affecting to dropouts in the higher education sector by reviewing the literature. Feature selection approach is then utilized in identifying important factors to better suit the predictive analysis. Application of classification algorithms selecting the most appropriate classifier and implementing the dropout prediction is done to support the predictive analysis. Next chapter on design provides a more detail view on the use of this approach.

Chapter 5

5 Design

5.1 Chapter Introduction

This section subtleties the analysis and design of the proposed system. Each step in the process, as briefed in Chapter 4, its“ inputs, outputs and the dependencies are described. The figure for the high-level design of the system is also provided (Figure 5.1).

5.2 Analysis and Design

5.2.1 Collection of student data

In this study, information on students’ Income, Demographics factors, Institution Characteristics, Psychological factors, Social Integration factors affects to student dropouts were collected and data preprocessing was connected to gauge the quality and appropriateness of data.

5.2.2 Pre-processing

The information utilized in this study was set up from the higher education institutions through a structured questionnaire. The survey has been developed dependent on hypothetical and observational grounds about factors influencing student performance. The survey included socio-demographics pointers (Age, Date of birth, Geographical area, status, Parents instruction, Parents occupation and Monthly income), Educational components (Performance in O/L Examination, A/L Examination , Prior ICT Education, English test Grade of Entrance test and so on.), Performance of institution, Psychological elements and Social Integration factors and Institutional elements, and so forth. Information was gathered from the diplomatists of SLIATE.

The data format is shown in Table: 5.1. Before the fundamental visit to review the records, a coding structure was made for each factor to be accounted for (e.g., for names of Geographical territories URBAN, SUBURBAN, RURAL). The socio-demographic data, school factors, family segment and attitude towards institution were inspected to support this investigation.

FACTOR	DESCRIPTION	POSSIBLE VALES
AGE	Age	{20-25 }
YEAR_B	YEAR OF BIRTH	YEAR
A_TYPE	Admission Type	{1ST ATTEMP, 2ND ATTEMPT, 3RD ATTEMPT}
GENDER	Gender	{MALE, FEMALE}
CITY	Geographic Location	{URBAN, SUB URBAN, RURAL}
STAY_LOC	Stay Location while Studying	{HOME, BORDING}
HOUSE_INCOME	House Hold Income	ANY CURRENCY VALUE
F_EDU	Father Education Level	{BELOWO/L, O/L, A/L, CERTIFICATE, DIPLOMA, DEGREE, POST GRADUATE,}
M_EDU	Mother Education Level	{BELOWO/L, O/L, A/L, CERTIFICATE, DIPLOMA, DEGREE, POST GRADUATE,}
FAMILY_STAT	Family Status	{LOW, MIDDLE, GOOD, HIGH, EXCELLENT}
WK_W_STD	Work While Study	{YES, NO}
O/L_PER	O/L Performance	{EXCELLENT, VERY_GOOD, GOOD, FAIR, LOW}
ICT_GRADE	ICT Grade at O/L	{A, B, C, S, F}
STREEM	A/L Streem	{ART, COMMERCE, BIO, MATHS, TECHNOLOGY}
ZSCORE	Z_Score	
PRIOR_EX	Prior Experience	{EXCELLENT, GOOD, FAIR, LOW, NO}
ENG_SCORE	English Grade at Entrance	{A, B, C, S, F}
ACC_PERFOR	Academic Performance at the Entrance Test	{EXCELLENT, HIGH, GOOD, MEDIUM, LOW}
SAT_LEVEL	Satisfaction Level of Study	{EPERCIEVEDXCELLENT, GOOD,LOW,NOT SATISFIED}
STRESS	Emotional Exhaustion	{LOW, MEDIUM, HIGH}
PERCIEVED_QLT	Perceived Quality of Education at the Institute	{LOW, MEDIUM, HIGH}
SOCO_INT	Social Integration	{LOW, MEDIUM, HIGH}
DROPOUT	Drop out	{YES,NO}

Table 5.1: Student Information Data Formats

5.2.3 Data Selection and Transformation

After collecting data, the collected dataset was set up to apply the data mining methods. Prior to utilization of prescribed techniques from the literature review, data preprocessing was connected to gauge the quality and appropriateness of information. For this, expel missing values; settling irregularities of data, choice of a relevant attribute from a dataset or evacuating insignificant parameters, distinguishing or expel exception esteems from the dataset, and settling irregularities of data were done. A portion of the insignificant parameters was expelled from the dataset, for example, age, year birth the WK_W_STD field containing just single esteem no, the marital status field containing one value- unmarried, etc.

A grade scale is utilized for assessment of student success at O/L. “EXCELLENT” students are viewed as the individuals who have a rate more prominent than 85, "VERY GOOD"- in the range somewhere in the range of 75 and 85, "GOOD"- in the range somewhere in the range of 65 and 75, and "FAIR" in the range somewhere in the range of 50 and 65, and "LOW" in the range beneath 65. A four-level scale is utilized in the yearly family income. Discretization was done on HOUSE_INCOME and classified the qualities into four canisters. A straight out target variable "Dropout" is developed; it has two conceivable values "Yes" (a student who totally pull back from their course) and "No"(students who are proceeding with their examination). To apply the classification, the complete dataset was converted to nominal values.

The final dataset utilized for the study contains 4000 examples each depicted with 20 attributes (1 output and 19 input variables), nominal. The study is constrained to the students’ data for two batches as of late gone out. Finally, the pre-processed information was converted into an appropriate setup to apply data mining techniques.

5.2.4 Feature extraction: Identifying important factors affecting to dropouts

During this pre-processing, the dataset was subject to feature selection and important factors affecting to apply classification algorithms were identified. As the output of this step, dataset with overall factors affecting to dropouts was available.

5.3 Model building by using various data mining technologies

The following stage followed to enter the pruned student dataset into WEKA data mining tool. This aided in assessing captivating outcomes by applying classification algorithms on the student training dataset.

After the authentic gathering, exploration, and transformation of data utilizing suitable measures, data mining classification techniques were connected to predict students' dropout rates of their study program. Decision Tree (J48), K-Nearest Neighbor (Lazy-IBK), Naïve Bayes and Rule-Based (ZeroR) were connected on the dataset and assessed by considering their Confusion metrics measures. The classification model was implemented by utilizing the WEKA data mining tool. There are a few classifiers accessible in WEKA yet J48 was utilized with the end goal of the investigation since it showed the most noteworthy exact prediction.

5.4 Deployment of GUI application to predict student dropouts

This study executes the standalone java application of the Dropout Prediction to foresee which students are probably going to drop out. Build up the application utilizing java WEKA API and container jar file will be made to deploy the application. Utilization of the application will predict "Survive" or "Retired" which the investigation goes for diminishing dropout rates of the students in both first and second of the academic year and improving the effectiveness of course.

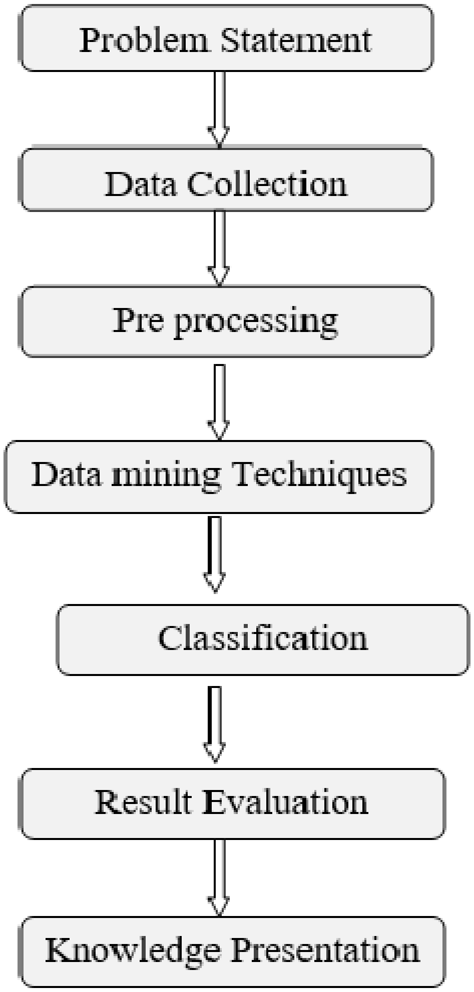


Figure 5.1: High-Level design of the proposed system

5.5 Chapter Summary

This chapter was written with the aim of providing a detailed view of the design of the system whereas the implementation related details will be described in the next section.

Chapter 6

6 Implementation of the Dropout Prediction System

6.1 Chapter Introduction

This section furnishes the implementation choices made regarding the design of the proposed framework as detailed in Chapter 5.

6.2 Implementation

6.2.1 Building up classifier models and evaluate for accuracy

In this study, four classification models are delivered: one using Decision Tree algorithm (J48), one using NaïveBayers, one using K-NearestNeighbor (Lazy-IBK), and one using Rule-based (ZeroR). The presentations of these models are surveyed on the test data with respect to accuracy, precision and recall.

6.2.2 Development of Dropout Prediction system

6.2.2.1 Development of GUI of Dropout Prediction system

This module implements the GUI screen for the client to choose the dataset and after that predict a new student’s DROPOUT feature from different attributes.

Here a basic GUI with a capacity to choose the required dataset to classify and enter attributes distinguished as adding to dropout choice was created to encourage predictive analysis of students’ profiles. Dataset was set up in .csv configuration, and interface of the prediction application enables the client to peruse and choose a dataset from the local file system.

6.2.2.2 Buildup the classifier model

Classify Model module takes the .csv document containing student dataset as input. In the event that the record has an alternate document expansion that is typically connected with the loader, it utilizes a loader straightforwardly which is fit for lording .csv document to the java WEKA API. At that point create a dataset memory structure by defining the format of the data by setting up the attributes and adding the actual data, row by row.

Then it builds the classifier utilizing a training dataset in the whole dataset at once and uses a trained classifier tree to mark every instance in an unlabeled dataset (test dataset) that gets loaded from disk.

After a classifier setup has been assessed and ended up being helpful, the j48 classifier was chosen to build classifier model as it ended up being most accurate classifier algorithm when it assessed with another tree top most well-known models (Naïve Bayers, K-Nearest, rule-based).

6.2.2.3 Classify the new data

The built classifier is used to make predictions and label previously unlabeled data. When the user enters a new student data system will classify new data instance and then predict a new student DROPOUT feature from other attributes

6.3 Chapter Summary

This chapter was a briefing on the implementation decisions that were made. Results of this study will be analyzed in the next chapter.

Chapter 7

7 Results and Evaluation

7.1 Chapter Introduction

This section talks about the assessment of the solution for seeing whether goals have been accomplished and present the outcomes from the assessment in diagrams, outlines, tables, and so on. Extra insights concerning the assessment may go as an appendix.

7.2 Results

7.2.1 Performance evaluation of classifiers using stratified cross-validation

In this study, four classification models have produced: one Decision Tree calculations (J48), one utilizing Navie Bayers, K- Nearest (lazy-IBK), and one utilizing Rule Based (ZeroR). Divided the test dataset of 2000 instances to 500, 600, 700, 800, 900, and 1000 and checked for accuracy changers for selected algorithms (Naive Bayas, Lazy-IBK, j48, Rule-ZeroR).

Re-evaluate model summary of each of the four classifiers are given in Fig. 7.1, 7.2, 7.3 and 7.4 separately.

7.2.1.1 Decision Tree algorithms (J48) Classifier

No of Instances	Accuracy
500	100
600	100
700	100
800	100
900	100
1000	100
2000	100

Table 7.1: Re-evaluate model summary of Decision Tree (j48) Classifier

7.2.1.2 NaiveBayers Classifier

No of Instances	Accuracy
500	96.0396
600	96.0396
700	96.0396
800	96.0396
900	96.0396
1000	96.0396
2000	96.0396

Table 7.2: Re-evaluate model summary of NaiveBayers Classifier

7.2.1.3 KNearestNeighbor (Lazy-IBK) Classifier

No of Instances	Accuracy
500	100
600	100
700	100
800	100
900	100
1000	100
2000	100

Table 7.3: Re-evaluate model summary of KNearestNeighbor (Lazy-IBK) Classifier

7.2.1.4 rule-based (ZeroR)

No of Instances	Accuracy
500	83.1683
600	83.1683
700	83.1683
800	83.1683
900	83.1683
1000	83.1683
2000	83.1683

Table 7.4: Re-evaluate model summary of rule-based (ZeroR) Classifier

7.2.2 Accuracy comparison by increasing the number of instances in test dataset

7.2.2.1 Decision Tree - J48 Classifier

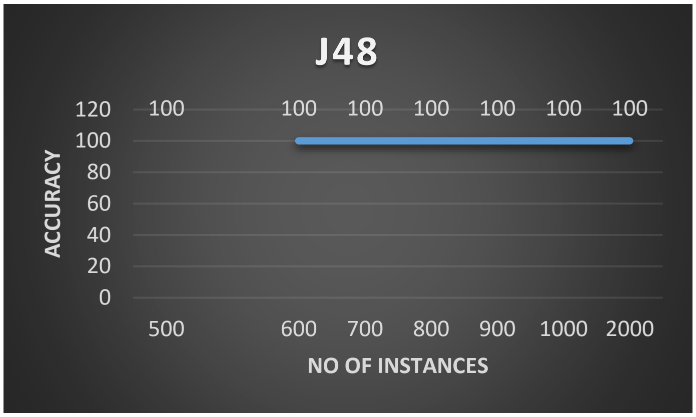


Figure: 7.1 Accuracy comparison by increasing the number of instances in test dataset of Decision Tree algorithms (J48) Classifier

7.2.2.2 NaiveByayes Classifier

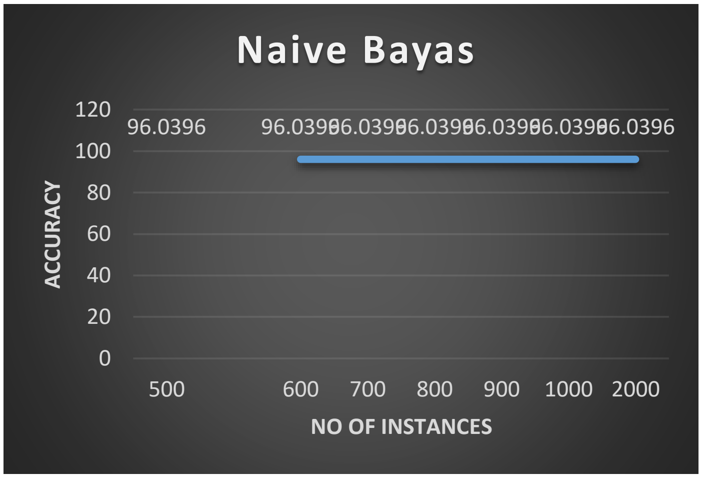


Figure: 7.2 Accuracy comparison by increasing the number of instances in test dataset of naive Bayes Classifier

7.2.2.3 KNearestNeighbor - Lazy.IBK Classifier

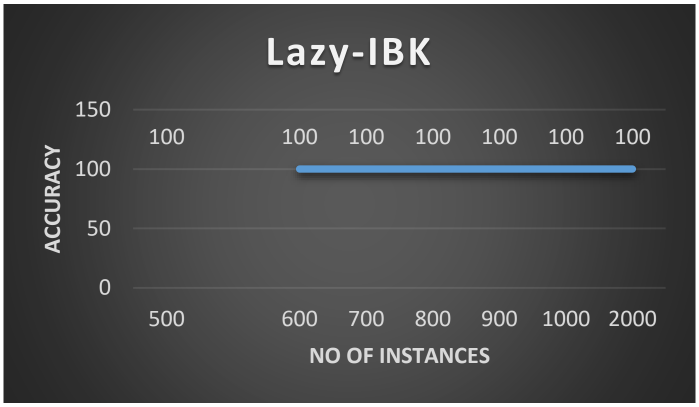


Figure: 7.3 Accuracy comparison by increasing the number of instances in test dataset of KNearestNeighbor - Lazy.IBK Classifier

7.2.2.4 Rulebased - ZeroR Classifier

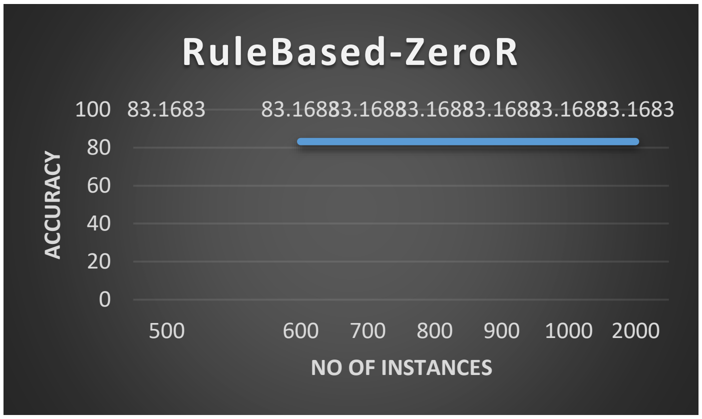


Figure: 7.4 Accuracy comparison by increasing the number of instances in test dataset of Rule-based - ZeroR Classifier

A performance comparison of all the classifiers associated is finished using assessment measures: accuracy, precision, and recall. Table: 7.5 concludes performance evaluation measures of all of the four classifiers. As can be seen from Table: 7.5, all classifiers give practically identical results on the test dataset. When we dissect the model performance, all of the techniques used in this investigation is convincing in requesting "Yes" and "No" dropout factor.

Model	Accuracy	Precision	Recall
Lazy.IBK	100	1.000	1.000
J48	100	1.000	1.000
Naive Bayes	96.0396	0.842	0.941
ZeroR	83.1683	0	0

Table: 7.5 Performance Evaluation Matrix of all Classifiers

This answers the first and second hypothesizes of this study which state that features selection techniques used in data mining can be used to select suitable factors affecting to dropouts in information technology higher education and Student Income, Demographics factors, Institution Characteristics, Psychological factors, Social Integration factors affects to student dropouts in information technology higher education since the accuracy esteems, which survey the viability of the models, are all at any rate roughly 83%. J48 and IBK classifier are the best in performance ZeroR is the most noticeably bad. Precision, which evaluates the predictive power, again demonstrated J48 and IBK as best classifiers. Recall esteems, which show the affectability and genuine positive rate (TPR) of the models, vary among classifiers. As indicated by recall, J48 and IBK are the best classifiers in performance and the Naïve Bayes is the third best. In addition, Zero is the most noticeably bad in performance as indicated by accuracy, precision, and recall.

Accordingly, this study reveals that despite the fact that all the classifiers are comparably great, from the highest classifiers J48 and IBK classifiers are nearly better. Further, decision trees really make the client see the rationale for the information to interpret and it is so easy to comprehend the data and settle on some great clarifications. At last, j48 can be considered as the remarkable classifier among all as per the given performance measures. This answers the third hypothesis of this study which expresses that j48 choice tree calculation best predicts the dropouts in higher education.

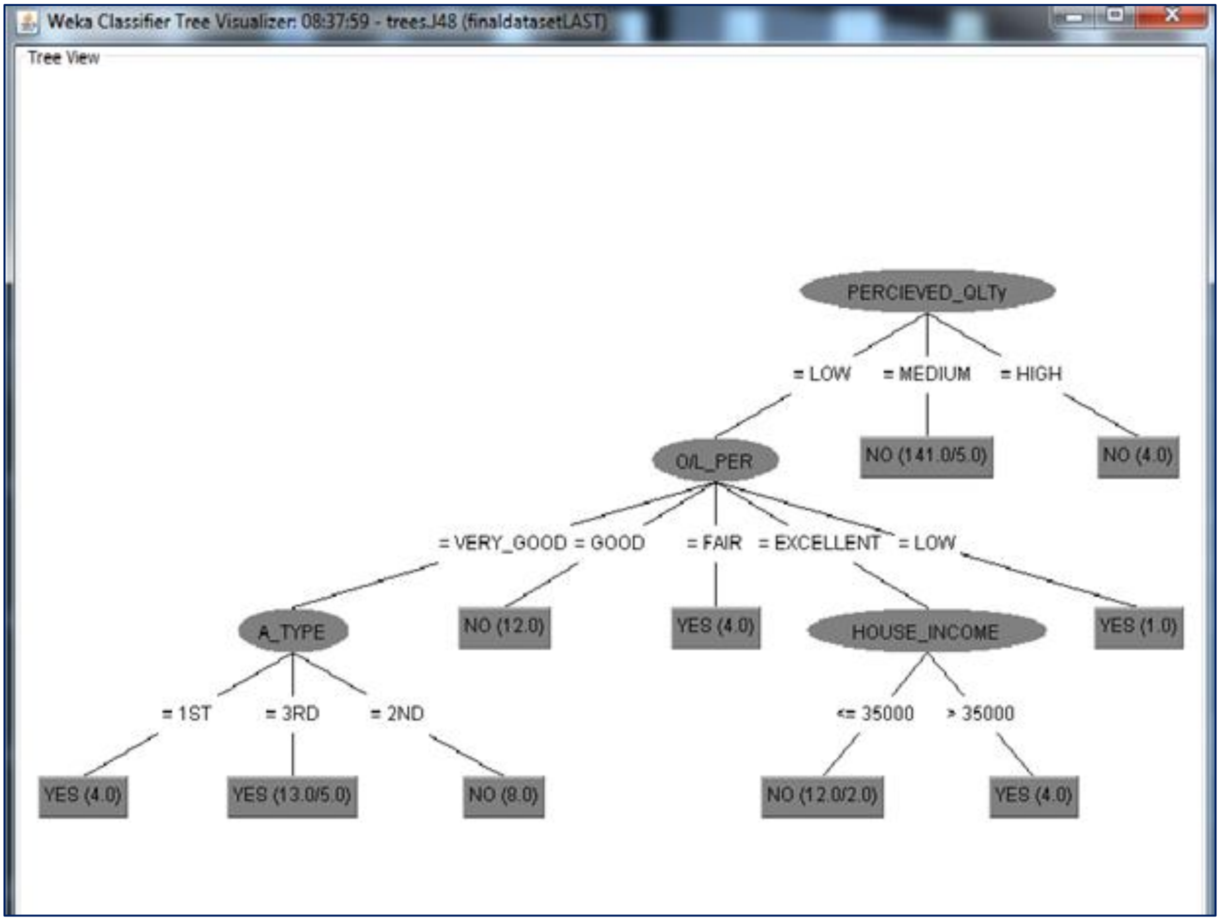
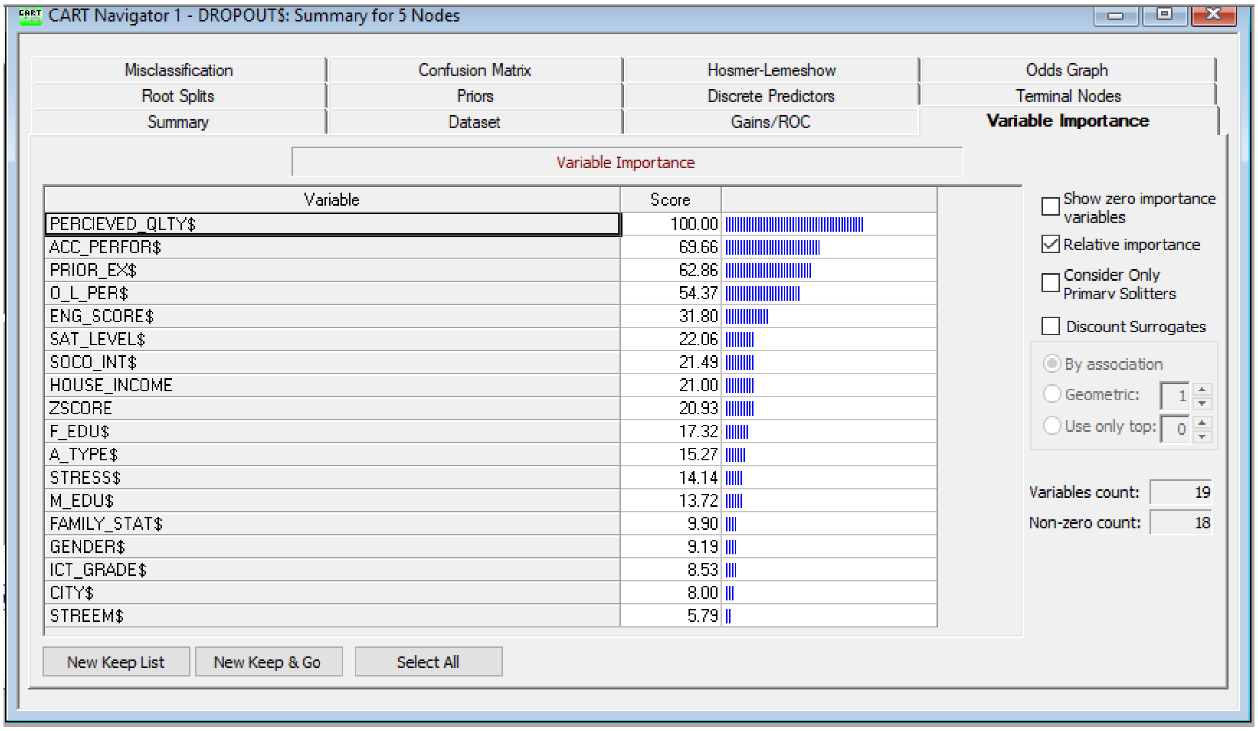


Figure 7.5. Decision Tree Diagram generated by j48 algorithm

Variable importance graph of the Decision Tree J48 classifier is given in Figure: 7.6. In J48 classifier, the top three important variables are PERCIEVED_QLTY, ACC_PERFOMANCE, PRIOR_EXPERIENCE which can be AS Figure 7.5.



Variable Importance

Variable	Score
PERCIEVED_QLTY$	100.00
ACC_PERFOR$	69.66
PRIOR_EX$	62.86
O_L_PER$	54.37
ENG_SCORE$	31.80
SAT_LEVEL$	22.06
SOCO_INT$	21.49
HOUSE_INCOME	21.00
ZSCORE	20.93
F_EDU$	17.32
A_TYPE$	15.27
STRESS$	14.14
M_EDU$	13.72
FAMILY_STAT$	9.90
GENDER$	9.19
ICT_GRADE$	8.53
CITY$	8.00
STREEM$	5.79

□ Show zero importance variables
☑ Relative importance
□ Consider Only Primary Splitters
□ Discount Surrogates
● By association
○ Geometric: 1
○ Use only top: 0

Variables count: 19
Non-zero count: 18

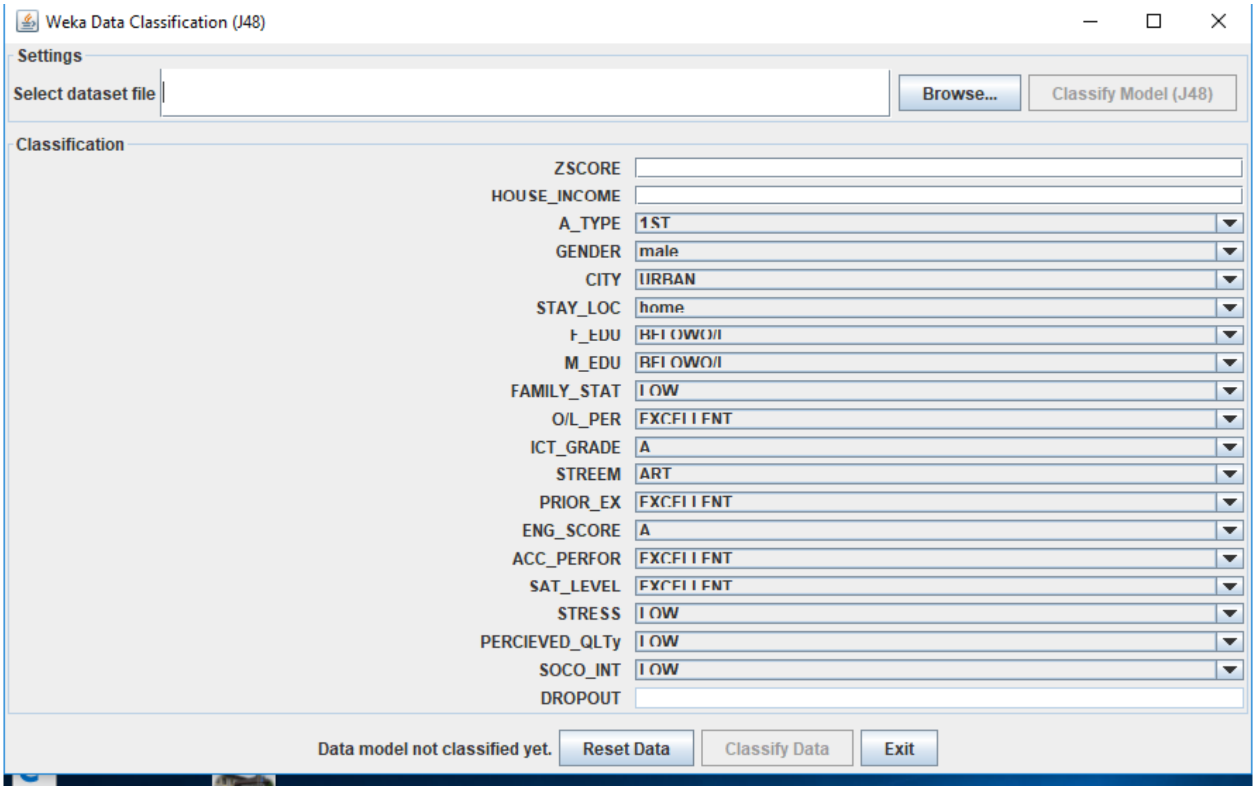
Figure: 7.6 Variable importance graph of the Decision Tree J48 classifier

7.3 Implemented the Predictive analysis model

WEKA helped essentially in finding concealed data from the training dataset. These recently learned predictive patterns for anticipating students' success would then be able to actualize in a working application to get the anticipated aftereffects of conceded students at the entrance to the course utilizing.

7.3.1 Browsing a file from local file system

The user can select csv file which contains dataset to be classified.



Select dataset file:

ZSCORE:

HOUSE_INCOME:

A_TYPE: 1ST

GENDER: male

CITY: URBAN

STAY_LOC: home

F_EDU: BELOWO/L

M_EDU: BELOWO/L

FAMILY_STAT: LOW

O/L_PER: EXCELLENT

ICT_GRADE: A

STREEM: ART

PRIOR_EX: EXCELLENT

ENG_SCORE: A

ACC_PERFOR: EXCELLENT

SAT_LEVEL: EXCELLENT

STRESS: LOW

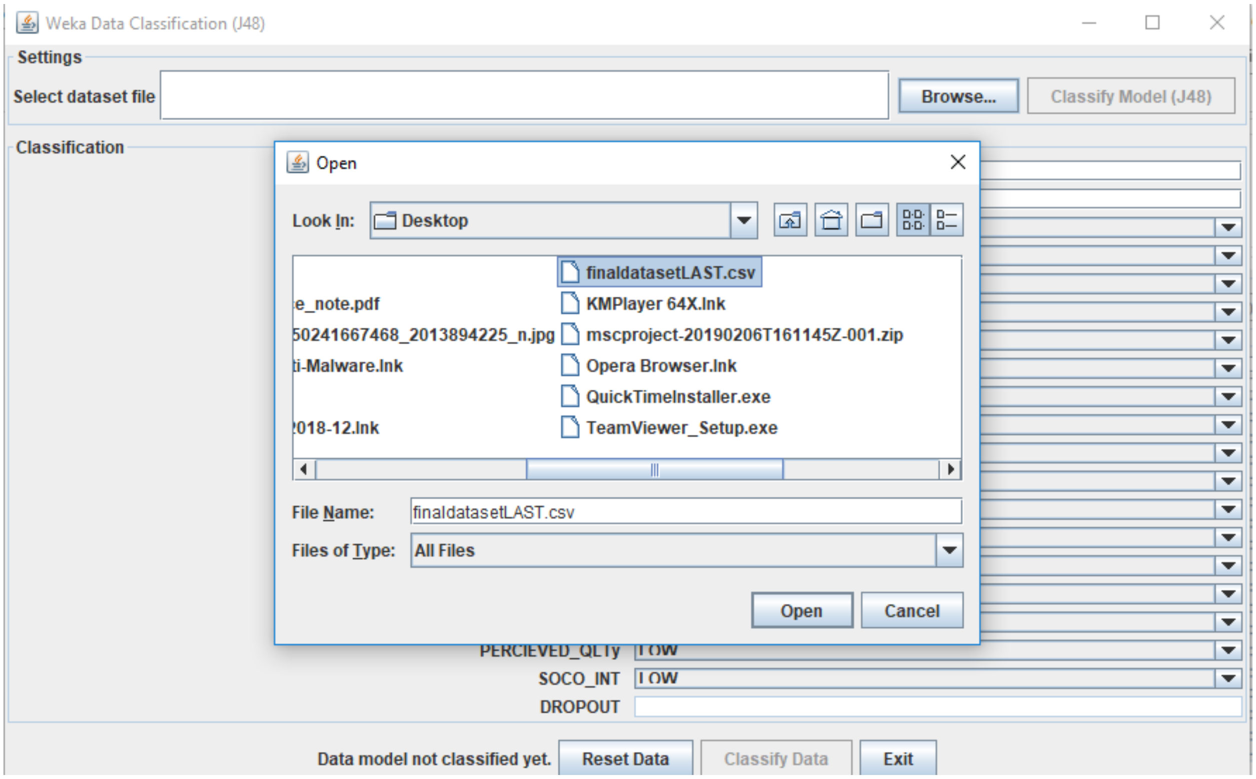
PERCIEVED_QLTy: LOW

SOCO_INT: LOW

DROPOUT:

Data model not classified yet.

Figure: 7.7a Browsing a file from local file system



Look In: Desktop

e_note.pdf
50241667468_2013894225_n.jpg
ti-Malware.lnk
2018-12.lnk
finaldatasetLAST.csv
KMPlayer 64X.lnk
mscproject-20190206T161145Z-001.zip
Opera Browser.lnk
QuickTimeInstaller.exe
TeamViewer_Setup.exe

File Name: finaldatasetLAST.csv

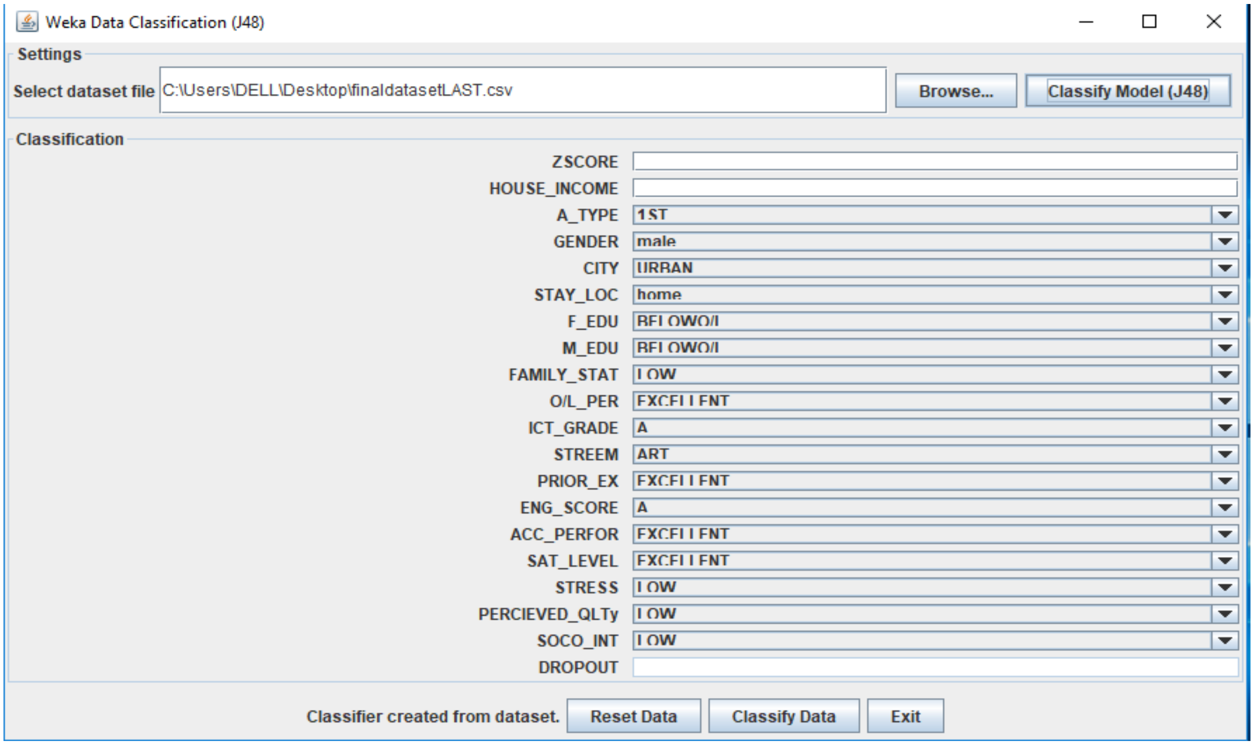
Files of Type: All Files

Data model not classified yet.

Figure: 7.7b User selects a file from local file system

7.3.2 Building the model using the J48 classifier

The user can select "Classify Model J48" to build the model for the given dataset. A message will appear prompting classifier created from the dataset.



Settings

Select dataset file C:\Users\DELL\Desktop\finaldatasetLAST.csv

Classification

ZSCORE

HOUSE_INCOME

A_TYPE 1ST

GENDER male

CITY URBAN

STAY_LOC home

F_EDU BELOWO/L

M_EDU BELOWO/L

FAMILY_STAT LOW

O/L_PER EXCELLENT

ICT_GRADE A

STREEM ART

PRIOR_EX EXCELLENT

ENG_SCORE A

ACC_PERFOR EXCELLENT

SAT_LEVEL EXCELLENT

STRESS LOW

PERCIEVED_QLTy LOW

SOCO_INT LOW

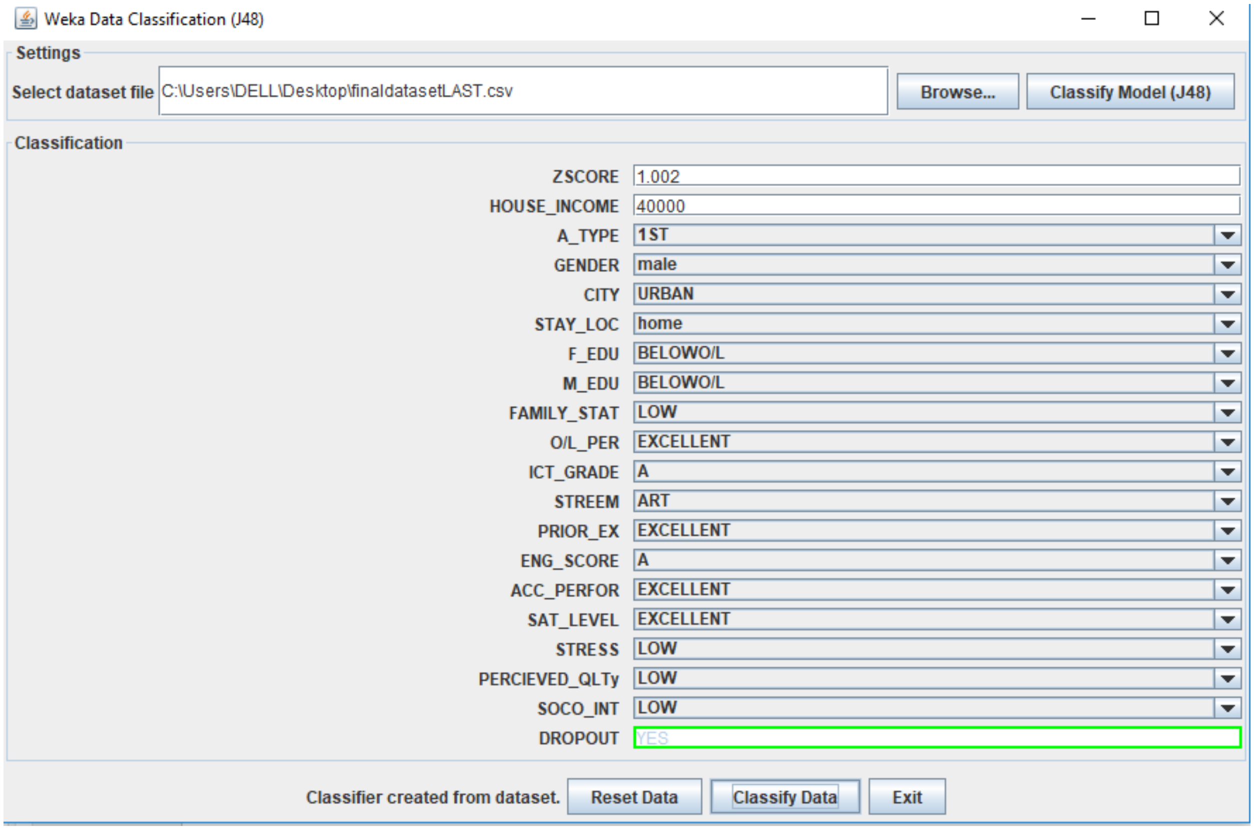
DROPOUT

Classifier created from dataset.

Figure: 7.8 Building the model using the Decision Tree J48 classifier

7.3.3 Classify new data

The user can fill up the required parameters in the form and select "classify data” button. It will classify new data record for a given student and predict the dropout feature. The field is highlighted in the screen.



Select dataset file C:\Users\DELL\Desktop\finaldatasetLAST.csv

ZSCORE	1.002
HOUSE_INCOME	40000
A_TYPE	1ST
GENDER	male
CITY	URBAN
STAY_LOC	home
F_EDU	BELOWO/L
M_EDU	BELOWO/L
FAMILY_STAT	LOW
O/L_PER	EXCELLENT
ICT_GRADE	A
STREEM	ART
PRIOR_EX	EXCELLENT
ENG_SCORE	A
ACC_PERFOR	EXCELLENT
SAT_LEVEL	EXCELLENT
STRESS	LOW
PERCIEVED_QLTy	LOW
SOCO_INT	LOW
DROPOUT	YES

Classifier created from dataset.

Figure: 7.9 Classify new data item

7.4 Chapter Summary

This chapter was an evaluation of the results of this study leading to the acceptance of all three hypotheses concluding that classification algorithms used in data mining can predictively analyze the dropouts in Information Technology higher education. Furthermore, the chapter also deduces that Decision Tree models such as J48, which are able to visualize the constructed model performs better understanding of important factors affects to dropouts in the context of this study.

Chapter 8

8 Conclusion and Further work

8.1 Chapter Introduction

This chapter explains the general accomplishments of this research, future work of this research and the restrictions recognized which may obstruct the materialness of the proposed arrangement.

8.2 Conclusion

This study could effectively foresee the student dropouts in data information technology higher education utilizing Decision tree calculation which can recognize the relationship between influencing factors and the dropout status with 100 % exactness.

In achieving above, a random sampling of a statistically acceptable volume of raw data forms one batch of students of Higher National Diploma in Information Technology was collected from Advanced Technological Institute Labuduwa. The study was able to identify strong correlations between identified dropout factors and the dropout feature of students. With successful training of a series of classification techniques like Decision Tree, K-Nearest Neighbor, Navie Bayers and Rule-based, the study found stronger relations between student dropout factors and dropout feature with more than 83% accuracy. J48 Decision Tree classifier and IBK K-Nearest Neighbor showed the highest accuracy of 100%.
Perceived quality, Prior academic performance, Prior experience in ICT, O/L performance and English score are the top most important factors affecting the prediction of dropout feature of the considered higher education sector.

These discoveries are in agreement with the early investigations [1], [6], and [7] which asserts that factors that characterize the academic performance and the prediction methods that can be utilized to decide the students' performance uncovers that Decision Tree, K-Nearest Neighbor, Naïve Bayes and Rule-Based have the expected precision to presciently dissect student profiles.

Additionally, the extracts of this study are in concurrence with the investigations [2], [8], [10] and [11] that they discovered student socioeconomics; student income; inspiration; performance in the institute ; student' psychological condition; institutional attributes; year of studies are influencing to dropouts in ICT higher education.

Dissimilar to past investigations, this study could execute a predictive analysis application utilizing Java WEKA API which can be utilized as a Decision Support System at the entrance to the courses and backing rethink the section necessities for the course considered. As it can distinguish ahead of time students who tend to need support most, the mediation projects can be updated to effectively affect dropout, particularly for the first year by adequately using the restricted help assets for the intercession programs.

8.3 Future Work

In this specific context, this work indicates how data mining techniques can be utilized to help early identification of potential dropouts, enabling the foundation to intercede. Further research with more information is important to maintain discoveries from this work, and feature selection would be utilized for expanding the predictive exactness and decreasing complexity.

Further association, rule mining system, ought to be utilized to find the connection between apparently random factors in the dataset and to relate to the high certainty what could be the essential variables influencing to decide the dropouts.

8.4 Limitations of the proposed solution

As most experts said information pre-preparation assessed to take 70-80% of the time and effort in the mining task. The more related data is accumulated and masterminded its quality and satisfaction, the better model's performance can be. More student records from different batches and various institutions have given increasingly noticeable authenticity to the results.

A better method to locate the correct classification criteria ought to be considered. Foreseeing dropout student commonly important to help the students improving their learning technique and to screen the students' academic success. Also, the investigations should consume thought on components relating to learning analytics of the advancement of students learning and the settings in which learning occurs since various foundations believe that the use of these examination improves learning and educating.

Utilizing the expanded accessibility of huge datasets around student action and advanced impressions left by student action in learning situations, learning analytics take further than data at present accessible can.

8.5 Chapter Summary

This chapter was an overview of the overall achievements of this study, further work to be addressed to fulfill the aim of this study and the possible loopholes of the approach.

References

[1] K. Limsathitwong, K. Tiwatthanont, and T. Yatsungnoen, "Dropout prediction system to reduce discontinue study rate of information technology students," *Proc. 2018 5th Int. Conf. Bus. Ind. Res. Smart Technol. Next Gener. Information, Eng. Bus. Soc. Sci. ICBIR 2018*, pp. 110–114, 2018.

[2] H. Gulati, "Predictive Analytics Using Data Mining Technique," *Comput. Sustain. Glob. Dev. (INDIACom), 2015 2nd Int. Conf.*, pp. 713–716, 2015.

[3] S. Sivakumar, S. Venkataraman, and R. Selvaraj, "Predictive modeling of student dropout indicators in educational data mining using improved decision tree," *Indian J. Sci. Technol.*, 2016.

[4] S. Ameri, M. J. Fard, R. B. Chinnam, and C. K. Reddy, "Survival Analysis based Framework for Early Prediction of Student Dropouts," *Proc. 25th ACM Int. Conf. Inf. Knowl. Manag. - CIKM '16*, pp. 903–912, 2016.

[5] P. Kaur, M. Singh, and G. S. Josan, "Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector," *Procedia Comput. Sci.*, vol. 57, pp. 500–508, 2015.

[6] R. S. J. D. Baker and K. Yacef, "The State of Educational Data Mining in 2009 : A Review and Future Visions," *J. Educ. Data Min.*, vol. 1, no. 1, pp. 3–16, 2009.

[7] P. Thakar, "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue," 2015.

[8] A. Gopalakrishnan, R. Kased, H. Yang, M. B. Love, C. Graterol, and A. Shada, "A multifaceted data mining approach to understanding what factors lead college students to persist and graduate," *Proc. Comput. Conf. 2017*, vol. 2018–Janua, no. July, pp. 372–381, 2018.

[9] J. Liang, C. Li, and L. Zheng, "Machine learning application in MOOCs: Dropout prediction," *ICCSE 2016 - 11th Int. Conf. Comput. Sci. Educ.*, no. Iccse, pp. 52–57, 2016.

[10] E. Yukselturk *et al.*, "Predicting Dropout Student : an Application of Data Mining Methods in an Online Education Program," *Eur. J. Open, Distance e-Learning*, vol. 17, no. 1, pp. 118–133, 2014.

[11] K. Kori *et al.*, "First-year dropout in ICT studies," *IEEE Glob. Eng. Educ. Conf. EDUCON*, vol. 2015–April, no. March, pp. 437–445, 2015.

[12] D. Oreški, M. Konecki, and L. Milić, “Estimating profile of successful IT student : data mining approach,” pp. 829–833, 2017.

[13] L. Khanna, S. N. Singh, and M. Alam, “Educational data mining and its role in determining factors affecting students academic performance: A systematic review,” *India Int. Conf. Inf. Process. IICIP 2016 - Proc.*, 2017.

[14] K. Maharani, T. B. Adji, N. A. Setiawan, and I. Hidayah, “Comparison analysis of data mining methodology and student performance improvement influence factors in small data set,” *Proc. - 2015 Int. Conf. Sci. Inf. Technol. Big Data Spectr. Futur. Inf. Econ. ICSITech 2015*, pp. 169–174, 2016.

Appendix

Appendix A - Selected Source Code

Source code for dropout prediction Application

Main Class that runs the data classification program

```
/**
 * This is the main driver program that runs the data classification
program
 * by instantiating DataClassificationUI class for user to interact.
 */
public class Main {

	public static void main(String[] args) {
		DataClassificationGUI gui = new DataClassificationGUI();

		gui.setSize(900,600);
		gui.setLocationRelativeTo(null);
		gui.setVisible(true);
	}

}
```

Data Classification GUI Class

```
import java.awt.BorderLayout;
import java.awt.Color;
import java.awt.Container;
import java.awt.FlowLayout;
import java.awt.Font;
import java.awt.GridLayout;
import java.awt.event.ActionEvent;
import java.awt.event.ActionListener;
import java.io.File;
import java.io.FileNotFoundException;
import java.io.IOException;

import javax.swing.DefaultComboBoxModel;
import javax.swing.JButton;
import javax.swing.JComboBox;
import javax.swing.JFileChooser;
import javax.swing.JFrame;
import javax.swing.JLabel;
import javax.swing.JOptionPane;
import javax.swing.JPanel;
import javax.swing.JTextField;
import javax.swing.border.Border;
import javax.swing.border.LineBorder;
import javax.swing.border.TitledBorder;

import weka.classifiers.Classifier;
import weka.classifiers.trees.J48;
import weka.core.Instance;
import weka.core.Instances;
import weka.core.SparseInstance;
import weka.core.converters.CSVLoader;

/**
 * This class implements the GUI screen for user to select the data
set and then predict
 * a new student DROPOUT feature from other attributes.
 */
public class DataClassificationGUI extends JFrame implements
ActionListener {

     private JTextField txtDatasetFile;
     private JButton btnBrowse;
     private JButton btnClassifyModel;
     private JButton btnReset;
     private JButton btnExit;
     private JButton btnClassify;
     private JLabel lblStatus;
     private Border highlightBorder = new LineBorder(Color.GREEN, 2);
     private Border noBorder;

     String[] fieldNames = {"ZSCORE",
"HOUSE_INCOME","A_TYPE","GENDER","CITY","STAY_LOC",

     "F_EDU","M_EDU","FAMILY_STAT","O/L_PER","ICT_GRADE","STREEM","PRI
OR_EX","ENG_SCORE",

     "ACC_PERFOR","SAT_LEVEL","STRESS","PERCIEVED_QLTy","SOCO_INT","DR
OPOUT"};

     String[][] nominalValues = {{"1ST","2ND","3RD"},
                                    {"male","female"},
                                    {"URBAN","SUB_URBAN","RURAL"},
                                    {"home","boading"},

{"BELOWO/L","O/L","A/L","CERTIFICATE","DIPLOMA","DEGREE","POST_GRADUAT
E"},

{"BELOWO/L","O/L","A/L","CERTIFICATE","DIPLOMA","DEGREE","POST_GRADUAT
E"},

{"LOW","MIDDLE","GOOD","HIGH","EXCELLENT"},

{"EXCELLENT","VERY_GOOD","GOOD","FAIR","LOW"},
                                    {"A","B","C","S","F"},
{"ART","COMMERCE","BIO","MATHS","TECHNOLOGY"},

{"EXCELLENT","GOOD","FAIR","LOW","NO"},
                                    {"A","B","C","S","F"},

{"EXCELLENT","HIGH","GOOD","LOW","MEDIUM"},

{"EXCELLENT","GOOD","LOW","NOT_SATISFIED"},
                                    {"LOW","MEDIUM","HIGH"},
                                    {"LOW","MEDIUM","HIGH"},
                                    {"LOW","MEDIUM","HIGH"}};
     private JTextField txtZScore;
     private JTextField txtIncome;
     private JTextField txtDropout;
     JTextField[] txtAttributes;
     JComboBox[] cmbxAttributes;

     private Classifier classifier;
     private Instances data;

     public DataClassificationGUI() {
           super("Weka Data Classification (J48)");
           txtAttributes = new JTextField[fieldNames.length];
           cmbxAttributes = new JComboBox[fieldNames.length-3];
           initGUI();
           noBorder = txtDropout.getBorder();
     }

     private void initGUI() {
           Container cont = this.getContentPane();
           cont.setLayout(new BorderLayout(5,5));

           JPanel p = new JPanel();
           p.setLayout(new BorderLayout(0, 0));
           p.setBorder(new TitledBorder("Settings"));
           p.add(new JLabel("Select dataset file "),
BorderLayout.WEST);
           txtDatasetFile = new JTextField(45);
           txtDatasetFile.setFont(new Font("Arial", Font.PLAIN, 12));
           p.add(txtDatasetFile, BorderLayout.CENTER);
           cont.add(p, BorderLayout.NORTH);

           JPanel p3 = new JPanel();
           p3.setLayout(new BorderLayout(5,5));
           p3.setBorder(new TitledBorder("Classification"));

           JPanel formPanel = new JPanel();
           formPanel.setLayout(new GridLayout(fieldNames.length, 2,
10, 5));
           txtZScore = new JTextField();
           JLabel lbl = new JLabel(fieldNames[0]);
           lbl.setHorizontalAlignment(JLabel.RIGHT);
           formPanel.add(lbl);
           formPanel.add(txtZScore);
           txtIncome = new JTextField();
           lbl = new JLabel(fieldNames[1]);
           lbl.setHorizontalAlignment(JLabel.RIGHT);
           formPanel.add(lbl);
           formPanel.add(txtIncome);
           for(int i = 0; i < cmbxAttributes.length; i++) {
                lbl = new JLabel(fieldNames[i+2]);
                lbl.setHorizontalAlignment(JLabel.RIGHT);
                formPanel.add(lbl);
                //txtAttributes[i] = new JTextField();
                cmbxAttributes[i] = new JComboBox();
                DefaultComboBoxModel model =
(DefaultComboBoxModel)cmbxAttributes[i].getModel();
                for(String option : nominalValues[i]) {
                     model.addElement(option);
                }
                formPanel.add(cmbxAttributes[i]);
           }
           lbl = new JLabel(fieldNames[fieldNames.length-1]);
           lbl.setHorizontalAlignment(JLabel.RIGHT);
           formPanel.add(lbl);
           txtDropout = new JTextField();
           txtDropout.setEnabled(false);
           formPanel.add(txtDropout);
           p3.add(formPanel, BorderLayout.CENTER);

           cont.add(p3, BorderLayout.CENTER);

           JPanel p4 = new JPanel();
           p4.setLayout(new FlowLayout());
           btnReset = new JButton("Reset Data");
           btnReset.addActionListener(this);

           lblStatus = new JLabel("Data model not classified yet.");
           p4.add(lblStatus);
           p4.add(btnReset);

           btnClassify = new JButton("Classify Data");
           btnClassify.addActionListener(this);
           btnClassify.setEnabled(false);
           p4.add(btnClassify);

           btnExit = new JButton("Exit");
           btnExit.addActionListener(this);
           p4.add(btnExit);

           cont.add(p4, BorderLayout.SOUTH);

           JPanel p2 = new JPanel();
           p2.setLayout(new FlowLayout());
           btnBrowse = new JButton("Browse...");
           btnBrowse.addActionListener(this);
           p2.add(btnBrowse);
           btnClassifyModel = new JButton("Classify Model (J48)");
           btnClassifyModel.addActionListener(this);
```

```
            btnClassifyModel.setEnabled(false);
            p2.add(btnClassifyModel);
            p.add(p2, BorderLayout.EAST);

            setDefaultCloseOperation(EXIT_ON_CLOSE);
            pack();
        }

        @Override
        public void actionPerformed(ActionEvent e) {
            Object src = e.getSource();
            if(src == btnBrowse) {
                selectDatasetFile();
            }
            else if(src == btnClassifyModel) {
                classifyModelForDataset();
            }
            else if(src == btnClassify) {
                classifyData();
            }
            else if(src == btnExit) {
                System.exit(0);
            }
            else if(src == btnReset) {
                resetForm();
            }
        }

        // Reset the new data form
        private void resetForm() {
            for(int i=0; i < fieldNames.length -1 ; i++) {
                txtAttributes[i].setText("");
            }
            txtAttributes[txtAttributes.length-1].setBorder(noBorder);
            txtAttributes[txtAttributes.length-1].setText("");
        }


        // Classify the new data
        private void classifyData() {
            try {
                Instance instance = new
    SparseInstance(fieldNames.length);
                instance.setDataset(data);
                instance.setValue(0,
    Double.parseDouble(txtZScore.getText()));
                instance.setValue(1,
    Double.parseDouble(txtIncome.getText()));
                for(int i = 0; i < cmbxAttributes.length; i++) {
                    String value =
    (String)cmbxAttributes[i].getSelectedItem();
                    System.out.println(i + " " + value);
                    instance.setValue(i+2, value);
                }
                double targetValue =
    classifier.classifyInstance(instance);
                String label =
    instance.dataset().classAttribute().value((int) targetValue);
                //txtAttributes[txtAttributes.length-
    1].setText(label);
                //txtAttributes[txtAttributes.length-
    1].setBorder(highlightBorder);
                txtDropout.setText(label);
                txtDropout.setBorder(highlightBorder);
            } catch (Exception e) {
                JOptionPane.showMessageDialog(this, "ERROR: " +
    e.getMessage(), "Error", JOptionPane.ERROR_MESSAGE);
                e.printStackTrace();
            }
        }

        // Allow user to browse and select data set file from local file
    system
        private void selectDatasetFile() {
            JFileChooser fch = new JFileChooser(".");
            if(fch.showOpenDialog(this) == JFileChooser.APPROVE_OPTION)
    {
                File f = fch.getSelectedFile();
                txtDatasetFile.setText(f.getAbsolutePath());
                btnClassifyModel.setEnabled(true);
            }
        }

        // Build Data Classification model using J48 algorithm from
    dataset
        private void classifyModelForDataset() {

            File f = new File(txtDatasetFile.getText());
            try {
                CSVLoader csvLoader = new CSVLoader();
                csvLoader.setSource(f);
                data = csvLoader.getDataSet();
                data.setClassIndex(fieldNames.length-1);
                classifier = new J48();
                classifier.buildClassifier(data);
                btnClassify.setEnabled(true);
                lblStatus.setText("Classifier created from dataset.");
            } catch (FileNotFoundException e) {
                JOptionPane.showMessageDialog(this, "ERROR: " +
    e.getMessage(), "Error", JOptionPane.ERROR_MESSAGE);
            } catch (IOException e) {
                JOptionPane.showMessageDialog(this, "ERROR: " +
    e.getMessage(), "Error", JOptionPane.ERROR_MESSAGE);
            } catch (Exception e) {
                JOptionPane.showMessageDialog(this, "ERROR: " +
    e.getMessage(), "Error", JOptionPane.ERROR_MESSAGE);
            }
        }
    }
```