2010

Longitudinal study of first-time freshmen using data mining

Ashutosh R. Nandeshwar
West Virginia University

LONGITUDINAL STUDY OF FIRST-TIME FRESHMEN USING DATA MINING

Ashutosh R. Nandeshwar

Dissertation submitted to the
College of Engineering and Mineral Resources
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in
Industrial Engineering

Majid Jaraiedi, Ph.D., Chair
Tim Menzies, Ph.D., Co-Chair
Robert C. Creese, Ph.D.
B. Goplakrishnan, Ph.D.
Mike Sperko

Department of Industrial and Management Systems Engineering

Morgantown, West Virginia

2010

KEYWORDS: data mining, student retention, predictive modeling, higher education

Abstract

LONGITUDINAL STUDY OF FIRST-TIME FRESHMEN USING DATA MINING

Ashutosh R. Nandeshwar

In the modern world, higher education is transitioning from enrollment mode to recruitment mode. This shift paved the way for institutional research and policy making from historical data perspective. More and more universities in the U.S. are implementing and using enterprise resource planning (ERP) systems, which collect vast amounts of data. Although few researchers have used data mining for performance, graduation rates, and persistence prediction, research is sparse in this area, and it lacks the rigorous development and evaluation of data mining models. The primary objective of this research was to build and analyze data mining models using historical data to find out patterns and rules that classified students who were likely to drop-out and students who were likely to persist.

Student retention is a major problem for higher education institutions, and predictive models developed using traditional quantitative methods do not produce results with high accuracy, because of massive amounts of data, correlation between attributes, missing values, and non-linearity of variables; however, data mining techniques work well with these conditions. In this study, various data mining models were used along with discretization, feature subset selection, and cross-validation; the results were not only analyzed using the probability of detection and probability of false alarm, but were also analyzed using variances obtained in these performance measures. Attributes were grouped together based on the current hypotheses in the literature. Using the results of feature subset selectors and treatment learners, attributes that contributed the most toward a student's decision of dropping out or staying were found, and specific rules were found that characterized a successful student. The performance measures obtained in this study were significantly better than previously reported in the literature.

DEDICATION

Dr. B. R. Ambedkar—the greatest social reformer of India
(1891–1956)
Only his unyielding determination, heroic efforts, and fearless fights could have led to the emancipation of the "untouchables" of India.

and

Mrs. Janabai Janbodhkar—my *Aaji*, meaning grandmother
(–2004)
Her integrity, honesty, wisdom, and independence have left an unforgettable impression on my mind.

Acknowledgments

There are many people who have helped me directly or indirectly, and I sincerely appreciate everyone's assistance.

First of all, many thanks to Dr. Jaraiedi for being such a fantastic advisor. Your policy of responding to e-mails within 24 hours worked extremely well for me, as I was away from the campus and I was still able to communicate with you without any problems. I appreciate all the feedback that you have given on this work.

Dr. Tim Menzies, thanks for everything. Not only you introduced me to LaTeX, AWK scripting, Tufte (along with the evils of powerpoint) and the importance of simplicity, but you also shaped this project to produce actionable theories. Your cheerfulness, enthusiasm, expertise, and of course, your Australian accent have helped me keep my sanity and finish this dissertation.

I would also like to thank Dr. Gopala and Dr. Creese for their valuable suggestions to make this work even better. I owe gratitude to Mike Sperko as well, for his tremendous support throughout my dissertation.

I am thankful to Adam Nelson for his genius shell and awk scripts to run the experiments repeatedly, and for his excellent teamwork. Thanks also go to wonderful and helpful IMSE staff.

I would like to thank all my friends from WVU who enriched my graduate student life in many ways. I do not want to reduce their help to one liners, but I want to acknowledge them in a short space, so here it goes: Subodh, you are the best buddy one could have; thanks for all your help. Devdatta, thanks for being there. Masoud, thanks for your wise remarks. Prasad, thanks for being an excellent roommate. Deepak G., thanks for your continuous encouragement and help. Also, thanks to all my buddies, who made my initial days (and years) in a new country very enjoyable: Tripathi, Veenu, Ram G., Srinivas V., Sujan, Michael J., Chakri, Dhruv, Amol, Archis, Girish, Rahul, Santosh, and Nitin. Special thanks to Dr. Chandran for helping me at the time when I needed the help most.

Many thanks to the wonderful colleagues at Kent State University for pro-

viding valuable feedback and sugary treats.

Of course, my sincerest appreciation goes to my parents and family members. Baba (Dad), thanks for encouraging me to prepare for the GRE again and apply to other US universities, although you were still recovering from your major accident and still had trouble forming correct words and meaningful sentences. Aai (Mom), thanks for your unconditional support and constant push to make the most of my life. I simply could not come this far without you two. Thanks to my brothers and sisters-in-law for your great help and guidance. Huge thanks to Utpalvarna, my dear wife, for believing in me and pushing me to achieve certain goals, and for your countless hours doing all the work without my help. Thanks to my kids, Asanga and Dinnaga, for teaching me new meaning of life and making my days and nights bright with your smiles and calls of "Da-Da."

Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

Abbreviation	Description	Definition
AGI	Adjusted Gross Income	page 66
ANN	Artificial Neural Networks	page 13
CRISP-DM	CRoss Industry Standard Process for Data Mining	page 6
CART	Classification and Regression Tree	page 15
EFC	Expected Family Contribution	page 44
ERP	Enterprise Resource Planning	page 4
ETL	Extract, Transform, and Load	page 4
FAFSA	Free Application for Federal Student Aid	page 89
GMDH	Group Method of Data Handling	page 47
HSGPA	High School GPA	page 44
PD	Probability of Detection	page 53
PF	Probability of False Alarm	page 53
RET1	Student retained after first-year	page 64
RET2	Student retained after second-year	page 64
RET3	Student retained after third-year	page 64
TAR3	Tarzan Treatment Learner	page 90

Chapter 1

Introduction and Research Objective

> There is nothing like looking, if you want to find something. You certainly usually find something, if you look, but it is not always quite the something you were after.
>
> J.R.R. Tolkien

1.1 Introduction

Following World War II, a great need for higher education institutions arose in the United States, and the higher education leaders built institutions on "build it and they will come" basis. After the World War II, enrollment in the public as well as the private institutions soared (Greenberg, 2004); however, this changed by 1990s, due to a significant drop in enrollment, universities were in a marketplace with "hypercompetition," and institutions faced the unfamiliar problem of receiving less applicants than they were used to receive (Klein, 2001).

Today higher education institutions are facing the problem of student retention, which is related to graduation rates; colleges with higher freshmen retention rate tend to have higher graduation rates within four years. The average national retention rate is close to 55% and in some colleges fewer than 20% of incoming student cohort graduate (Druzdzel and Glymour, 1994), and approximately 50% of students entering in an engineering program leave before graduation (Scalise

et al., 2000). Tinto (1982) reported national dropout rates and BA degree completions rates for the past 100 years to be constant at 45 and 52 percent respectively with the exception of the World War II period (see Figure 1.1 for the completion rates from 1880 to 1980). Tillman and Burns at Valdosta State University (VSU) projected lost revenues per 10 students, who do not persist their first semester, to be $326,811. Although gap between private institutions and public institutions in terms of first-year students returning to second year is closing, the retention rates have been constant for a long period for both types of institutions(ACT, 2007, see Figure 1.2).The National Center for Public Policy and Higher Education (NCPPHE) reported the U.S. average retention rate for the year 2002 to be 73.6% (NCPPHE, 2007). This problem is not only limited to the U.S. institutions, but also for the institutions in many countries such as U.K and Belgium. The U.K. national average freshmen retention for the year 1996 was 75% (Lau, 2003), and Vandamme (2007) found that 60% of the first generation first-year students in Belgium fail or dropout.

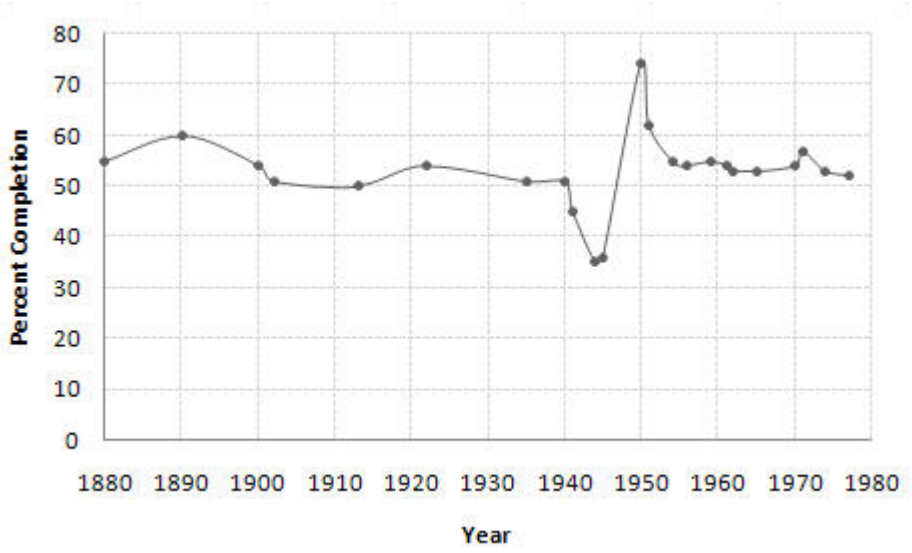
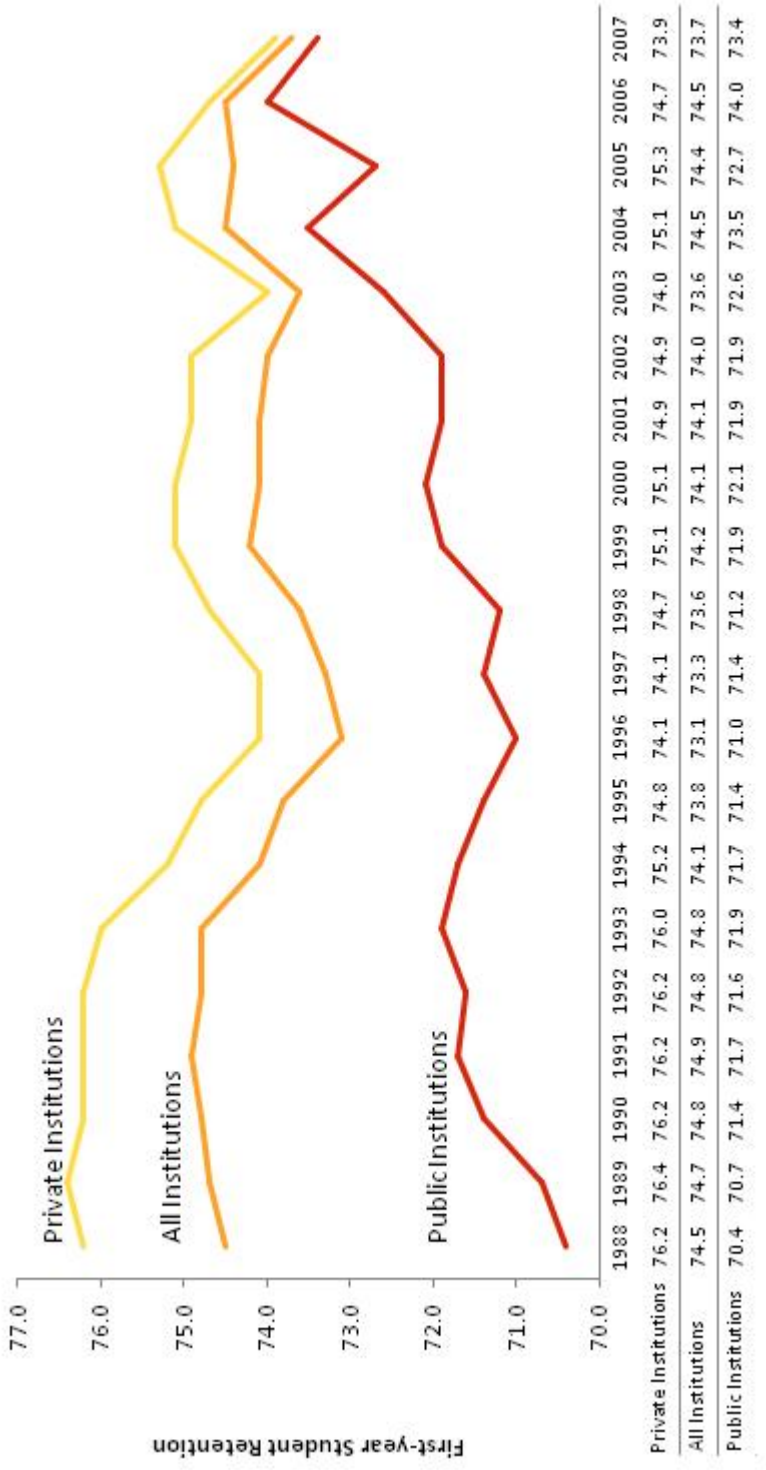


Figure 1.1: BA Degree Completion Rates for the period 1880 to 1980, where Percent Completion is the Number of BAs Divided by the Number of First-time Degree Enrollment Four Years Earlier (Tinto, 1982)

Theoretical models of student departure, such as, Tinto's student dropout model (Tinto, 1975), described the conceptual stages of a dropout from a college, which studied interaction between an individual and the academic and social system of the college. While the researchers widely accept this model and the model explains the problem, it is difficult to implement this model using universities' data warehouses. In addition, data warehouses cannot capture the social aspect of a student's experience at a college or university.



	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Private Institutions	76.2	76.4	76.2	76.2	76.2	76.0	75.2	74.8	74.1	74.1	74.7	75.1	75.1	74.9	74.9	74.0	75.1	75.3	74.7	73.9
All Institutions	74.5	74.7	74.8	74.9	74.8	74.8	74.1	73.8	73.1	73.3	73.6	74.2	74.1	74.1	74.0	73.6	74.5	74.4	74.5	73.7
Public Institutions	70.4	70.7	71.4	71.7	71.6	71.9	71.7	71.4	71.0	71.4	71.2	71.9	72.1	71.9	71.9	72.6	73.5	72.7	74.0	73.4

Figure 1.2: Percentage of First-Year Students at Four-Year Colleges Who Return for Second Year (ACT, 2007)

Predictive modeling of student persistence using traditional methods, such as, linear and logistic regression, fail to produce results with high accuracy, and are prone to the problems of linearity, correlation of attributes, missing data, and vastness of data.

Universities' enterprise resource planning (ERP) systems collect vast amounts of data. Typically, these data consist of demographical, financial, and academic information; later, these data reside in some form of data warehouses. However, this massive data storage, often, does not transform into knowledge or information to enable administrative decision-making. This abundance of data makes the predictive modeling of high-risk students using data mining a perfect case. In addition, data mining techniques are robust and work well with missing or correlated data. As the business world benefited tremendously by data mining, and data mining supported marketing campaigns and quality assurance (Luan and Serban, 2002), it presents an opportunity to the higher education institutions to employ the same techniques to solve some of the major problems faced by the higher education administrators today.

1.2 Data Mining

1.2.1 What is Data Mining?

Although data mining definitions change with the area of the researcher, the definitions by some of the well-known researchers are apt for this research. Hand et al. (2001) defined data mining as "the science of extracting useful information from large data sets or databases." Witten and Frank (2005) defined data mining as "the process of discovering patterns in data. The process must be automatic or (more usually) semiautomatic." Berry and Linoff (1997) defined data mining as "the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules."

Data mining is also known as knowledge discovery in databases (KDD), and this discovery process is shown in Figure 1.3. Enterprise Resource Planning (ERP) systems hold massive amounts of data, which usually consists of information, such as, demographic, financial, payroll, others. The data entry people working in each functional area enter this information in ERP systems. Database administrators load this information in databases using Extract, Transform, and Load (ETL) tools. Data analysts or miners analyze these databases, understand the data or work with the domain experts, develop prediction, classification, or clustering models, evaluate the models, and implement them; using this approach, data miners transform information into tangible knowledge for decision-making.

Areas of computer science, statistics, database technologies, machine learning, and others form the field of data mining. Statistics influenced the field of data mining tremendously; so much that Kuonen (2004) asked whether data mining

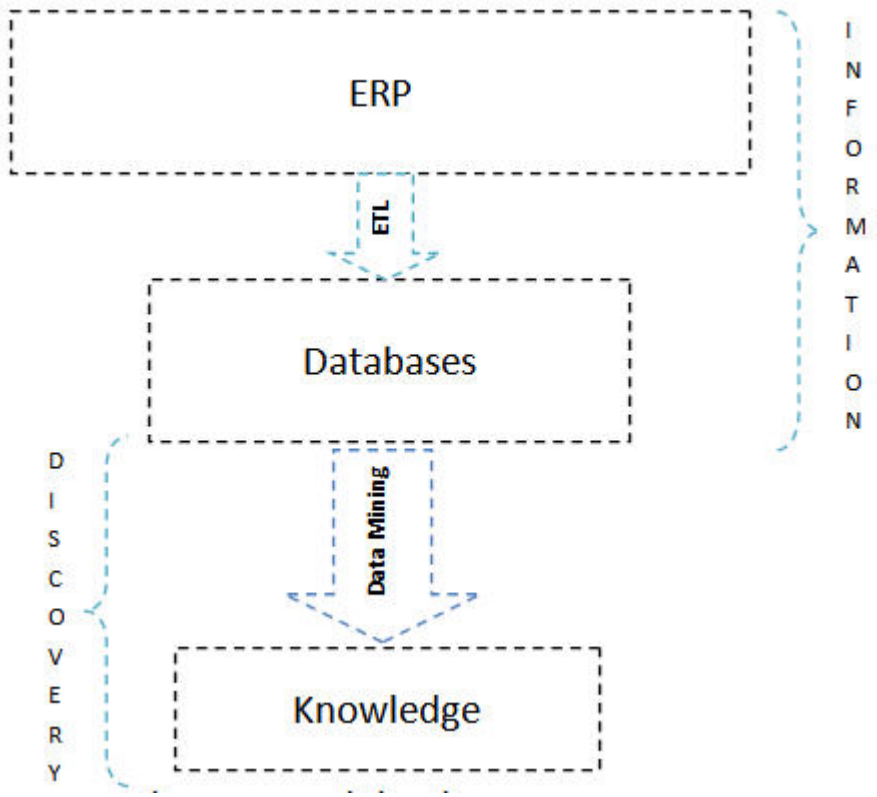


Figure 1.3: Data to Knowledge

is "statistical déjà vu." Amalgamation of statistics and computer science started data mining; however, data mining as a field is evolving on its own. Han and Kamber (2006) described the overlap of multiple disciplines as shown in Figure 1.4.

> Facts are cheap, information is plentiful - knowledge is precious.
>
> Fortune cookie saying

1.2.2 Data Mining Methodology

Data mining is a non-linear process of data selection and cleaning, data transformation, pattern, and model evaluation. To refine the model, data miners usually apply the output of a step as an input to any other step. Han and Kamber (2006) illustrated this non-linear process as shown in Figure 1.5. Although the progression from databases to knowledge in Figure 1.5 seems to be linear, the dotted and thick arrows show the process flow from any node to another node.

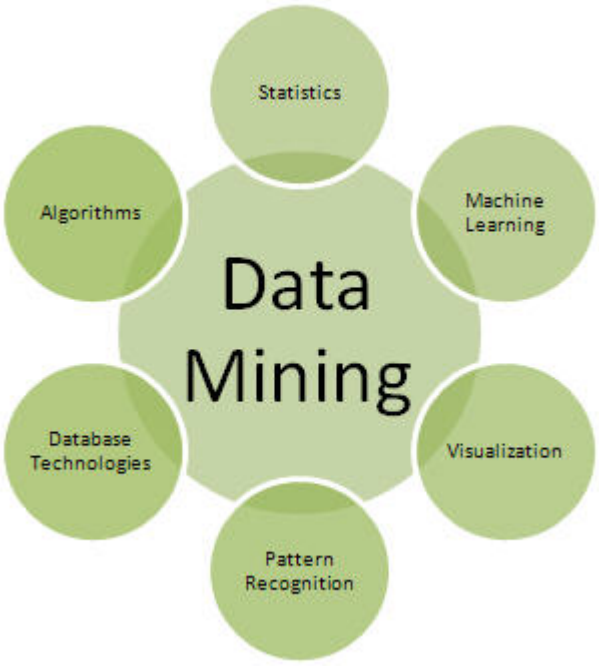


Figure 1.4: Data Mining-Confluence of Multiple Disciplines

1.2.2.1 CRoss Industry Standard Process for Data Mining (CRISP-DM)

DaimlerChrysler (then Daimler-Benz), SPSS (then ISL), and NCR, in 1996, worked together to form the CRoss Industry Standard Process for Data Mining (CRISP-DM). Their philosophy behind creating this standard was to form non-propriety, freely available, and application-neutral standards for data mining. Figure 1.6 shows CRISP-DM version 1.0, and it illustrates the non-linear (cyclic) nature of data mining. Standard's phases include, business understanding, data understanding, data preparation, modeling, evaluation, and deployment.

Business Understanding: Business understanding is the initial phase of data mining process, where the business group defines project objectives, and the data miner transforms these objectives into data mining definitions. In addition to the project objectives, a preliminary plan is designed in this phase to achieve these objectives. Berry and Linoff (1997) advised data miners to break general goals into more specific ones, and to achieve that business knowledge is very important. Identifying the input and target variables using the business objectives is a key process in this phase. Correct understanding of the business objectives is imperative in this process, for example, a person who is likely to make late payments can be a "good customer" for a credit card company, and

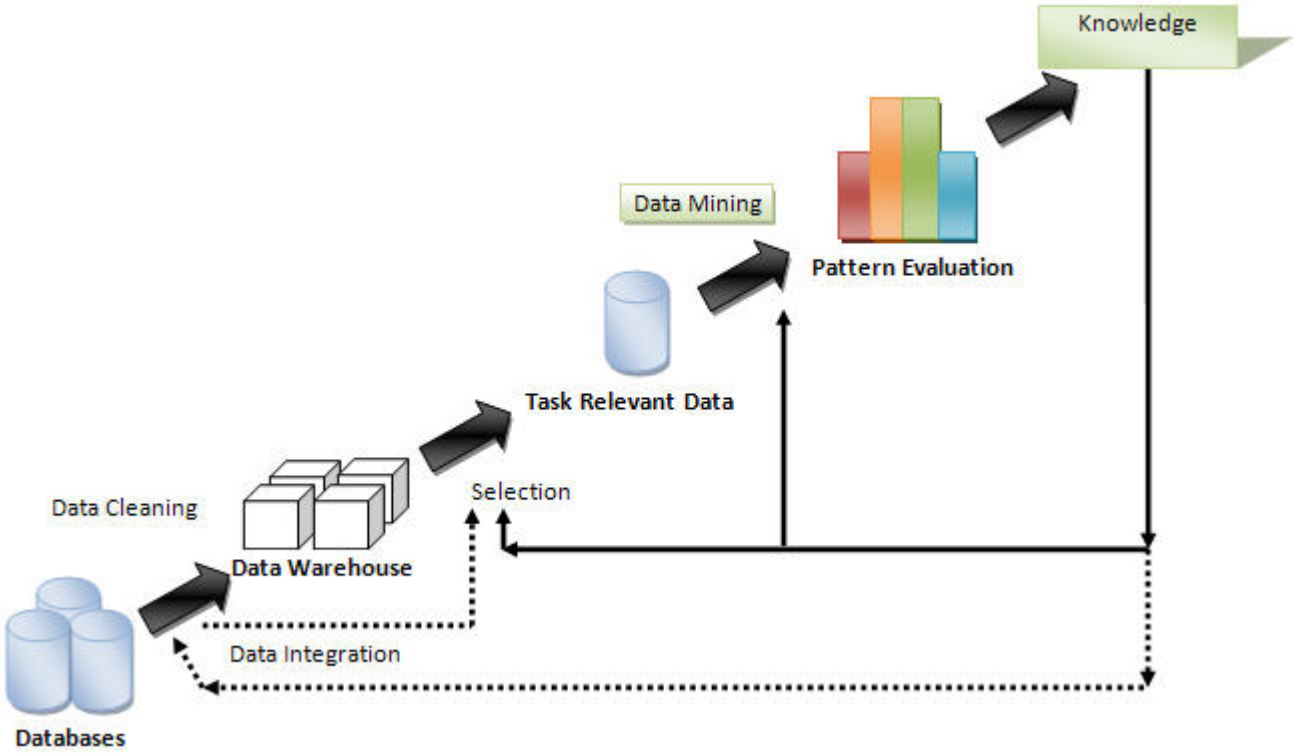


Figure 1.5: Knowledge Discovery Process

unless the data miners have this knowledge they will not be able to transform this to data mining objectives.

Data Understanding: Data understanding is the phase where the domain expertise is very important, and it is a part of the business understanding. Initial exploring of data, identifying data quality problems, and discovering insights into the data are the phases of data understanding. Data and business understanding are very critical to the data mining process, as some of the attributes in the data might appear trivial to the data miners, where, in reality, those attributes might be significant. Although domain expertise is imperative for data mining, it can create hindrances while selecting attributes, as data mining algorithms might find some patterns in the excluded attributes; and the cyclic nature of data mining lies here.

Examining distributions, relation of attributes, and descriptive statistics are the basic steps of data understanding phase. Examining relation of attributes is useful for generating derived variables. Examining distributions and descriptive statistics is useful for finding disparities and irregularities in the data.

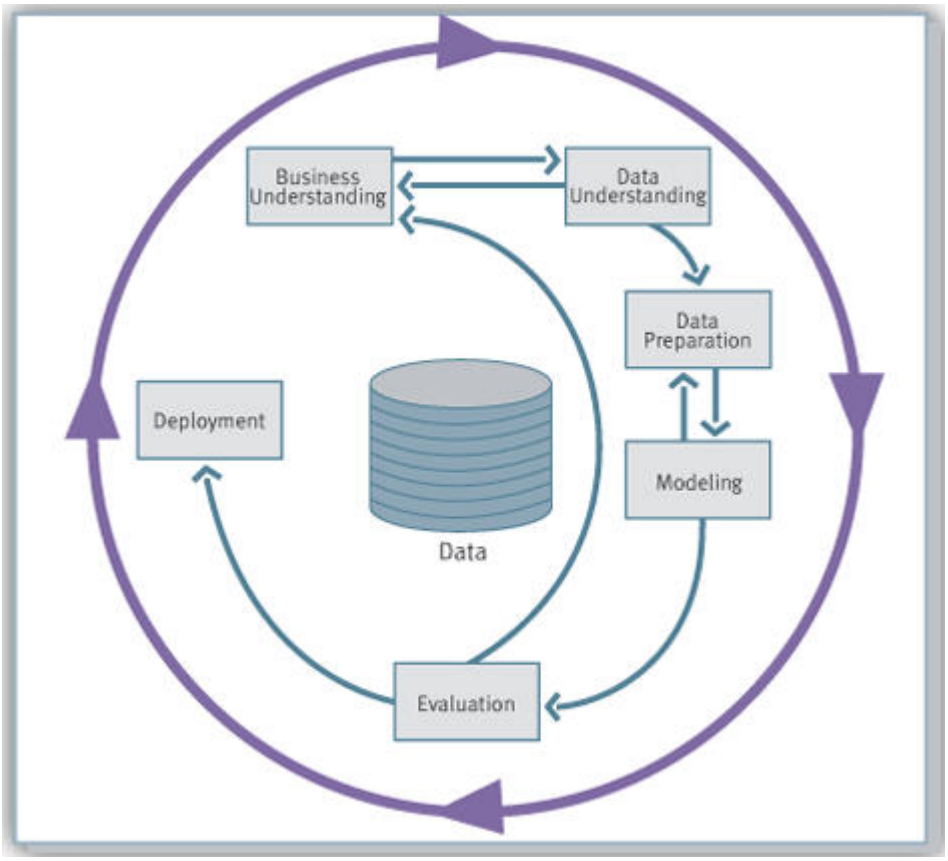


Figure 1.6: CRISP-DM Model Version 1.0

Data Preparation: Data preparation is the most labor-intensive process of data mining. This phase includes preparation of the raw data to a final dataset for modeling; it involves initial attribute selection and transformation using the data and the business understanding. To prepare a final dataset, treatment of dirty data and missing values is critical using manual or automatic processes. Some of the data mining algorithms, such as, naïve bayes, handle missing data very well; however, replacing missing values with the mean, or modeling the data to predict the missing values are common and good practices.

Modeling: This is the core process of data mining, where models transform input into output; Berry and Linoff (1997) illustrated this process as shown in Figure 1.7. There are several data mining techniques for the same problem, and the evaluation phase is useful to selecting the best model. The best model, sometimes, might not be best in performance, but simplest in explanation. Brinkman and McIntyre (1997) cautioned on generating complicated models, "policymakers may not have confidence in a forecast if they do not understand its conceptual

basis or accept its assumptions", or as the famous Occam's Razor describes:

Entia non sunt multiplicanda praeter necessitatem
Or
Entities should not be multiplied more than necessary

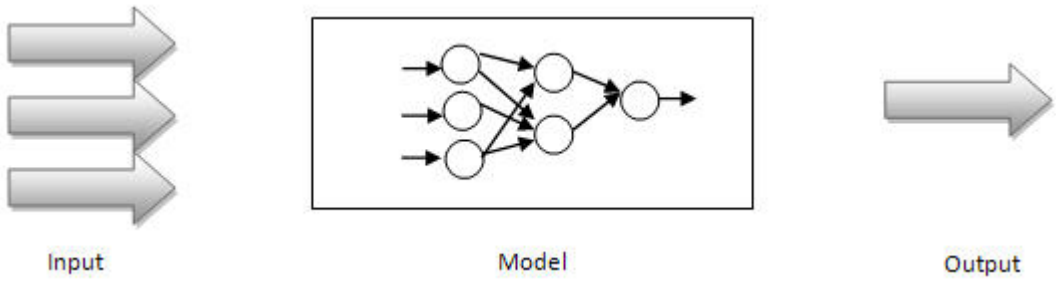


Figure 1.7: Modeling Process

Menzies (2006) illustrated the explanation and performance systems as shown in Figure 1.8. As the name suggests, the explanation systems offer explanation on how a conclusion was reached; performance systems produce results with high accuracy, but offer no explanation. Menzies et al. (2007a) explained the trade-off between efficiency and explanation of the models: "sometimes the explanatory power must be decreased in order to increase the efficacy of the predictor." They offered ensemble techniques as a solution to explain a model while producing high precision results. These techniques included discretization, cross-validation, and feature subset selection (FSS).

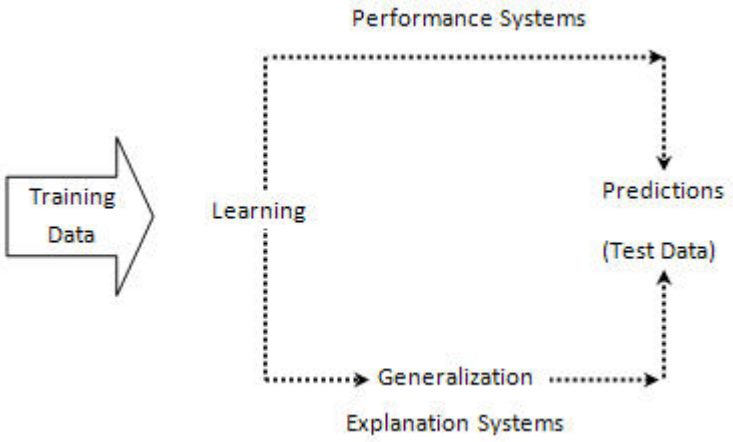


Figure 1.8: Performance vs. Explanation Systems

		Predicted	
		Yes	No
Actual	Yes	True Positive (TP)	False Negative (FN)
	No	False Positive (FP)	True Negative (TN)

Table 1.1: Possible Outcomes of a Two-class Prediction

Evaluation: Before deployment of the model, this phase evaluates the model for quality and effectiveness. This phase also evaluates the closeness of the model from the business objective, and checks whether all important business matters are considered or not. Evaluating the results of the model also determine the use of the data mining model for deployment. Some of the tools to evaluate models are confusion matrix, lift chart, and minimum description length (MDL); later sections provide explanation on these tools. Researchers use the confusion matrix, given in Table 1.1, to evaluate different models; some of the evaluation criteria are: recall, precision, accuracy or overall correct classification rate, and F-score or harmonic mean of precision and recall.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{1.1}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1.2}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1.3}$$

$$F\text{-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1.4}$$

Deployment: After the evaluation is complete, the model is ready for deployment; however, the project does not end here, analysts generate reports to present the information that users can easily understand, or set up similar models for different units.

1.2.3 Data Mining Terminology

1.2.3.1 Records or Instances

Records are the number of rows present in a file, which is to be analyzed by the use of data mining. Records can be sequential or random depending on the algorithm used for data mining. For a typical data mining task, required number of records is usually high.

1.2.3.2 Fields, Attributes, Features, or Variables

As many fields influenced data mining, finding different names for a single entity is inevitable. All of these are common names of the columnar data in a file.

1.2.3.3 Data or Dataset

Data or dataset are a collection of records across different fields. Researchers, to represent the files, loosely use the term data.

1.2.3.4 Learners or Techniques

The tools used for data mining modeling are learners or techniques. These learners differ by the type of output they produce, such as, prediction, classification, clusters, and associations.

1.2.3.5 Input Variables

The variables or attributes used for modeling in order to produce an output.

1.2.3.6 Output or Target Variables

The attributes on which the modeling techniques learn are output or target variables; however, some data mining techniques, such as, clustering and association, learn without target variables, instead, only the input variables are used to produce generic rules of the existing patterns in the dataset.

1.2.3.7 Training, Validation, and Test Data Set

Usually, application of a data mining technique involves creation of three partitions of the available dataset. Models are built on the training dataset, the models are compared or fine-tuned on the validation dataset, and the performance of the models on unseen data is checked on the test dataset.

An example data file on weather and the decision to play golf is shown in Figure 1.9. This file has 14 records and five variables. In this example, the fields outlook, temperature, humidity, and windy are input variables, and the field play is an output variable.

1.2.4 Data Mining Modeling Techniques

There are different types of modeling techniques for different types of tasks, and there are different types of modeling techniques for a single problem. Table 1.2 is a list of some of the data mining techniques by the task type.

Row ID	Outlook	Temperature	Humidity	Windy	Play
1	sunny	85	85	FALSE	no
2	sunny	80	90	TRUE	no
3	overcast	83	86	FALSE	yes
4	rainy	70	96		
	rainy	75	80	FALSE	yes
11	sunny	75	70	TRUE	yes
12	overcast	72	90	TRUE	yes
13	overcast	81	75	FALSE	yes
14	rainy	71	91	TRUE	no

Figure 1.9: Weather Data

Data Mining Tasks	Data Mining Techniques
Classification or Prediction	**Function Based**
	Linear Regression
	Logistic Regression
	Neural networks
	Tree Based
	CART
	J48
	M5′
	Rule Based
	OneR
	JRip
	PART
	Other
	Naive Bayes
Clustering	K-means
Association	Apriori

Table 1.2: Data Mining Techniques by Task

1.2.4.1 Classifiers

Linear Regression: Statistics heavily use linear regression, and it works the best when all the variables are numeric, the data are non-linear in nature, and there are no missing values. The general linear regression model (Neter et al., 1989), with normal error terms, is given in Equation 1.5.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1} + \epsilon \tag{1.5}$$

where, $\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1}$ are parameters, $X_1, X_2, \ldots, X_{p-1}$ are input variables, and ϵ are independent and identically normally distributed error terms with mean

$= 0$ and variance σ_ϵ^2.

The general linear regression model given in Equation 1.5 is represented in vector-matrix form in Equation 1.6, and in matrix terms, the general linear model is given in Equation 1.7. The parameters, $\beta_0, \beta_1, \ldots, \beta_p - 1$, are estimated by using Equation 1.8.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np-1} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{1.6}$$

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{1.7}$$

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{1.8}$$

where, n is the total number of observations, and $\widehat{\beta}$ are the estimated parameters.

Logistic Regression: Logistic regression is best suitable for modeling when the output variable is dichotomous, which can take the value of probability of success equal to one (q) and probability of failure to zero $(1-q)$. The probability of the dependent variable $(\mathbf{Y})$ or probability of success, given the probability of the input variables (x), is given in Equation 1.9.

$$P\{\mathbf{Y} = 1|x\} = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}} \tag{1.9}$$

where, $\beta_0, \beta_1, \beta_2, \ldots, \beta_{p-1}$ are parameters, and $x_1, x_2, \ldots, x_{p-1}$ are input variables

A simplified model using θ is given in Equation 1.10, and the logistic model is given in Equation 1.11. The regression parameters are estimated using maximum-likelihood.

$$P\{\mathbf{Y} = 1|x\} = \theta \tag{1.10}$$

where, $\theta = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}$

$$\text{logit}\theta = \log\frac{\theta}{1-\theta} = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1}x_{p-1} \tag{1.11}$$

Neural Networks: Although the working mechanism of the human brain influenced artificial Neural Networks (ANN) or Multi-Layer Perceptron (MLP) models, these models are very similar to linear regression models. A collection of neurons or nodes is a layer, and there are many layers in an ANN; each neuron in a layer is fully connected to all other neurons in the following layer. The first layer receives the input, hence called an input layer. The output of the last layer is the

output of the network. Hidden layers are the layers between the input and output layers. It is a common practice to use either one or two hidden layers. Figure 1.10 is a representation of a feed forward network with configuration as one input layer with three inputs, one hidden layer with two nodes, and one output layer with single output abbreviated as 3-2-1 network (Nandeshwar, 2006). The input

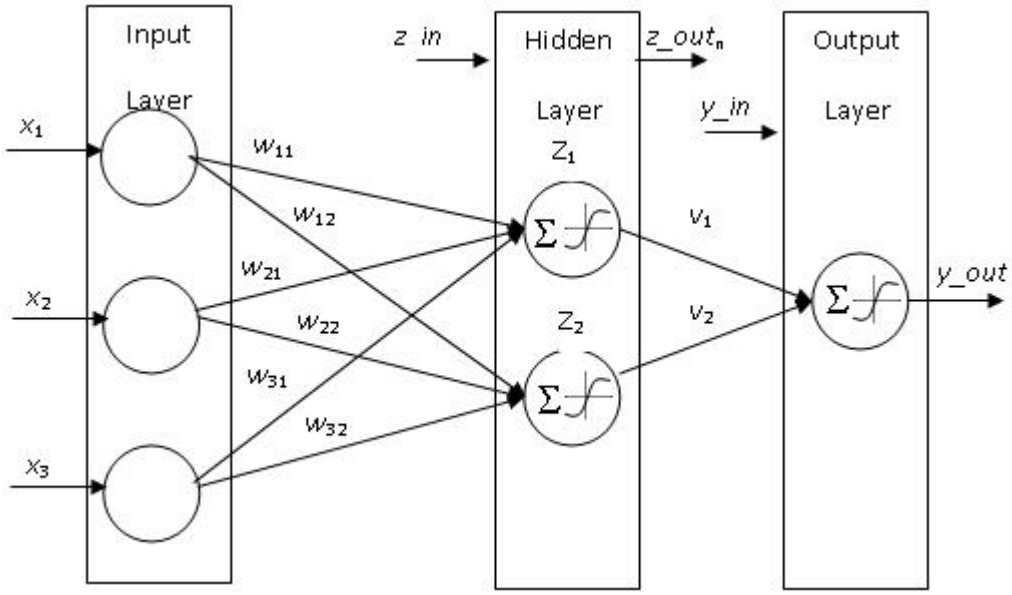


Figure 1.10: Feed-forward Network with 3-2-1 Architecture

layer receives signals as $X_1, X_2,$ and X_3. Initially, random or fixed weights are assigned to the connections between all the neurons in all the layers, which are denoted by matrix **W** and **V**. Matrix **W** denotes the weights between the input layer and the hidden layer, and matrix **V** denotes the weights between the hidden layer and the output layer. The summation of the multiplication of the inputs of a layer with the weights of a layer is the input of the next layer. Matrix **W** is multiplied with the input signals and then summed up in the hidden layer. An activation function, given in Equation 1.12, is applied to this summation to give new input signals for the next layer. The most popular activation function is the sigmoid function or the logistic function, given by Equation 1.13 and illustrated by Figure 1.11.

$$y = f(x) \tag{1.12}$$

where, y = output of the function, $f()$ = linear, identity, or non-linear function, and x = input to the function.

$$f(x) = \frac{1}{1 + e^{-x}} \tag{1.13}$$

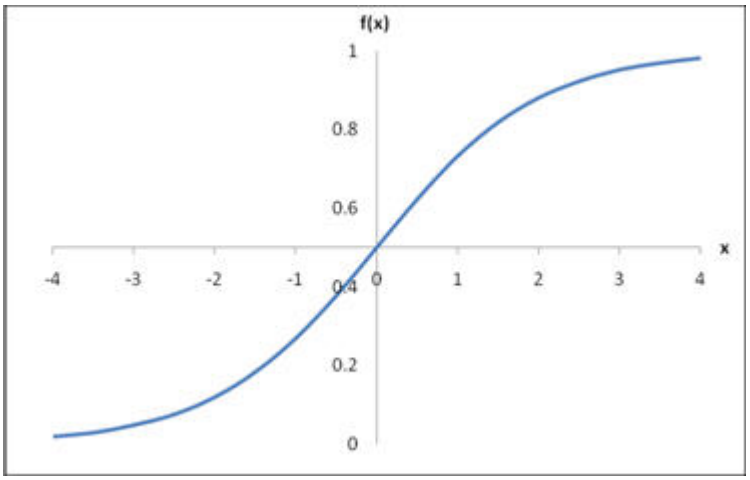


Figure 1.11: Sigmoid or Logistic Activation Function

The output of the activation function is again fed-forward and multiplied by the weights between hidden and output layer, i.e. matrix V. This multiplied signal is again sent through the activation function, Equation 1.12, to give the output or result of the network. Fausett (1994) provided mechanics of a feed-forward neural network with one hidden layer as shown in Figure 1.12.

Backpropagation algorithm is the most common learning or training algorithm of ANNs. Artificial neural network learn by example, and backpropagation algorithm "trains" the neural network by looping through the data and constantly updating the weights to minimize the difference between the actual and the predicted data. Training is stopped when the maximum number of iterations or epochs, iterations in machine learning language, or acceptable difference between the actual and the predicted data is reached.

Decision Trees: Decisions tree are a collection of nodes, branches, and leaves. Each node represents an attribute; this node is then split into branches and leaves. Decision trees work on the "divide and conquer" approach; each node is divided, using purity information criteria, until the data are classified to meet a stopping condition. Gini index and information gain ratio are two common purity measurement criteria; Classification and Regression Tree (CART) algorithm uses Gini index, and C4.5 algorithm uses the information gain ratio (Quinlan, 1986a, 1996). The Gini index is given by Equation 1.14, and the information gain is given by Equation 1.15.

$$I_G(i) = 1 - \sum_{j=1}^{m} f(i,j)^2 = \sum_{j \neq k} f(i,j) f(i,k) \tag{1.14}$$

n	Number of input units.
nh	Number of hidden layer neurons.
x_i	Activations of units X_i: For input units X_i, x_i = input signal;
w_{ij}	Weights between the input layer and the hidden layer.
v_{jk}	Weights between the hidden layer and the output layer.
z_in_j	Input to the hidden layer: $z\_in_j = \sum_{i=1}^{n} w_{ij} x_i$
z_out_j	Output of the hidden neurons: $z\_out_j = f(z\_in_j)$
y_in_k	Input to the output layer: $y\_in_k = \sum_{j=1}^{nh} v_{jk} z\_out_j$ Note: If there is only one output unit then subscript k is removed
y_out_k	Output of the network: $y\_out_k = f(y\_in_k)$

Figure 1.12: Feed-forward Network with one Hidden Layer

$$I_E(i) = -\sum_{j=1}^{m} f(i,j) \log_2 f(i,j) \tag{1.15}$$

where, m is the number of values an attribute can take, and $f(i,j)$ is the proportion of class in i that belong to the j^th class.

Figure 1.13 is an example of construction decision tree using the Titanic data and the JMP software. Based on the impurity, JMP selected the attribute sex (male and female) as the root node, then for attribute value sex = female, JMP created one more split on class (first, second, third, and crew). In order to reduce the impurity, JMP created a split on the root node of sex =male for the attribute age (child and adult).

Rules: Construction of rules is quite similar to the construction of decision trees; however, rules first cover all the instances for each class, and exclude the instances, which do not have class in it. Therefore, these algorithms are called as covering algorithms, and pseudocode of such algorithm is given in Figure 1.14 reproduced from Witten and Frank (2005).

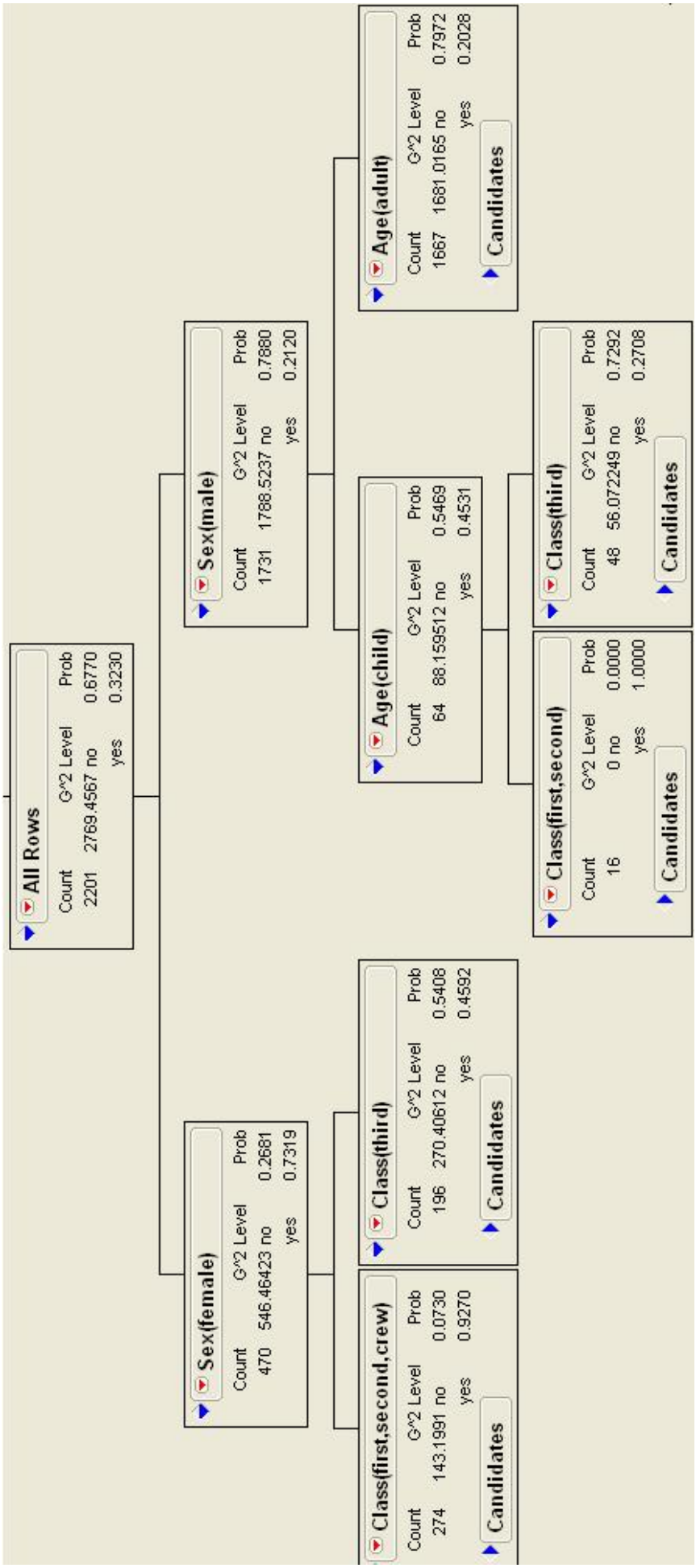


Figure 1.13: Construction of Decision Tree by JMP

```
For each class C
  Initialize E to the instance set
  While E contains instances in class C
  Create a rule R with an empty left-hand side that predicts class C
  Until R is perfect (or there are no more attributes to use) do
  For each attribute A not mentioned in R, and each value v,
   Consider adding the condition A=v to the LHS of R
   Select A and v to maximize the accuracy p/t
   (break ties by choosing the condition with the largest p)
   Add A=v to R
   Remove the instances covered by R from E
```

Figure 1.14: Pseudocode for a Basic Rule Learner

1.2.4.2 Feature Subset Selection (FSS)

Feature subset selection is a method to select relevant attributes (or features) from the full set of attributes as a measure of dimensionality reduction. Although some of the data mining techniques, such as decision trees, select relevant attributes, their performance can be improved, as the experiments have shown(Witten and Frank, 2005, p. 288). Two main approaches of feature or attribute selection are the filters and the wrappers (Witten and Frank, 2005). A filter is an unsupervised attribute selection method, which conducts an independent assessment on general characteristics of the data. It is called as a filter because the attributes are filtered before the learning procedure starts. A wrapper is a supervised attribute selection method, which uses data mining algorithms to evaluate the attributes. It is called as a wrapper because the learning method is wrapped in the attribute selection technique. In an attribute selection method, different search algorithms are employed, such as, genetic algorithm, greedy step-wise, rank search, and others.

1.2.5 Discretization

Some of the classifiers work well with discretized variables, such as tree and rule learners, therefore, discretizing numerical attributes is a very important preprocessing step. In addition, methods often produce better results (or run faster) , if the attributes are prediscretized(Witten and Frank, 2005, p. 287). There are two types of discretizers: unsupervised and supervised.

1.2.5.1 Unsupervised Discretization

Similar to unsupervised learning, unsupervised discretization works without the knowledge of the class attribute. Although unsupervised discretization is easy to

understand and arguably fast, it risks the danger of excluding some important information (for the learners) as a result of discrete intervals being too short or too long(Witten and Frank, 2005, p. 298). Some of the unsupervised discretization methods are:

1. Equal Interval Binning: as the name says, this discretization method divides the attribute in equal (predetermined arbitrary) intervals.

2. Equal Frequency Binning: this method is also called as histogram equalization, because the attributes are discretized in such a manner so that each intervals gets equal number of instances.

3. Proportional k-interval Discretization (PKID) (Yang and Webb, 2001): Yang and Webb (2003) warned that proportional k-interval discretization worked better for larger datasets, and suggested weighted proportional k-interval discretization. The proportional k-intervals are calculated using the Equation 1.16.

$$k = \sqrt{N} \tag{1.16}$$

   where, N is the number of instances.

1.2.5.2 Supervised Discretization

One of the best and state of the art supervised discretization method is Fayyad and Irani's (1992) minimum description length (MDL) criterion and entropy-based discretization. This discretization method is based on the idea of reducing the impurity by splitting (*cut point*) the intervals where the information value is smallest. The numeric attribute values are sorted in the ascending order, and a split is created where the subintervals are as pure as possible.

1.2.6 Bias

As data mining algorithms train and try to generalize the solutions, the generalization faces the problem of bias, and different algorithms face different type of bias. Some of the common biases are search bias, overfitting avoidance bias, sample bias, and language bias.

1.2.6.1 Search Bias

As data mining algorithm seek the optimal solution, which is defined by some criteria, such as, simplicity or best fit, a search bias is created. Different algorithms use different search heuristic, thus create search bias while searching for the optimal solution. For example, the results would be different if the criterion of optimal solution is highest performance rather than the criterion of simplest model.

1.2.6.2 Overfitting Avoidance Bias

Over generalization of the data makes the learning phase prone to poor performance on unseen data, therefore, data mining algorithm employ overfitting avoidance strategies. For example, decision trees use pruning and neural networks use penalties. These overfitting avoidance strategies create a bias, as techniques respond differently to each overfitting strategy.

1.2.6.3 Sample Bias

Sample bias, as the name suggests, occurs when data available for training are not representing the population fairly. Data itself creates the bias rather than the data mining algorithm. For example, sample containing only East Coast data for predicting something on national basis will cause sample bias (Menzies, 2006).

1.2.6.4 Language Bias

The structure and the working of an algorithm itself create language bias. Different algorithms behave differently with respect to the input and the style of generalizing. For example, some algorithms cannot take numbers as input, classification algorithms find pattern between the input and the output attributes, whereas, association algorithms find pattern between the input attributes.

1.3 Need for Research

As mentioned in Section 1.1, higher education institutions face tremendous challenge of student retention. Traditional methods used by researchers for solving this problem do not provide accurate solutions, as these methods face the problems of missing data, non-linearity of attributes, correlation, and massive amounts of data, whereas, data mining algorithms excel when presented with large amounts of data, and are robust enough to handle other problems.

Although application of data mining in the business world is a success story, the field of higher education is still experimenting with data mining. In the reviewed literature, only two research studies on the application of data mining in higher education explored other important options of data mining, especially, feature subset selection and evaluation: Barker et al. (2004) used principal component analysis to reduce the number of variables, but noted that the reduced data sets produced "much worse" results than the full data sets, and DeLong et al. (2007) mentioned the usage of attribute evaluation techniques, such as Chi-square gain, gain ratio, and information gain, however, did not provide comparative results.

Stewart and Levin (2001) noted, "the significance of data mining in sectors such as education have yet to be vindicated." Luan and Serban (2002) com-

mented, "suffice it to say that higher education is still a virgin territory for data mining." Chang (2006) commented, "although data-mining technologies have been applied widely and effectively in the business world, their use is relatively new to higher education." Herzog (2006) commented, "published studies on the use and prediction accuracy of data mining approaches in institutional research are few." From the above quotes, it is evident that there is still plenty of scope for experimentation and research in this field.

Lack of technical expertise has somewhat hindered the higher education researchers from exploring the data mining options fully; most of the researchers on higher education are social scientists, and most of the current research in data mining is done via "point-and-click" methods using various data mining software (Clementine, Enterprise Miner, etc). Therefore, there is a great need of thorough research in the field of application of data mining to the higher education data, especially in retention.

Tinto's (1975; 1988) theoretic model of student departure and other models based on Tinto's model attempted to find attributes that affect student's decision on departure. These attributes consisted of demographical, precollege experience, and family background information. Although these attributes, indeed, affect student's decision on departure, in order to produce prediction models, data mining algorithms might not need all of these attributes, and data mining tools can generate simplified and high performance models.

As the results produced by some of the data mining algorithms are not explainable, researchers term these as "black box" techniques. In the reviewed literature, it is apparent that existing research in this field has not attempted dimensionality reduction, as a way to increase the explanatory power. As Menzies et al. (2007a) suggested, use of ensemble techniques, such as, discretization, cross-validation, and feature subset selection, can produce high performance and good explanation models. Tinto (2006) noted "In the world of action, what matters are not our theories per se, but how they help institutions address pressing practical issues of persistence. Unfortunately, current theories of student leaving are not well-suited to that task." Therefore, it is critical to not only generate high-performance models, but also explainable models that can be translated into actions.

Need for research can be summarized as:

1. In the field of higher education and data mining, thorough research using various data mining tools, especially for student retention, is nonexistent.

2. Researchers in this field have not generated explainable high performance models using the ensemble techniques mentioned by Menzies et al. (2007a).

1.4 Research Objectives

The major research objectives of this study were:

1. To study attributes affecting student's drop-out decision.
2. Select attributes using different feature subset selection (FSS) techniques, such as, wrappers and filters.
3. Develop various data mining predictive models, such as, regression, decision tree, rule based, and neural networks, on data with all attributes and selected attributes. In addition, study the discretization effects using different discretization techniques.
4. Evaluate and compare these models using win-loss tables (Hall and Holmes (2003)), cross-validation, and quartile charts.
5. Generate explainable, but high performance, models to implement on the current data.

Chapter 2

Literature Review

> Data! Data! Data! ...I can't make bricks without clay.
>
> Sherlock Holmes

2.1 Theoretical Models of Student Dropouts

Researchers in higher education have extensively studied the theoretical models on the student dropouts problem developed by Spady (1970; 1971), Tinto (1975; 1988), and Bean (1980). These theoretical models led to the development of statistical models using linear and logistic regression (Pascarella and Terenzini, 1979, 1980; Gillespie and Noble, 1992; Brinkman and McIntyre, 1997; Beil et al., 1999; Brunsden et al., 2000). This section covers theoretical models developed by Spady (1970; 1971), Tinto (1975; 1988), and Bean (1980).

2.1.1 Spady's Model of Student Dropouts

2.1.1.1 Introduction

Spady's theoretical model (1970; 1971) (shown in Figure 2.1) was based on Durkheim's theory of suicide (Durkheim, 1951) and it focused on the interaction between student attributes and the influences caused due to the university environment. Spady argued that this interaction provides the student with the opportunity of incorporating into the academic and the social systems of the university; and the success derived in the academic and the social systems influence student's dropout decision. In the academic system, the successes in the form of rewards are grades and intellectual development. In the social system, normative

congruence and friendship support are the successes or the rewards. Spady defined normative congruence as, "attitudes, interests, and personality dispositions that are basically compatible with the attributes and influences of the environment." Spady further added that normative congruence and friendship support resembled the major social components of social integration in Durkheim's theory of suicide. Spady (1971) tested the theoretical model using multiple regression

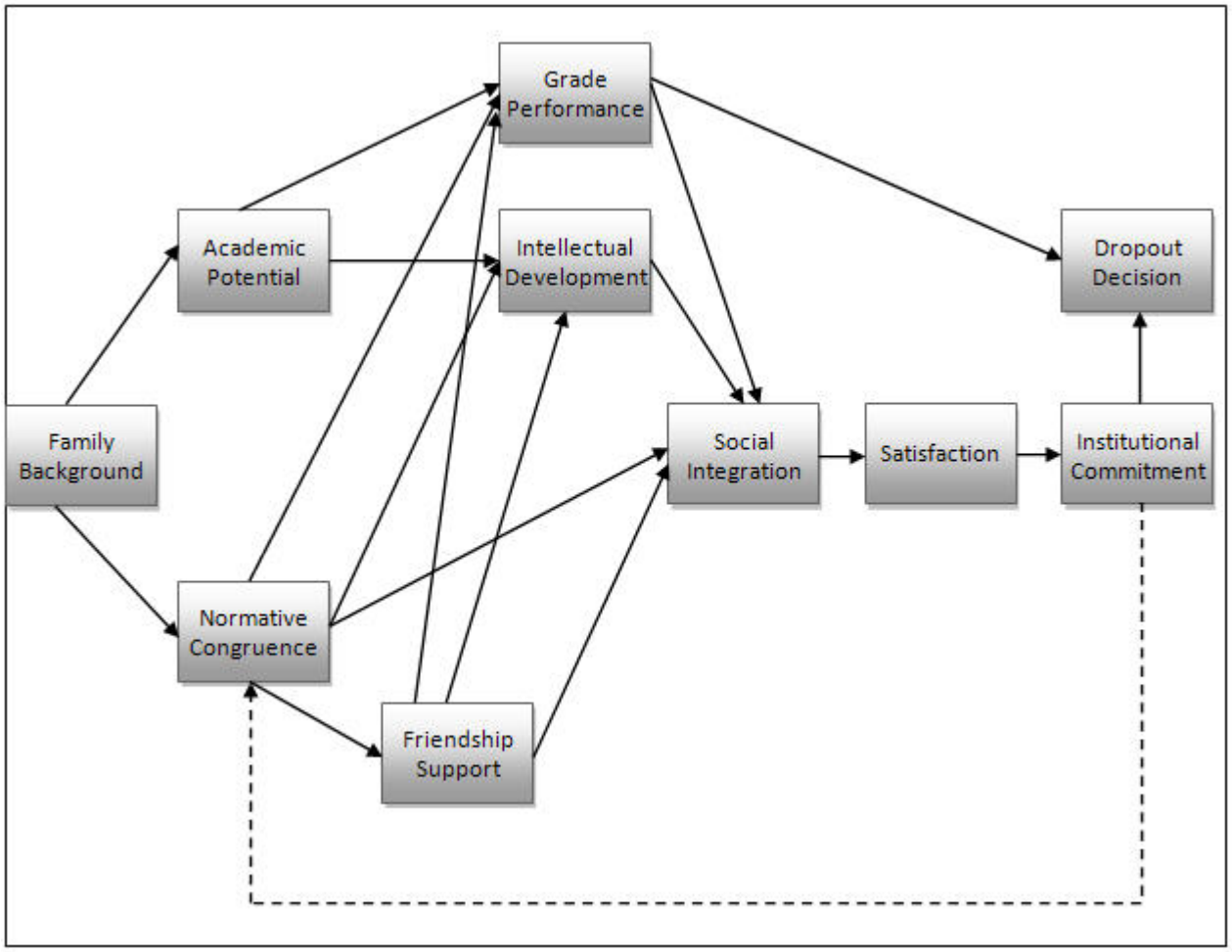


Figure 2.1: Spady's Theoretical Model (Spady, 1971)

with the longitudinal data of 683 first-year students. Spady collected these data using surveys and admissions data. Some characteristics of these students were:

- Sixty-two percent were men and 38% women
- Two-thirds attended schools that send over 50% of graduates to college
- More than one-third ranked in the upper 2% of the graduating class

- Two-thirds scored above 90^{th}percentile for all American students on SAT verbal and math

2.1.1.2 Variables

Table 2.1 is a list of variables from Spady's model on student dropout. These variables were from nine main components of the theoretical model, and each component had a cluster of other variables. Spady analyzed the model by adding these variables or cluster of variables in the step-wise multiple regression model.

2.1.1.3 Analysis

Spady analyzed the regression model by comparing the percentage of explained variance (R^2) for different combinations of dependent variables by either adding one cluster variable, or deleting one cluster variable from the regression model. The stepwise and unique contributions of variable clusters to the explained variance in first-year dropouts by sex is given in Table 2.2. Some of the key findings of this experiment were:

- Deleting institutional commitment from the full regression model reduced the explained variance (in first-year dropouts) by 12% for the women and 2.52% for the men
- Grades accounted for 5.91% of the explained variance for the men and 1.26% for the women
- Grade performance was the most important component of the dropout process for the men, followed by institutional commitment, social integration, extremes in independence from family, friendship support
- Institutional commitment was the most important component of the dropout process for the women, followed by being a natural science major, having high intellectual development, earning low grades, having unsatisfactory faculty contacts

2.1.1.4 Conclusion

After analyzing the data and the results, Spady revised the theoretical model, given in Figure 2.2, to match the consistent aspects of the data. Solid arrows in the Figure 2.2 depict that at least one element in a component has a statistically significant relationship with the dependent variable on the other end of the arrow for both men and women. This revised model indicated that friendship support for the women is directly dependent on elements in family background

Components	**Variables**	**Sub-variables**
Family background	Cosmopolitanism	Religious-ethnic origin
		Degree of urbanization
		Father's education
		Mother's education
		Father's occupation
	Family relationships	Parental martial stability
		Student's general happiness at home
		Freedom from family rule
		Psychological independence from parents
Normative congruence	Patterns of relationships	
	Personality dispositions	
	Measures of intellectual, moral, and vocational values	
	Attitude towards the university	
Academic potential	SAT verbal and math scores	
	High school rank	
	High school quality	
Friendship support	Quality and quantity of student's relationships with peers	
	Structural relations	Heterosexual relations
		Extracurricular involvements
		Faculty contacts
Intellectual development	Student's simulation in coursework	
	Expansion of intellectual perspectives	
	Ability to think systematically	
	Perceived excellence in academic work	
Grade performance	GPA	
Social integration	Sense of compatibility or dissonance with the university and its students	
Satisfaction	Student's satisfaction with the college experience	
Institutional commitment	Importance of graduating from the university	

Table 2.1: Variables from Spady's Model

Variable	Men		Women	
	Stepwise contribution	Unique contribution	Stepwise contribution	Unique contribution
Cosmopolitanism	0.45	0.22	3.18	1.33
Family relationships	1.54	1.67	0.84	0.66
High school experiences	2.91	1.61	4.10	2.17
Academic potential	1.62	0.21	1.28	0.31
Personality dispositions	2.22	0.50	3.47	2.87
Value orientations	0.63	0.88	2.85	1.39
Chicago dispositions	0.19	0.16	0.09	0.57
Subcultural orientations	1.61	1.06	2.75	3.62
Structural relations	5.64	1.82	5.03	2.92
Intellectual development	4.00	0.32	0.12	1.92
Grade performance	6.06	5.91	1.28	1.26
Social integration	1.89	0.81	1.03	0.01
Satisfaction	0.02	0.06	0.79	0.02
Institutional commitment	2.52	2.52	11.97	11.97
Total explained variance	31.32		38.79	

Table 2.2: The Stepwise and Unique Contributions of Major Variable Clusters to the Explained Variance in Dropouts

and normative congruence. Extracurricular participation and heterosexual relationship created strong friendships for both the sexes. For the men, the analyses indicated that the students with more conventional values, attitudes, and more socially oriented high school experiences were more likely to establish close relationships with others than the students without such experiences.

One of the most significant conclusions from this study was that the subjective intellectual growth of both men and women was apparently unrelated to their previous high school performance and measured intellectual capabilities. Spady concluded that women's decision to quit the college before the second year was pragmatic and rational, as their reaction and behavior rested on intrinsic, subjective, and social criteria, where academic and performance factors played a secondary role. Whereas, men reflected a sensitivity towards their roles as achievers within the formal academic system, and men quit the college based on extrinsic factors.

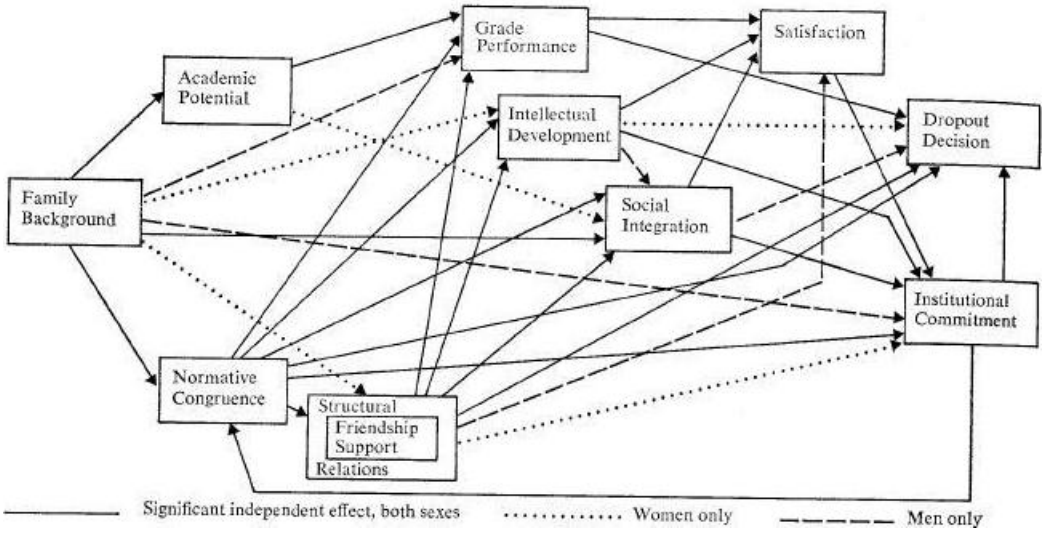


Figure 2.2: Spady's Revised Theoretical Model of the Undergraduate Dropout Process (Spady, 1971)

2.1.2 Tinto's Model of Student Dropouts

2.1.2.1 Introduction

Tinto's (1975) research paper on student dropouts is perhaps the most cited paper[1] in the field of student retention. Tinto's model like Spady's model (Spady, 1970, 1971) was based on Durkheim's theory of suicide (Durkheim, 1951). Tinto argued that the student's decision to leave or continue college was based on the student's integration in social and academic system; failure in any one of them was possibly a cause of the termination of the college. This model is given in Figure 2.3. Tinto argued that the dropout process, as depicted in Figure 2.3, was a "longitudinal process of interactions between the individual and the academic and social systems of the college during which a person's experiences in those systems (as measured by his normative and structural integration) continually modify his goal and institutional commitments in ways which lead to persistence and/or to varying forms of dropout"(Tinto, 1975, p. 94).

2.1.2.2 Variables

Tinto insisted that in order to develop a predictive model of student dropout the model should include individual characteristics and dispositions relevant to

[1]In the area of student retention, amongst the famous models on student dropouts of W. Spady (1970; 1971), V. Tinto (1975), and J. Bean (1980), researchers cited V. Tinto (1975) 949 times, W. Spady (1970; 1971) 337 times, and J. Bean (1980) 244 times. (Data from Google Scholar: `http://scholar.google.com` as of 02/21/08.)

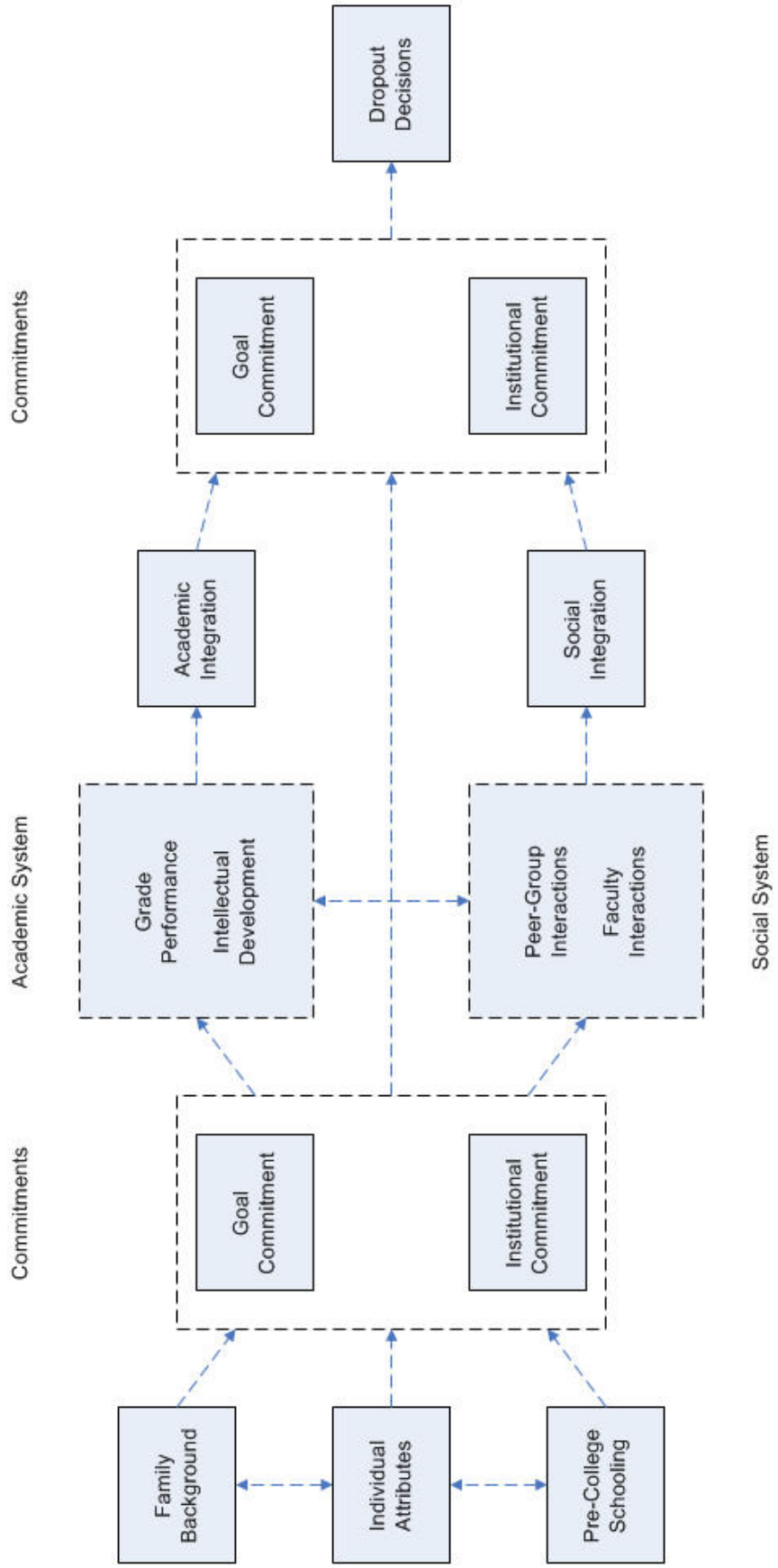


Figure 2.3: Tinto's Model of Student Dropouts (Tinto, 1975)

educational persistence. Researchers measure the individual characteristics and attributes in the forms of social status, high school experiences, community of residence, sex, ability, race, and ethnicity. Tinto suggested that in the predictive models, researchers should include expectational and motivational attributes of individuals. Researchers measure these attributes in career and educational expectations and levels of motivation for academic achievement of the individuals. Education expectation of an individual along with educational goal commitment was a very important input variable in Tinto's model, as students bring these aspirations to the college environment and it predicts how the individuals interact with the environment.

Precollege experiences, such as grade-point average, academic and social attainments, were important factors in this model, and along with these experiences individual characteristics and commitments, a student's integration in the academic and social system, Tinto argued, was in direct relation with the continuance of that student in the college. This integration causes a revision in the student's commitment towards the college and academic aspirations, and these new commitments derive student's decision to quit or continue college education. If either goal commitments or institutional commitments are low, the student is likely to dropout from that institution. Variables from different clusters in Tinto's model are shown in Table 2.3.

2.1.3 Bean's Model of Student Dropouts

2.1.3.1 Introduction

Bean developed this model (shown in Figure 2.4) using path analytic techniques, which the author called a "casual model", of student dropouts based on findings on employee attrition in work organizations (Bean, 1979, 1980); the basic assumption was that the reasons for which students leave college were similar to the reasons for which employees leave work. Bean studied Spady's and Tinto's models of student dropouts that were based on the theory of suicide, and noted that there was insufficient evidence on the link between dropping out and suicide. Bean criticized previous research because of the following reasons:

- Previous research ignored other literature and excluded other determinants of student attrition.
- Previous research ignored the distinction between analytic variables and demographic variables. Previous studies ignored the "directional causality" and discreetness of the variables.

Cluster	Variable	Measure	Explanation
Family Background	Characteristics of the family	Family's socioeconomic status	Children from lower status families exhibit higher rates of dropout than children from higher status families
		Parent's education	
	Quality of relationships within the family		Parents of persisters tend to enjoy more open, democratic, supportive, and less conflicting relationship with their children
Individual Characteristics	Measured ability	Grade performance in high school	
		Standardized test results	
	Gender		
Past Educational Experiences	Performance in high school	GPA or class rank	
	Characteristics of the high school		These characteristics affect the individual's aspirations, expectations, and motivation for college education
Goal Commitment	Terms of educational plans		
	Educational expectations or career expectations		
Academic Integration	Performance in college environment	Grade performance	
	Intellectual development		
Social Integration	Informal peer group associations		Friendship and faculty support are important social awards that affect educational and institutional commitment
	Semi-formal extracurricular activities		
	Interaction with faculty		
Institutional Characteristics	Resources		
	Facilities		
	Structural arrangement		
	Composition of its members		
	Quality of college		

Table 2.3: Variables in Tinto's Model

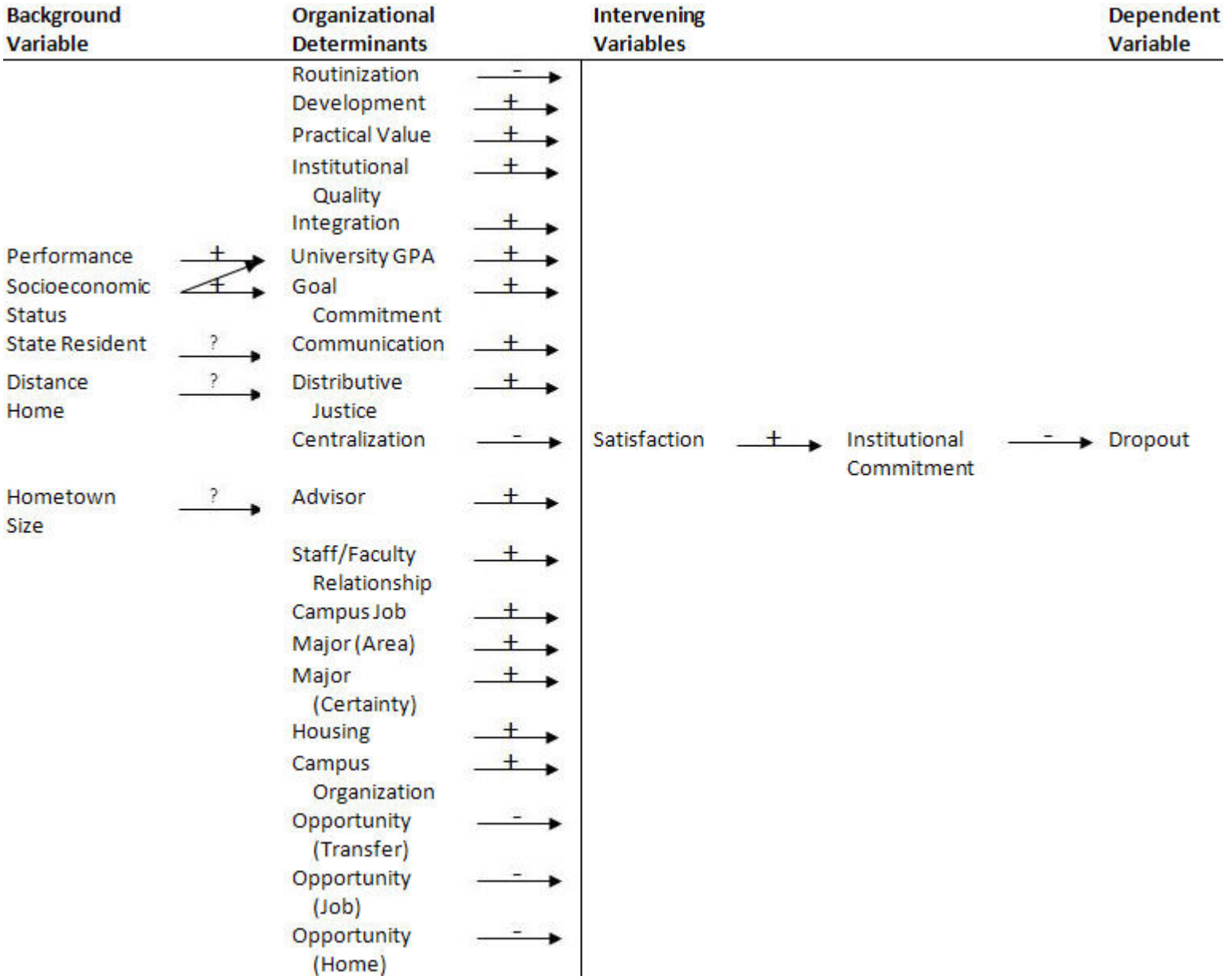


Figure 2.4: Bean's Casual Model of Student Dropout (Bean, 1980)

Cluster	Variable	Definition
Background Variables	Performance	The degree to which a student has demonstrated past academic achievement
	Socioeconomic status	The degree to which a student's parents have achieved status through occupational level
	State resident	Being a resident of the state where the college is located
	Distance home	Distance to a student's parents' home
	Hometown size	size of the community where a student spent the most time while growing up
Organizational determinants	Routinization	The degree to which the role of being a student is viewed as repetitive
	Development	The degree to which a student believes that he/she is developing as a result of attending college
	Practical value	The degree to which the student perceives that his/her education will lead to employment
	Institutional quality	The degree to which the college is perceived as providing good education
	Integration	The degree to which a student participates in primary or quasiprimariy relationships
	University GPA	The degree to which a student has demonstrated a capability to perform at college
	Goal commitment	The degree to which obtaining the bachelor's degree is perceived as being important
	Communication	The degree to which information about a being student is viewed as being received
	Distributive justice	The degree to which a student believes that he/she is being treated fairly by the institution
	Centralization	The degree to which a student believes that he/she participates in the decision making process
	Advisor	The degree to which a student believes that his/her advisor is helpful
	Staff/faculty relationship	The amount of informal contact with faculty members
	Campus job	The necessity of having a campus job to stay in school
	Major (area)	The area of one's field of study
	Major (certainty)	The degree to which a student is certain of what he/she is majoring in
	Housing	Where a person lives while attending college
	Campus organizations	The number of memberships in campus organizations
	Opportunity	The degree tow which alternative roles exists in the external environment
Intervening variables	Satisfaction	the degree to which being a student is viewed positively
	Institutional commitment	The degree of loyalty toward membership in an organization

Table 2.4: Definition of Variables from Bean's Model (Bean, 1980)

2.1.3.2 Variables

Variables and their definitions used in Bean's model are given in Table 2.4. Arrows in the model shown in Figure 2.4 were of casual relationship, and the signs on top of the arrow show the type of relationship (positive or negative). Bean noted that the GPA for students was similar to the salary for employees as a performance measure. Many other variables were consistent with Tinto's model (Bean, 1980, p. 156), but were derived from Price's1977 turnover model in work organizations.

2.1.3.3 Analysis

To test this model, Bean provided questionnaires to the freshmen; out of 2,587 new freshmen 1,111 had returned the questionnaires, and out of these questionnaires, the author selected two homogeneous samples of 366 men and 541 women. Bean selected only the students who were under 22 years of age, Caucasian race, U.S. citizen, and single. Bean used multiple regression and path analysis to analyze and test the casual model of student dropouts.

Using multiple regression, Bean found that for women institutional commitment, institutional quality, and routinization were statistically significant. Using these clusters of variables, Bean's model had the R^2 value of 0.22. For men, institutional commitment, routinization, satisfaction, and communications were statistically significant. Bean found that for women, institutional commitment was more than $4\frac{1}{2}$ times as important as institutional quality. The author found that the amount of explained variance for women ($R^2 = 0.22$) was twice the amount of explained variance for men ($R^2 = 0.9$).

Using the coefficient (β) values, Bean removed nonsignificant variables from the regressions equations to create parsimonious models. Bean regressed on all the clusters variables and kept important variables in the model using R^2 values; the path models of student attrition for women and men are shown in Figure 2.5 and Figure 2.6 respectively.

2.1.3.4 Conclusion

Using this sample of data, Bean found that institutional commitment was the primary variable influencing dropout. In addition, the author found that variables: routinization, opportunity (transfer, job, home), university GPA, practical value, institutional quality, and satisfaction were important in this model. Institutional quality and opportunity (transfer) were the two most important variables that influenced institutional commitment for men and women. Bean noted that performance was the only important background variable along with routinization, development, and university GPA. According to Bean, this model performed better ($R^2 = 0.12$ for men and $R^2 = 0.21$ for women) than the earlier models except

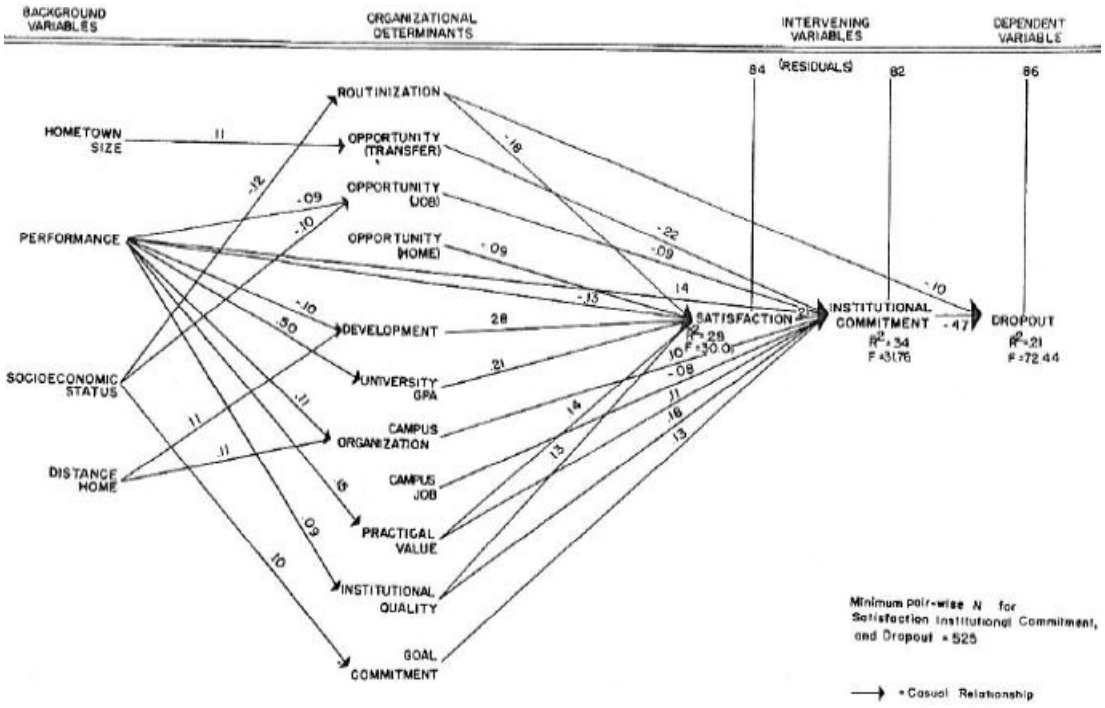


Figure 2.5: Bean's Path Model of Student Attrition for Women (Bean, 1980)

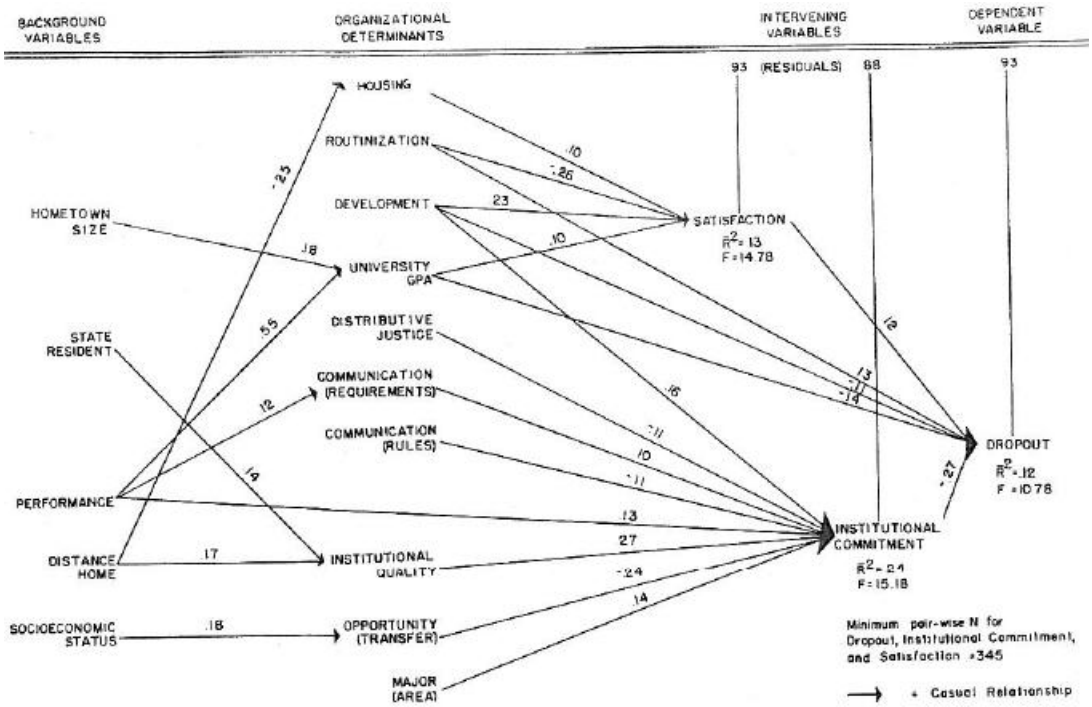


Figure 2.6: Bean's Path Model of Student Attrition for Men (Bean, 1980)

that of Spady's model ($R^2 = 0.31$ for men and $R^2 = 0.39$ for women)(Bean, 1980, p. 179).

2.1.4 Studies Based on Theoretical Models

2.1.4.1 Studies by Terenzini and Pascarella

Terenzini and Pascarella extensively analyzed Tinto's model (Tinto, 1975) of student dropout (Terenzini and Pascarella, 1980; Pascarella and Terenzini, 1979, 1980). In (Terenzini and Pascarella, 1980), the authors summarized results from six studies performed on freshmen at Syracuse University from 1974 to 1976. Terenzini and Pascarella performed discriminant analysis and stepwise multiple regression to construct a validity on Tinto's model. Out of these six studies, two of them focused on the faculty interaction component of Tinto's model. Summary of results from this study are given in Table 2.5.

Terenzini and Pascarella (1980)concluded on these points:

- the quality and impact of a student's peer group relations was the most important factor for women for persistence.
- pre-college characteristics of students were significant factors in student's attendance behavior.
- the frequency of students' informal contact with faculty members was consistently related to freshmen year persistence.

2.1.4.2 Study by Stage

This study by Stage (1989) focused on analysis of college withdrawal using Tinto's framework, and it examined associations among background characteristics, commitment levels, institutional involvement and motivational orientations (certification, cognitive, and community service). Stage (1989) collected the data via surveys sent to the freshmen students. Some of the variables used in this study are given in Table 2.6. The author used logistic regression to find significant relationships between variables and to provide equations model.

Stage (1989) used LISREL (Jöreskog and Sörbom, 1989) to model the data using logistic regression. The final model had the chi-square value of 458.38 with 424 degrees of freedom. The author used stepwise logistic regression to select variables with a p value less than 0.1 and p value greater than 0.15 to remove a variable. Table 2.7 shows the variables that were statistically significant predictors of persistence.

Some of the conclusions of this study were:

- In the certification group, positive effects for male students and low measures of mother's education were found towards persistence.

Characteristics / Results	**Study 1**	**Study 3**	**Study 5**	**Study 6**
Sample size: Leavers	63	90	61	61
Sample size: Stayers	63	428	436	436
Validation sample	253	NA	29 Leavers and 237 Stayers	29 Leavers and 237 Stayers
Analytical methods	3 discriminate function analyses	setwise multiple regression	setwise discriminate function analysis	setwise discriminate function analysis
Total variance explained	24.60%	25.60%	30.90%	47.6% for men and 55.3% for women
Significant main effects variables	Demand / challenge	No. of Faculty contacts	Institutional & goal commitments	Men: institutional & goal commitment; discuss intellectual matters
	No. of Faculty contacts	Affective appeal of academic program	Interactions with faculty	Women: peer group relationships; faculty interactions
	Interest value	Dullness of academic program	Faculty concern with student development and teaching	

Table 2.5: Summary of Results from Terenzini and Pascarella Studies (Terenzini and Pascarella, 1980)

Cluster	Variable or Survey Question
Background Characteristics	Mother's education
	Father's education
	Age
	Sex
	Ethnicity
Goal Commitments	It is important for me to graduate from college
	I have no idea at all what I want to major in
Initial Commitments	It is important for me to be enrolled
	It is likely that I will register at this university next fall
Academic Integration	Academic Development Scale
	Faculty concern scale
	GPA
	Credits earned during the first semester
	Hours spent on academic extra-curricular activities
Social Integration	Peer Group Relations Scale
	Informal Faculty Relations Scale
	Residency
	Campus employment
	Hours spent on social activities
	Hours spent on intercollegiate athletics

Table 2.6: Variables in Stage's Study (Stage, 1989)

Subgroup	Independent Variable
Certification	Mother's education
	Gender (female)
	Academic integration
	Institutional commitment
	Ethnicity × academic integration
	Ethnicity × social integration
Cognitive	Mother's education
	Academic integration
	Institutional commitment
Community Service	Institutional commitment
	Goal commitment
	Gender × Social integration

Table 2.7: Selected Variables from Stage's Model(Stage, 1989, p. 395)

- In the cognitive group, students with high levels of mother's education were likely to persist.
- Results agreed with Tinto's claim that background effects influenced persistence.
- Statistically significant interaction effects were found between ethnicity and social integration, ethnicity and academic integration, and gender and social integration.
- In the certification group, minorities with high levels of academic integration were not likely to persist as majority students.
- Academic integration significantly (positively) influenced persistence.

2.1.4.3 ACT Research Report

Gillespie and Noble (1992) studied Tinto's model of persistence using predictor variables from five institutions. The authors used linear and logistics regression to develop the prediction models, and the primary aim of these prediction models was to identify high-risk students and intervening them to keep them in school. The predictor variables used in this study are given in Table 2.8.

Gillespie and Noble computed correlations between each predictor variable and the output variable; variables that had a correlation coefficient greater than or equal to 0.10 and statistically significant were included in the prediction model. If the included variables had large amounts of missing data or were similar to other variables were eliminated from the model, then the authors excluded these variables from the model.

Some of the important variables for all institutions were: goal commitment, institutional commitment, academic fit ins:/ integration, and high school preparation. For some institutions, plans to work while in school was important in predicting persistence. In addition, this study found that if the students' satisfaction with their employment opportunities decreased over time, they were more likely to persist. The authors found that the results from this study were consistence with previous research using Tinto's model.

2.1.4.4 Study by Dey and Astin

Dey and Astin (1993) created prediction models for student retention using logistic regression, probit analysis, and linear regression. The authors collected data on behavioral and motivational items from surveys. Table 2.9 shows all of the variables used in this study.

Dey and Astin found that the results from linear regression were close to that of logistic regression or probit analysis. Multiple R for logit, probit, and

Cluster	Variable
Background Information	Demographic characteristics
	Academic development
	Nature of high-school preparation
	Extracurricular participation
	Financial
	Family attitudes towards education
	Academic and personal needs
	Self-reported physical health
	Self-reported personality characteristics
Initial commitment to Institution	Purpose for enrolling
	Institutional choice
	Importance of selected institutional characteristics
	Full-time/part-time enrollment
Initial and subsequence academic goal commitment	Expected degree and strength of expectations
	Certainty of career aspirations
	Commitment to and value placed on college education
	Actual vs expected progress in reaching academic goals
	Satisfaction with academic progress and services
	Absenteeism
Student/institution academic fit	Does the institution meet the academic expectations of the student
	Course enrollment, completion and grades
	Need for remediation
	Perception of relationships with faculty
Student/institution social fit	Amount of friendship, peer support
	Social relationships with faculty and staff
	Comfort and satisfaction with the environment
	Extracurricular activities
Student/institution financial fit	Amount of immediate family contribution
	Hours/week spent working
	Loans required to meet expenses

Table 2.8: Predictor Variables in ACT Research Study (Gillespie and Noble, 1992)

Variables
Age
Concern about ability to finance college education
Hours per week spent
•Studying/homework
•Socializing with friends
•Talking with teachers outside of class
•Exercising/sports
•Partying
•Working (for pay)
•Volunteer work
•Student club/groups
•Watching TV
•Hobbies
Average high school grades
Reasons for attending college
•To be able to get a better job
•To gain a general education and appreciation of ideas
•To improve my reading and study skills
•There was nothing better to do
•To make a more cultured person
•To learn more about things that interest me
•To prepare myself for graduate or professional school
•My parents wanted me to go
•I couldn't find a job
•Wanted to get away from home
Female student

Table 2.9: Variables used in Dey and Austin's Study1993

regression were 0.354, 0.351, and 0.323. In large samples, the fit of predictions based on linear regression were equal or better as the fits that were obtained with logistic regression or probit analysis.

2.2 Other Studies

Waugh et al. (1994) studied the predictive values of ethnicity, SAT/ACT scores, and high school GPA towards retention and graduation rates. The authors found that high school GPA had moderate correlation with graduation (0.22) and retention/graduation (0.21); however, SAT (0.10) and ACT scores (0.01) had no relationship with retention. In addition, African-American students with low GPAs were noted as vulnerable to dropping out.

Variables
Age
Sex
Ethnicity
Residency
College
High School GPA
SAT Score
First Quarter GPA
Participation in Education Opportunities Program
Enrollment in Freshman Orientation Course

Table 2.10: Variables Used in the Study by Murtaugh et al. (1999)

Murtaugh et al. (1999) created prediction models on the retention of university students using survival analysis. The authors used demographic and academic variables, which are given in Table 2.10, for 8,867 students. The results indicated some of the important variables: age, residency, high school performance, and enrollment in the Freshman Orientation Course; high school GPA had superior predictive value than SAT score. The authors found that that in-state students had lower attrition rates than non-residents.

Herzog (2005) studied the effect of different variables, such as student demographics, high school preparation, college experience, and financial aid status, on student return, dropout/stopout, and transfer from the university (see Table 2.11). The author used multinomial logistic regression to study these effects. The author found that the out-of-state students had twice the odds of dropping out than the in-state students. Parental income for upper-income students faced lower dropout odds. In the first term, the middle-income students with high levels of unmet need faced twice th risk of dropping out. The author noted that gender had no impact on retention and that the grade point average was a strong predictor of student persistence.

Researchers have conducted longitudinal studies to study the effects of academic variables on student retention (Gillespie and Noble, 1992; Felder et al., 1998; Beil et al., 1999; Ishitani and DesJardins, 2002; Ishitani and Snider, 2004; Snider and Boston, 2004). Longitudinal studies unlike cross-sectional studies track the same cohort for a time period. Beil et al. (1999) studied effects of academic integration, social integration, and commitment on student retention. The authors found that even though academic and social integration were important, when commitment was considered in the logistic regression model, it was a significant predictor of retention, and academic and social integration were insignificant; however, academic and social integration influenced commitment, in

Clusters	Variables
Student Demographics	Age
	Sex
	Ethnicity
	Residency
	Parent Income
High School Preparation	Composite Index
College Experience	On-campus Living
	Credit load
	GPA
	Math requirement
	First-year math grades
	Remedial course enrollment
	Peer challenge score
	Class selection
	Use of recreational facilities
Financial Aid Status	Package
	Eligibilty type
	Source
	Amount
	Remaining need
	Second-year offers

Table 2.11: Variables in the Study by Herzog (2005)

turn, affected retention. In addition, the authors cautioned on the multicollinearity between academic and social integration.

Ishitani and DesJardins (2002) studied national survey data using longitudinal methods (event history modeling) to research the factors that have effect on student departure at specific period of time. The authors found these variables to be statistically significant: family income, mother's education, self-educational aspiration, first-year GPA, SAT total scores, institutional type, and financial aid.

Ishitani and Snider (2004) studied the effects of college preparation programs on student retention. The authors noted the significant influence of student aspirations, parental encouragement, parent's education, and high school grades. Using survival analysis, the authors found that the students who took SAT/ACT preparation courses were more likely to persist, students who talked their parents about going to college were more likely to persist, lower levels of family income, parental education and being a first-generation college student affected the persistence negatively.

Researchers have studied the effect of financial aid and need on persistence and enrollment (Braunstein et al., 1999; John, 2000; Bresciani and Carson, 2002).

Braunstein et al. (1999); John (2000) noted that financial aid indeed had influenced the decisions of enrollment and persistence, however, it was difficult to understand the whys and the hows of the process. College debt had an influence on whether students could afford to continue their enrollment or re-enrollment.

Bresciani and Carson (2002) examined the effects of unmet financial need and amount of gift aid to the student persistence, and defined unmet need as: "unmet need is the amount of money that is left after all the aid that is awarded to a student has been subtracted from his or her need amount." Financial aid offices calculate the need amount by subtracting the expected family contribution (EFC) from the cost of attendance at a college. Although R^2 value obtained using linear regression remained around 0.022, the results explained the fact that likeliness of persistence decreased with the amount of unmet need.

Beeson and Wessel (2002) studied the impact of working on campus on the persistence of freshmen. The authors found that the freshmen, who worked on campus, persisted at slightly higher rates from fall to spring of their first year, and year to year; however, the authors did not find working on campus statistically significant towards graduation or persistence at the studied university.

DesJardins et al. (2002) affirmed that minority students, older students, and low family income students had high probabilities of dropping out of the college. The authors noted that high GPA lowered the risk of dropout, but the effect diminished over time, and that the financial aid was an insignificant factor for increasing graduation, however, it indeed reduced the student stopout.

Lotkowski et al. (2004) conducted a comprehensive literature search and identified more than 400 studies on student retention, and selected academic and non-academic factors from 109 studies pertaining to retention. The authors used stepwise multiple regression to identify the factors that had the strongest relationships with college retention; they found that high school GPA (HSGPA) had the strongest relationship with college retention in the academic factors and academic related skills in the non-academic factors. Other factors are given in Table 2.12 in the order of importance from highest to lowest.

2.3 Data Mining in Education

Various researchers have applied data mining in different areas of education, such as enrollment management (Gonzlez and DesJardins, 2002; Chang, 2006; Antons and Maltz, 2006), graduation (Eykamp, 2006; Bailey, 2006), academic performance (Naplava and Snorek, 2001; Pardos et al., 2006; Vandamme, 2007; Ogor, 2007), gifted education (Ma et al., 2000; Im et al., 2005), web-based education (Minaei-Bidgoli et al., 2003), retention (Druzdzel and Glymour, 1994; Sanjeev and Zytkow, 1995; Massa and Puliafito, 1999; Stewart and Levin, 2001; Veitch, 2004; Barker et al., 2004; Salazar et al., 2004; Superby et al., 2006; Sujitpara-

Variables	**Description**	**Strength of Relationships**
Academic-related skills	Time management skills, study skills, and study habits (taking notes, meeting deadlines, using information resources).	Strong
Academic self-confidence	Level of academic self-confidence (of being successful in the academic environment).	Strong
Academic goals	Level of commitment to obtain a college degree.	Strong
Institutional commitment	Level of confidence in and satisfaction with institutional choice.	Moderate
High school grade point average	Cumulative grade point average student average (HSGPA) earned from all high school courses.	Moderate
Social support	Level of social support a student feels that the institution provides.	Moderate
Contextual influences	The extent to which students receive financial aid, institution size and selectivity.	Moderate
Socioeconomic status	Parents educational attainment and family income.	Moderate
Social involvement	Extent to which a student feels connected to the college environment, peers, faculty, and others in college, and is involved in campus activities.	Moderate
ACT Assessment score	College preparedness measure in English, mathematics, reading, and science.	Moderate
Achievement motivation	Level of motivation to achieve success.	Weak
General self-concept	Level of self-confidence and self-esteem.	Weak

Table 2.12: Strength of Relationships of Academic and Non-Academic Factors with Retention (Lotkowski et al., 2004)

pitaya, 2006; Herzog, 2006; Atwell et al., 2006; Yu et al., 2007; DeLong et al., 2007), and other areas (Intrasai and Avatchanakorn, 1998; Baker and Richards, 1999; Thomas and Galambos, 2004). Luan and Serban (2002) listed some of the applications of data mining to higher education, and provided some case studies to showcase the application of data mining to the student retention problem.. Delavari and Beikzadeh (2004); Delavari et al. (2005) proposed a data mining analysis model to used in higher educational system (refer to Table A.1), which identified various research areas in higher education that could use data mining.

2.3.1 Data Mining for Enrollment Management

Gonzlez and DesJardins (2002) used artificial neural networks (ANN) to predict application behavior, and compared the results with logistic regression. The ANN model correctly classified 80.2% of prospective students, and the logistic regression model correctly classified 78% of prospective students. Chang (2006) used neural networks, Classification And Regression Tree (CART), and logistic regression to predict admissions yield. CART, neural network, and logistic regression obtained 74%, 75%, and 64% probability of correct classification respectively. Antons and Maltz (2006) used decision trees, neural networks, and logistic regression to predict the enrollees out of the applications. For the real data, the logistic regression model correctly classified 66% of the admitted applicants, however, it correctly classified only 49% of the enrollees and 78% of non-enrollees.

Nandeshwar and Chaudhari (2007) used ensemble data mining techniques to find the reasons of student enrollment using student admissions (demographic and academic) data. Using feature subset selection and discretization techniques, Nandeshwar and Chaudhari (2007) were able to reduce the number of variables to one from 287, and the authors were able to explain the student enrollment decision using very simple rule based models with an accuracy around 83%. The authors found that the accepted applicants decided to enroll if they received any amount of financial aid.

2.3.2 Data Mining for Graduation

Eykamp (2006) used data mining to study the effects of taking advance placement classes reduced the time to degree. Bailey (2006) developed data mining model to predict the graduation rates using the Integrated Postsecondary Education Data System (IPEDS)[1]. IPEDS is a National Center for Education Statistics (NCES) initiative that collects data from most of the higher-education institutions. The author collected data from the IPEDS for 5,771 institutions on various areas, such as, faculty salaries, staff headcount, financial aid, and institutional characteristics.

[1] http://nces.ed.gov/ipeds/

The objective of this study was to determine the institutional areas that influences graduation using CART. The best relationship between actual and predicted graduate rate, given by Pearson's correlation (r), was 0.885.

2.3.3 Data Mining for Academic Performance

Naplava and Snorek (2001) applied Group Method of Data Handling GMDH on student application data to predict the success of new students at the Czech Technical University of Prague. The authors used neural networks, combinatorial algorithm , and Multi-layered Iterative Algorithm (MIA) to predict the academic performance. Schumann (2005) studied high school data to predict academic performance using data mining.

Pardos et al. (2006) used Bayesian networks to develop prediction models to asses skill models for student testing. Using the question sets from the Massachusetts Comprehensive Assessment System (MACAS), the authors created ASSISTment, an online tutoring system, for 8^{th}grade mathematics students to test the grain size of the skills. The authors found that the medium-sized (39 skills) produced the best model to track student performance.

Vandamme (2007) applied discriminant analysis, neural networks, random forests, and decision tree to predict students' academic success. The authors divided the dependent variable in three categories: low risk, medium risk, and high risk students. Using the data collected from questionnaires, the overall correct classification rates for decision trees, neural networks, and discriminant analysis were 40.63%, 51.88%, and 57.35% respectively.

Ogor (2007) developed a methodology to deploy a student performance assessment and monitoring system using data mining techniques. The author developed rule induction and neural network models to predict academic performance using student demographic information and course assessment data.

2.3.4 Data Mining for Gifted Education

Ma et al. (2000) developed data mining models for selecting the right students for remedial classes from the Gifted Education Programme (GEP) in Singapore. Using association rule mining, the authors predicted weak students from the GEP cohort and suggested remedial classes for these students, whereas, traditionally, the administrators used a cutoff score on tests to select students for remedial courses (the authors argued that this method selected "too many" students).

As the current tests for identifying gifted students were unable to identify the "potentially gifted" students, Im et al. (2005) developed neural network models to identify such students in Korea. The authors created questionnaires to collect the data on students to measure the capabilities in the areas of scientific attitude, leadership, morality, creativity, etc. In addition, the authors build a model to

evaluate the similarities between students' characteristics and students' type of giftedness to create a giftedness quotient.

2.3.5 Data Mining for Web-Based Education

Minaei-Bidgoli et al. (2003) used data mining to predict the final grades of students based on the features extracted from students' logged data in an education web-system at Michigan State University. The authors developed classification models to find any patterns in the student usage data, such as time spent on problems, reading the supporting material, total number of tries, and others. The authors used quadratic Bayesian classifier, nearest neighbor, Parzen-window, multi-layer perceptron, and decision tree. In addition, the authors used Genetic Algorithm (GA) to select features to maximize the classification accuracy. The authors found that classifiers with GA for feature selection increased the accuracy by 10 to 12 percentage points.

2.3.6 Data Mining for Other Applications

Intrasai and Avatchanakorn (1998) developed an academic planning application using genetic algorithm. This application allowed administrators to search for suitable locations to open new campuses in the rural areas of Thailand. From the existing university data, this application extracted clusters of useful information to help administrators on deciding which majors to offer and which place to build the facility depending on the student population density in the area and travelling distance. Baker and Richards (1999) developed forecasting models for educational spending using linear regression and neural networks. Linear regression and neural networks models achieved an average R^2 value of 0.99.

Thomas and Galambos (2004) used regression and decision trees to investigate how students' characteristics and experiences influenced their satisfaction in public research university. The stepwise (forward and backward) linear regression models resulted in R^2 values in the range of 0.37 to 0.58. Using decisions tree algorithm (CHAID), the authors explained the satisfaction of students in different areas; the author noted that the rules from these trees supported Tinto's theory that the effects of social integration may compensate for weak academic integration. Beitel (2005) applied data mining tools to predict program evaluations for primary school courses.

2.3.7 Data Mining for Student Retention

Druzdzel and Glymour (1994) were the first to apply knowledge discovery algorithm to study the student retention problem. The authors applied TETRAD II[2], a casual discovery program developed at Carnegie Mellon University, to the

[2] http://www.phil.cmu.edu/projects/tetrad/index.html

U.S. news college ranking data to find the factors that influenced student retention, and they found that the main factor of retention was the average test score. Using linear regression, the authors found that test scores alone explained 50.5% of the variance in freshmen retention rate. In addition, they concluded that other factors such as student-faculty ratio, faculty salary, and university's educational expense per student were not casually (directly) related to student retention; and suggested that to increase student retention universities should increase the student selectivity.

Sanjeev and Zytkow (1995) used 49er, a pattern discovery process developed by Żytkow and Zembowicz (1993), to find patterns in the form of regularities from student databases related to retention and graduation. The authors found that academic performance in high school was the best predictor of persistence and better performance in college, and that the high school GPA was a better predictor than the ACT composite score. In addition, they found that no amount of financial aid influenced students to enroll for more terms.

Massa and Puliafito (1999) applied Markov chains modeling technique to create predictive models for the student dropout problem. By tracking the students for 15 years, the authors created state variables for the number of exams appeared, average marks obtained, and the continuation decision. Using data mining, Stewart and Levin (2001) studied the effects of student characteristics to persistence and success in an academic program at a community college. They found that the student's GPA, cumulative hours attempted, and cumulative hours completed were the significant predictors of persistence, and that young males were a high risk group.

Veitch (2004) used decision trees (CHAID) to study the high school dropouts. Using 25-fold cross-validation, the overall misclassification rate was 15.79%, and 10.36% of students, who did drop out were classified as non-dropouts. In this study, GPA was the most significant predictor of persistence. Salazar et al. (2004) used clustering algorithms and C4.5 to study graduate student retention at Industrial University of Santander, Colombia. The authors found that the high marks in the national pre-university test predicted a good academic performance, and that the younger students had higher probabilities of a good academic performance.

Barker et al. (2004) used neural networks and Support Vector Machines (SVM) to study graduation rates; the first-year advising center (University College at University of Oklahoma) collected data via a survey given to all incoming freshman. It is worthwhile to note that Barker et al. (2004) excluded all the missing data from the study, which constituted for approximately 31% of the total data. Overall misclassification rate was approximately 33% for various dataset combinations. The authors used principal component analysis to reduce the number of variables from 56 to 14, however, reported that the results using

Model	Model Description	Training data	Validation data	Testing data
Decision Tree 1	Entropy split criterion	91%	90%	88%
Decision Tree 2	Chi-square split criterion	84%	83%	82%
Decision Tree 3	Gini Index split criterion	84%	83%	82%
Logistic Regression	Stepwise regression	78%	77%	73%

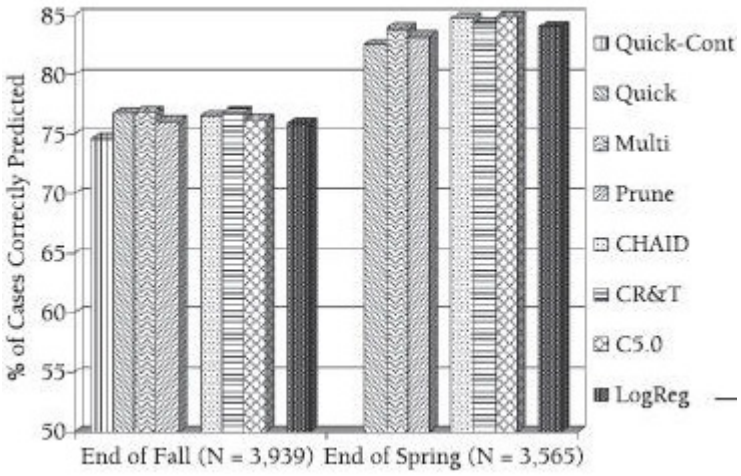
Table 2.13: Precision Rates Obtained (Atwell et al., 2006)

the reduced datasets were "much worse" than the complete datasets.

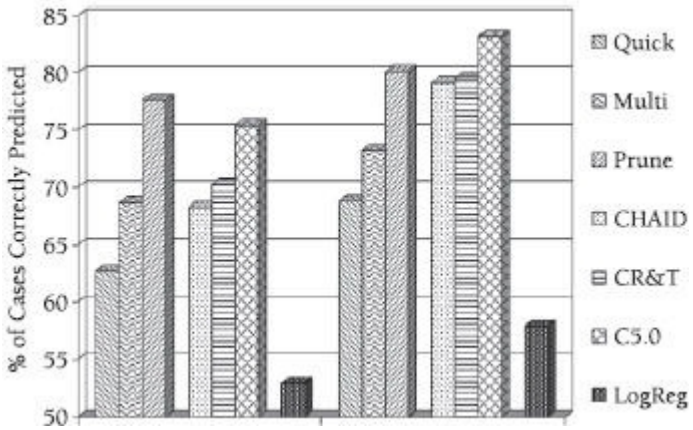
Superby et al. (2006) applied discriminant analysis, neural networks, random forests, and decisions trees to survey data at the University of Belgium to classify new students in low-risk, medium-risk, and high-risk categories. The authors found that the scholastic history and socio-family background were the most significant predictors of risk. The overall classification rates for decision trees, random forests, neural networks, and linear discriminant analysis were 40.63%, 51.78%, 51.88%, and 57.35% respectively.

Using the National Student Clearinghouse (NSC) data, Sujitparapitaya (2006) differentiated between stopout, retained, and transfer students. The overall classification rates for the validation sets using logistic regression, neural networks, C5.0 were 80.7%, 84.4%, and 82.1% respectively. Herzog (2006) used American College Test's (ACT) student profile section data, NSC data, and the institutional student information system data for comparing the results from the decision trees, the neural networks and logistic regression to predict retention and degree-completion time. The author substituted mean average ACT scores for missing scores. Decision trees created using C5.0 performed the best with 85% correct classification rate for freshmen retention, 83% correct classification rate for degree completion time (three years or less), 93% correct classification rate for degree completion time (six years or more) for the validation datasets.

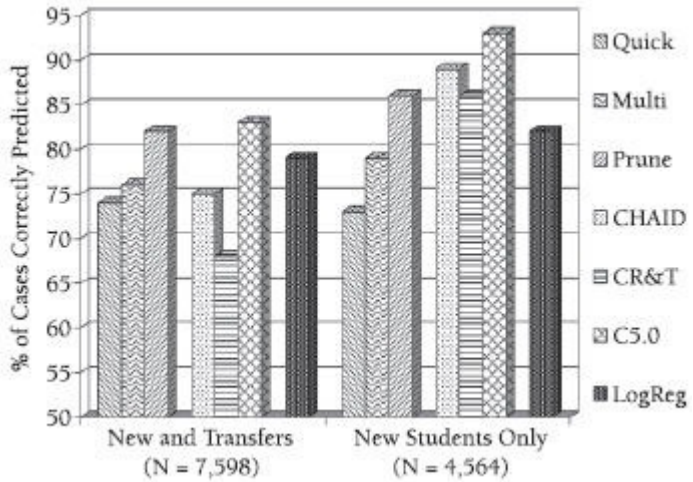
Atwell et al. (2006) used University of Central Florida's student demographic and survey data to study the retention problem with the help of data mining. In this study, university retained approximately 82% of the freshmen from the study, and it used 285 variables to create data mining models. The authors used nearest neighbor algorithm to impute more than 60% observations with missing values. Using decision trees with the entropy split criterion, the authors obtained precision of 88% for the not-retained outcome using the test data, and the actual retention rate for this test data set was 82.61%; other results from this study are given in Table 2.13.



(a) Freshmen Retention



(b) Degree Completion Time (three years or less)



(c) Degree Completion Time (six years or more)

Figure 2.7: Results Comparison for Freshmen Retention and Degree Completion Time (Herzog, 2006)

Yu et al. (2007) studied the data from Arizona State University using decision trees, and included variables, such as demographic, pre-college academic performance indicators, current curriculum, and academic achievement. Some of the important predictor variables were accumulated earned hours, in-state residence, and on campus living.

To study the retention problem using data mining for the admissions data, DeLong et al. (2007) applied various attribute evaluation methods, such as Chi-square gain, gain ratio, and information gain, to rank the attributes. In addition, the authors tested various classifiers, such as naïve Bayes, AdaBoost M1, BayesNet, decision trees, and rules, and noted that AdaBoost M1 with Decision Stump classifier performed the best in terms of precision and recall, hence, used this classifier for further experimentation. The authors balanced the class variable (retained and not retained) and obtained over 60% classification rates for both retained and not retained outcome. The authors concluded that the number of programs that the student applied to that specific institution and the student's order of program admit preference were the most significant predictors of retention.

Pittman (2008) compared various data mining techniques (artificial neural networks, logistic regression, Bayesian Classifiers, and decision trees) applied to the student retention problem, and also used attribute evaluators to generate rankings of important attributes. The author concluded that logistic regression performed the best in terms of ROC-curve area.

2.4 Customer Retention in the Business World

The applications of data mining in the business world are plenty, such as knowledge discovery in National Basketball Association (NBA) data (Bhandari et al., 1997), forecasting in airline business (Hueglin and Vannotti, 2001), direct marketing for charity (Chan et al., 2002), identification of early buyers (Rusmevichientong et al., 2004), application in physics (Roe et al., 2005), and the customer retention or *churn* analysis (Eiben et al., 1998; Smith et al., 2000; Ng and Liu, 2000; Bin et al., 2007).

Eiben et al. (1998) studied mutual fund investment data using logistic regression, rough data models, and genetic programming to predict customer retention. The authors found that genetic programming performed the best in terms of accuracy, and the rough data models provided meaningful information of the variables. Ng and Liu (2000) applied feature selection to create predictive models of customer retention for a confidential service provider using data mining on the data that had 45,000 transactions per day. Smith et al. (2000) applied neural networks, clustering, and decision trees to the various stages of insurance claims

patters. The authors found that neural networks provided the best results for the test set. Bin et al. (2007) used decision trees to predict customer churn in the telecommunication market in China. In some of the trained models, the recall and precision rate for the test set were 95% and 82% respectively.

Ngai et al. (2008) presented a literature review of papers published in peer-reviewed publications on the topic of customer relationship management and data mining. They found that out of 87 articles, 54 articles (61.2%) were on customer retention, which possibly means that in the domain of customer relationship management, researchers are applying data mining techniques to the customer retention area than other areas. These techniques included clustering sequence discovery, neural networks, decision tress, logistic regression, and association rules.

2.4.1 Assessing the State of the Art

Table 2.14 lists techniques used in the studied literature, where the cohort sizes were available, along with the reported performance measures. Some of the notable points in the literature were:

- Witten and Frank (2005) recommends the practice to divide the data into a train and test set, learn on the train set, then assess the learned theory on the test set. If a theory is tested on the train data itself, this test can over-estimate theory performance.

  For example, Glynn et al. (2003) result of Table 2.14 seems impressive with a 83% accuracy on a data set with a 49.08% retention rate; however, these results were obtained using the training data, whereas the test should have been repeated using some *hold-out* test set.

- All the regression studies from 1971 to 1999 reported R^2 values under 0.6. The maximum value of R^2 is one and R^2 values under 0.6 indicate very weak predictive abilities.

- The accuracy reported in the literature were very close to the *ZeroR* theoretically lower-bound on performance. ZeroR is a baseline classifier that simply returns the majority class. For example, Herzog (2006) studied a data set with a 83.5% retention rate, therefore, ZeroR would be correct in 83.5% of cases. The 85.4% accuracy of Herzog's data miners was very close to the ZeroR lower-bound.

The last three results of Table 2.14 did not report their accuracies. However, these can be calculated in the following way. Let A, B, C, D be the true negatives, false negatives, false positives, and true positives respectively of a predictor that some student will attend some year of university. Zhang and Zhang (2007)

and Menzies et al. (2007b) defined the relationships among various performance measures:

$$\begin{aligned} pd = recall &= D(B+D) && (2.1)\\ pf = false\ alarm &= C/(A+C) && (2.2)\\ prec = precision &= D/(C+D) && (2.3)\\ acc = accuracy &= (A+D)/(A+B+C+D) && (2.4)\\ neg/pos &= (A+C)/(B+D) && (2.5) \end{aligned}$$

- "Recall" measures how much of the target was found.
- The "false alarm" rate measures how what fraction of non-targets triggered the learned theory.
- "Precision" comments on how many targets are found in the data selected by the theory.
- "Accuracy" comments on how many of the targets and non-targets were accurately labeled by the learned theory.

In an ideal result, we can obtain high recall, low false alarms, high precision, and high accuracies. As discussed by Zhang and Zhang (2007) and Menzies et al. (2007b), recall, false alarm, accuracy, and precision values are inter-related; thus, high recall, low false alarms, high precision, and high accuracies are not possible. These inter-relationships are shown below:

$$\left(prec = \frac{D}{D+C} = \frac{1}{1+\frac{C}{D}} = \frac{1}{1+\frac{neg}{pos}\cdot\frac{pf}{recall}}\right) \Rightarrow \left(pf = \frac{pos}{neg}\cdot\frac{(1-prec)}{prec}\cdot recall\right) \quad (2.6)$$

Using these equations, missing performance measures can be found given other measures.

$$\begin{aligned} D &= recall * pos && (2.7)\\ C &= pf * neg && (2.8)\\ A &= C * 1/(pf-1) && (2.9)\\ acc &= (A+D)/(neg+pos) && (2.10) \end{aligned}$$

Using these equations, missing performance measures were found of the last three results of Table 2.14:

- In Atwell et al. (2006), the the precision varied from 73% to 88%. Using these equations, estimated false alarm (pf) values were between 2% and 8%

(recall values of 65% to 90% were assumed). It is very rare to achieve such very low false alarm rates, especially from noisy data relating to student retention. Hence, the Atwell et al. (2006), results are somewhat surprising.

- In DeLong et al. (2007), the precision varied from 57% to 60%. From these equations, estimated false alarm rates were in the range of 49% to 63% (recall values of 65% to 90% were assumed). For this type of study, these were very high false alarm rates.

- In Pittman (2008), the reported precision varied from 44% to 63%. For $0.78 \leq acc \leq 0.81$, $neg = 17139$ and $pos = 21136 - neg$, the equations obtained $prec \leq 50$. Thus the reported precision values ≥ 50 were unattainable and should be reviewed.

2.5 Summary

Retention research goes back to early 70's, and it is still ongoing; however, with the higher computing speeds and new algorithms, data mining research is giving a new perspective to this century-old problem. Different researchers built predictive models based on the theoretical framework of Spady (1970), Tinto (1975), and Bean (1979). These theoretical models concluded that student's integration with the university along with the past academic performance were key areas for student retention. Some other important variables were: high school GPA, ACT/SAT scores, on/off campus housing, socio-economic status, and parent's education. Table 2.14 lists performance obtained and techniques used in the literature, where the cohort size was available.

Although the use of data mining in the field of education is at a nascent stage, few researchers have applied data mining in the areas of graduation, enrollment management, and retention. This data mining research, however, lacks in-depth analysis of different learners, discretization methods, feature subset evaluation, and building high performance and explanation systems. Figure 2.8 provides a visual perspective on the terms discussed in the studied papers of this literature review.

Figure 2.8: Tag Cloud of the Papers Studied in the Literature Review. (Bolder and bigger fonts represent high frequency terms)

Author (Year)	Notes	Cohort Size	Retained (#)	Retained (%)	Measure of Accuracy	Coeffes Used?	Techniques Used
Spady (1971)		683	615	90.04%	R^2 of .3132 for men and .3879 for women	Yes	Multiple regression
Bean (1980)		906	769	84.88%	R^2 of .22 for women and 0.09 for men	Yes	Multiple regression
Terenzini (1980)	study 1	379	60	15.80%	R^2 of .246	Yes	discriminate analyses
	study 3	518	428	82.63%	R^2 of .256	Yes	Multiple regression
	study 5	763	673	88.20%	R^2 of .309	Yes	discriminate analyses
	study 6	763	673	88.20%	R^2 of .476 for men and .553 for women	Yes	discriminate analyses
Stage (1989)		323	294	91.00%		Yes	Logistic regression
Dey and Astin (1993)		947	152	16.00%	Multiple R 0.354, 0.351, and 0.323	Yes	logit, probit, and regression
Murtaugh et al. (1999)		8667	5200	60%	estimated ret prob 59.3%	Yes	Survival Analysis/ Hazard regression
Bresciani and Carson (2002)		3535	3121	88.30%	R^2 of 0.022	Yes	Logistic regression
Glynn et al. (2003)	any dropout; not only first-year; accuracies based on the training data	3244	1592	49.08%	overall accuracy of 83%	Yes	Logistic regression
Herzog (2005)		5261	4014	76.30%	77.4% accuracy	Yes	Logistic regression
		4298	3314	77.10%		Yes	
		4671	4040	83.50%	85.4% accuracy	Yes	
Sujitparapitaya (2006)		2,444	1943	79.50%	81.6% accuracy on training; 80.7% on validation	Yes	Logistic regression
		2,445	1994	79.50%	83.9% accuracy on training; 82.1% on validation		Neural Network
		2,445	1994	79.50%	85.5% on training; 84.4% on validation		C4.5
Herzog (2006)		8,018	6037	75.29%	accuracy close to 75%		Neural Networks; CHAID, C4.5, CR&T; Logistic regression
Atwell et al. (2006)	training	3,829	3149	82.24%	precision for drop-outs 91, 84, 84, 78		decision trees (entropy, chi-sq, gini) and logistic regression
	test	5,990	4,881	81.49%	precision for drop-outs 88, 82,82,73		
DeLong et al. (2007)				50%	precision varied from 57% to 60%		AdaBoost M1 with Decision Stumps
Pittman (2008)		21136	17139	81.10%	overall accuracy of 78-81%; not-retained precision from 44-63%		Logistic regression, neural network, bayes, J48

Table 2.14: Techniques and Accuracies Reported in Literature

Chapter 3

Data and Experiment

> It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.
>
> Richard Feynman

3.1 Data

Data used in this study were from Kent state university , a mid-size public university, and were extracted from the student information system on official census dates. These data consisted all first-year freshmen's demographic, academic, and financial aid information (more than 100 attributes), as of the census reporting dates (after two weeks of semester starting date). The attributes used in this study along with the descriptive statistics are given in Table 3.1 and 3.2.

Attribute	Description	Min	1st Qu.	Median	Mean	3rd Qu.	Max.	Missing
AGE	Age Of Student	0	18	18	18.5	19	64	
HS_GPA	High School Gpa	0	2.73	3.1	3.08	3.49	4.92	
HS_RANK	High School Rank	0	27	70	93.4	136	2256	
HS_SIZE	High School Class Size	0	2	135	164	270	9016	
HS_PERCENT	High School Percentile	0	38	59	56	78	100	
StdsFromSameHS	Number Of Students From Same Hs	1	4	12	19.3	25	113	
COMP_READ	Compass Read Score	0	69	84	71.4	91	99	
COMP_WRITE	Compass Write Score	0	47	78	65.4	92	991	
ACT_ENGL	ACT English Score (old)	0	0	0	0.0497	0	34	
ACT_MATH	ACT Math Score (old)	0	0	0	0.0485	0	32	
ACT_COMP	ACT Comprehensive Score (old)	0	0	0	0.0523	0	32	
ACT1_ENGL	ACT English Score (new)	0	15	20	17.4	23	440	
ACT1_MATH	ACT Math Score (new)	0	16	19	17.3	23	510	
ACT1_COMP	ACT Comprehensive Score (new)	0	17	20	17.7	23	35	
SAT_VERB	SAT Verbal Score	0	0	0	149	420	800	
SAT_TOT	SAT Total Score	0	0	0	298	860	1580	
ACTEQUIV	ACT Equivalent Of The SAT Score	10	10	10	13.4	18	35	
MaxACT	Max Of ACT Score And ACT Equivalent	10	18	21	20.7	24	35	
CUR_ERLHRS	Total Enrolled Hours	1	14	15	14.8	16	22	
AvgClassSize	Average Class Size Of All The Courses For A Student	1	51	83	84.7	114	484	
LrgEnrlClssCnt	Count Of Large Class Size Courses	0	1	1	1.4	2	5	
DiffCrsesCnt	Count Of Difficult Courses	0	1	2	1.92	3	6	
PercentileRankHSGPA	Percentile Of Hs Gpa Among All Freshmen	0	24.5	49.5	49.6	74.8	100	
PercentileRankMaxACT	Percentile Of Max ACT Among All Freshmen	0	21.8	44.7	46.3	72.1	100	
TotalClasses	Total Number Of Classes	1	5	6	5.81	6	11	
NoTenureFacCnt	Count Of Courses Taught By No Rank Faculty	0	3	4	3.9	5	11	
NoTenureFacRatio	Ratio Of Courses Taught By No Rank Faculty To The Total Courses	0	50	66.7	66.1	100	100	
NTTFacCnt	Count Of Courses Taught By Ntt Faculty	0	0	1	0.937	2	7	

Table 3.1 continued on next page

Attribute	Description	Min	1st Qu.	Median	Mean	3rd Qu.	Max.	Missing
NTTFacRatio	Ratio Of Courses Taught By Ntt Faculty To The Total Courses	0	0	16.7	16.6	33.3	100	
TTFacCnt	Count Of Courses Taught By Tenured/tenure-track Faculty	0	0	1	0.97	2	7	
TTFacRatio	Ratio Of Courses Taught By Tenured/tenure-track Faculty To The Total Courses	0	0	16.7	17.2	33.3	100	
FacExpLT1Ratio	Ratio Of Courses Taught By Faculty W/ Less Than 1 Yr Experience To The Total Courses	0	0	0	7.04	16.67	100	3802
FacExp1to5Ratio	Ratio Of Courses Taught By Faculty W/ Experience Between 1 & 5 Yrs To The Total Courses	0	16.7	33.3	31.3	42.9	100	3802
FacExp6to10Ratio	Ratio Of Courses Taught By Faculty W/ Experience Between 6 & 10 Yrs To The Total Courses	0	0	16.7	14.8	20	100	3802
FacExp11to15Ratio	Ratio Of Courses Taught By Faculty W/ Experience Between 11 & 15 Yrs To The Total Courses	0	0	16.7	14.9	20	100	3802
FacExp16to20Ratio	Ratio Of Courses Taught By Faculty W/ Experience Between 16 & 20 Yrs To The Total Courses	0	0	16.7	14	20	100	3802
FacExp21to25Ratio	Ratio Of Courses Taught By Faculty W/ Experience Between 21 & 25 Yrs To The Total Courses	0	0	0	7.12	16.67	100	3802
FacExp25to30Ratio	Ratio Of Courses Taught By Faculty W/ Experience Between 25 & 30 Yrs To The Total Courses	0	0	0	2.14	0	100	3802
FacExpGT31Ratio	Ratio Of Courses Taught By Faculty W/ Experience Greater Than 31 Years To The Total Courses	0	0	0	4.64	0	100	3802
DistanceFromCampus	Distance Between Home And Campus	0	15.3	29.4	63.5	65.5	4577.6	207

Table 3.1 continued on next page

Attribute	Description	Min	1st Qu.	Median	Mean	3rd Qu.	Max.	Missing
TotalFinAidOffered	Total Financial Aid Offered	15	2813	5800	5909	8600	29136	17800
FinAidAwardType_G	Financial Aid Amount Of Grants	21	1237	2006	2561	3473	20163	25304
FinAidAwardType_J	Financial Aid Amount Of Jobs	25	900	900	962	1000	3780	33357
FinAidAwardType_L	Financial Aid Amount Of Loans	22	1750	4000	4662	7313	17250	21576
FinAidAwardType_S	Financial Aid Amount Of Scholarship	15	750	1100	1861	2500	23063	26445
FinAidAwardType_W	Financial Aid Amount Of Waiver	303	3441	3977	4114	4215	12263	33213
FinAidSTUDENT_AG	Student's Adjusted Gross Income	-21651	2668	4694	6798	7630	724724	13309
FinAidSTUDENT_WA	Student's Wage	-20527	1118	3564	4264	6009	561500	13281
FinAidSPOUSE_WAG	Spouse's Wages	0	0	0	992	0	223894	26664
FinAidPARENT_AGI	Parent's Adjusted Gross Income	-551115	41447	68549	74100	95422	930323	11246
FinAidFATHER_WAG	Father's Income	-192939	17795	43762	46337	64585	732975	15731
FinAidMOTHER_WAG	Mother's Income	-55755	8173	23000	25881	38472	533395	14602

Table 3.1: Summary of Numeric Attributes

Attribute	Description	Values						
RESIDENCY	Residency Status	N: 4013	R:29699					
STATE_ORIG	Home State	OH :29584	PA : 2208	NY : 466	XX : 161	MI : 156	(Other): 1036	NA's : 101
COUNTYCODE	County Code	SUM : 5756	CYH : 5348	PRT : 3346	STK : 1988	LKE : 1195	(Other):13114	NA's : 2965
INTERNATAL	International Student Indicator	N :22295	Y : 107	NA's:11310				
GENDER	Gender	F:20769	M:12943					
ETHNIC	Ethnicity Code	A: 532	B: 3028	F: 242	H: 458	N: 129	W:28383	Z: 940
HS_CODE	High School Code	364845 : 935	362778 : 779	361770 : 548	362650 : 425	364940 : 397	363487 : 354	(Other):30274
ADMIT_DEG	Admit Degree	XX :17644	BS : 6937	BA : 5650	BSE : 2743	BFA : 435	BM : 223	(Other): 80
ADMIT_MAJ	Admit Major	EXPL : 6111	PBMT : 1755	PJMC : 1754	FPAG : 1458	PNUR : 1252	PSYC : 1210	(Other):20172
FTPT	Full-Time/Part-Time Indicator	FT:33000	PT: 712					
MAJOR_1	Current Major	EXPL : 6153	PBMT : 1750	PJMC : 1743	FPAG : 1478	PNUR : 1242	PSYC : 1206	(Other):20140
PreMajorInd	Pre-Major Indicator	N:21216	Y:12496					
ExploratoryMajorInd	Exploratory Major Indicator	N:27559	Y: 6153					
DEGREE_1	Current Degree	XX :17724	BS : 6911	BA : 5629	BSE : 2744	BFA : 437	BM : 193	(Other): 74
LIVEONCAMP	On-Campus Indicator	N: 7583	Y:26129					
EnrolledinSummer	Enrolled in Summer Indicator	N:32309	Y: 1403					
FacExpLT1Cnt	Count Of Courses Taught By Faculty W/ Less Than 1 Yr Experience	0.0 :19986	1.0 : 8092	2.0 : 1575	3.0 : 220	4.0 : 34	(Other): 3	NA's : 3802
FacExp1to5Cnt	Count Of Courses Taught By Faculty W/ Experience Between 1 & 5 Yrs	2.0 :9221	1.0 :9068	3.0 :5173	0.0 :4059	4.0 :1810	(Other): 579	NA's :3802
FacExp6to10Cnt	Count Of Courses Taught By Faculty W/ Experience Between 6 & 10 Yrs	0.0 :12311	1.0 :11622	2.0 : 4574	3.0 : 1134	4.0 : 221	(Other): 48	NA's : 3802
FacExp11to15Cnt	Count Of Courses Taught By Faculty W/ Experience Between 11 & 15 Yrs	0.0 :12289	1.0 :11456	2.0 : 4645	3.0 : 1209	4.0 : 266	(Other): 45	NA's : 3802
FacExp16to20Cnt	Count Of Courses Taught By Faculty W/ Experience Between 16 & 20 Yrs	0.0 :13433	1.0 :10642	2.0 : 4412	3.0 : 1125	4.0 : 249	(Other): 49	NA's : 3802
FacExp21to25Cnt	Count Of Courses Taught By Faculty W/ Experience Between 21 & 25 Yrs	0.0 :20316	1.0 : 7417	2.0 : 1664	3.0 : 391	4.0 : 101	(Other): 21	NA's : 3802
FacExp25to30Cnt	Count Of Courses Taught By Faculty W/ Experience Between 25 & 30 Yrs	0.0 :26556	1.0 : 3143	2.0 : 170	3.0 : 37	4.0 : 4	NA's: 3802	
FacExpGT31Cnt	Count Of Courses Taught By Faculty W/ Experience Greater Than 31 Years	0.0 :23101	1.0 : 5917	2.0 : 799	3.0 : 87	4.0 : 6	NA's: 3802	
FinAidOfferedInd	Financial Aid Offered Indicator	Y :15912	NA's:17800					

Table 3.2 continued on next page

Attribute	Description	Values						
FirstGenInd	First Generation Indicator	N :14110	Y : 9804	NA's: 9798				
FinAidDEPENDENCY	Dependency Status 1:59	D :21135	I : 2774	NA's: 9801				
FinAidFATHER_ED	Father's Education Level	1.0 : 598	2.0 :11434	3.0 :10135	4.0 : 1549	NA's: 9996		
FinAidMOTHER_ED	Mother's Education Level	1.0 : 385	2.0 :11762	3.0 :10552	4.0 : 1108	NA's: 9905		
FinAidSTUDENT_MA	Student's Martial Status 3:59	M : 404	S : 31	U :23460	NA's: 9813			
FinAidSTUDENT_HO	Student's Household Size	0.0 :18985	1.0 : 1939	4.0 : 811	2.0 : 754	3.0 : 728	(Other): 697	NA's : 9798
FinAidSTUDENT_TA	Student's Tax Form Type	1.0 : 3146	2.0 :13805	3.0 : 27	4.0 : 40	NA's:16694		
FinAidPARENT_MAR	Parents's Martial Status 504:59:00	M :15598	S : 4365	U : 990	W : 536	NA's:11718		
FinAidPARENT_HOU	Parents's Household Size	4.0 :8502	3.0 :5338	5.0 :4817	2.0 :1837	6.0 :1610	(Other):1810	NA's :9798
FinAidPARENT_TAX	Parents's Tax Form Type	1.0 :15529	2.0 : 3584	3.0 : 20	4.0 : 34	NA's:14545		
ANTH18	Enrolled in Anthropology Course	N:33117	Y: 595					
BSCI10	Enrolled in Biological Science-Course	N:32404	Y: 1308					
CHEM10	Enrolled in Chemistry Course	N:31233	Y: 2479					
ENG10	Enrolled in English Course	N:31903	Y: 1809					
ENG11	Enrolled in English Course	N:33256	Y: 456					
GEOL11	Enrolled in Geology Course	N:33280	Y: 432					
LEST16	Enrolled in Leisure Studies Course	N:33527	Y: 185					
MATH10	Enrolled in Math 100 Level Course	N:28094	Y: 5618					
MATH11	Enrolled in Math 110 Level Course	N:26925	Y: 6787					
MATH12	Enrolled in Math 120 Level Course	N:31695	Y: 2017					
MATH14	Enrolled in Math 14 Level Course	N:32488	Y: 1224					
PEP15	Has taken Physical Ed. Course?	N:33081	Y: 631					

Table 3.2: Summary of Categorical Attributes

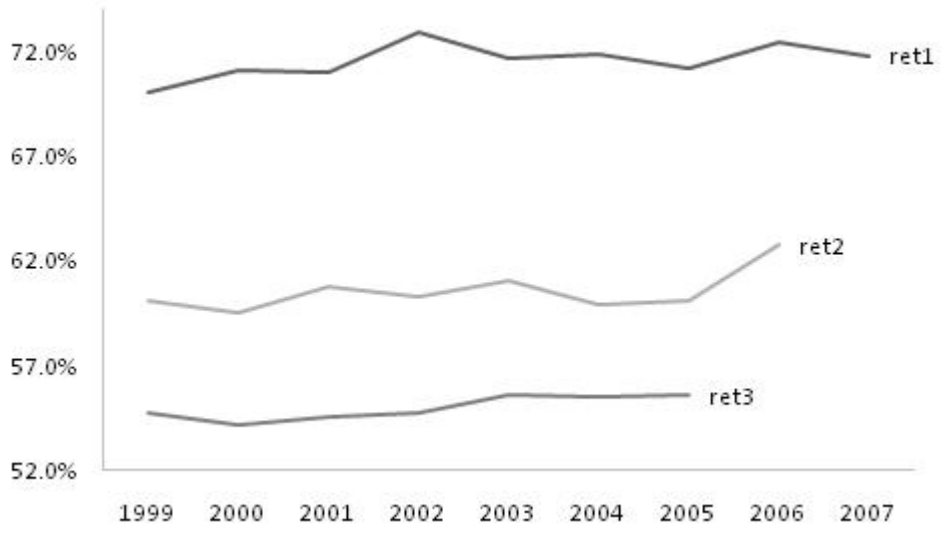


Figure 3.1: Retention Rates by Cohort Years. RET1 is first-year retention; RET2 is second-year retention; and RET3 is third-year retention

As the higher education administrators may design effective policies when the students begin their studies, it is important to note that the emphasis of this study was on detecting patterns based only on the first-term data, and that too only beginning of the term data. Three dependent variables were created: RET1, if the student returned after one year; RET2, if the student returned after two years; and RET3, if the student returned after three years. These retention rates by the cohort-years are given in Figure 3.1; note that there was no significant change in the retention rates over the years.

The overall distribution of these dependent variables is given in Table 3.3. For the studied time period, the overall first-year retention rate was 71.3%, the second-year persistence rate was 60.4%, and the third-year persistence rate was 54.8%.

	RET1		**RET2**		**RET3**	
	Count	Percentage	Count	Percentage	Count	Percentage
retained=**Y**	24,039	71.3%	18,055	60.4%	14,362	54.8%
retained=**N**	9,673	28.7%	11,857	39.6%	11,854	45.2%
Total	33,712	100%	29,912	100%	26,216	100%

Table 3.3: Distribution of Dependent Variables

In the Integrated Postsecondary Education Data System (IPEDS) , for the U.S. only, degree-granting, Doctoral degree offering, 4-year and above institutes

(excluding University of Phoenix-Online Campus), and cohort size greater than 3,000, we found that the full-time freshmen retention rate had a range from 59% to 96%, and the cohort size had a range from 3,117 to 8,025. In this list of institutions, Kent state university ranked 38 in the full-time retention percentage and 26 in the cohort size (Department of Education, 2010). Thus, Kent state data are representative of other similar size universities, and the data mining approach could be generalized to other universities.

3.1.1 Attribute Groups

Based on the research present in literature, some attribute groups were created to compare the performance between the attributes selected by algorithms and the attributes grouped by active hypotheses. Table 3.4 lists attributes that were grouped together under each hypothesis.

If the results of learners using algorithmically selected attributes are better than the results of learners using attributes from the active hypotheses, then these results are evidence against those hypotheses. Active hypotheses identified in the literature were:

H1: The *financial aid* hypothesis. Sanjeev and Zytkow (1995) found that no amount of financial aid influenced students to enroll for more terms; whereas Herzog (2005) found that upper-income students had reduced dropout odds compared to those from middle and lower incomes. According to John (2000), "the research literature remains ambiguous" regarding the influence financial aid on recruitment and retention.

H2: The *academic performance* hypothesis. Although there is no doubt that high school GPA and high school preparedness has a significant impact on persistence, researchers have often questioned the effects of standardized college entrance examinations (ACT/SAT). Waugh et al. (1994) found that SAT and ACT scores had no relationship with retention, whereas Murtaugh et al. (1999) found that SAT scores had some predictive value, although inferior compared to high school GPA. DesJardins et al. (2002) noted that high GPA lowered the risk of dropout, but the effect diminished over time, and that the financial aid was an insignificant factor for increasing graduation, however, it indeed reduced the student stopout. In their comprehensive literature review, Lotkowski et al. (2004) found that high school GPA had the strongest relationship with college retention in the academic factors, but ACT assessment scores had a moderate impact.

H3: The *faculty tenure and experience* hypothesis. Ehrenberg and Zhang (2005) found that for every 10 percentage point increase in the percentage of part-time faculty and not on tenure-track full-time faculty, there was a 3-5 per-

centage point reduction in the institution's graduation rate. Jacoby (2006) found similar results at community colleges that increase in the ratio of part-time faculty had a negative impact on the graduation rates.

3.1.2 Data Exploration

To find any existing patterns in the data, data exploration using visualization was performed. Figure 3.2 shows retention possibilities over three years with the averages of ACT English score, ACT Math score, Max of ACT and SAT scores, HS GPA, Percentile Rank of HS GPA amongst other freshmen, and Parent's adjusted gross income (AGI) . Although the average values of these attributes were close to each other for both first-generation and non-first-generation student, there was a vast gap in the AGI of parents. It also shows a upward trend of averages all the attributes, meaning that students who succeeded to their third-year had higher average scores and higher parents' AGI. The average percentile rank of HS GPA plot is particularly interesting, because it shows approximately 20 points difference between those who did not enroll in three years and those who did enroll. This attribute measures the students' percentile of their HS GPA amongst the incoming freshmen cohort; this attribute is important as it measures the academic performance gap amongst the freshmen.

After discretizing the data using minimum description length criterion (Fayyad and Irani, 1992), visual data analysis was performed to find any interesting patterns in the data before running any learners. Figure 3.3 shows the third-year retention percentage for different education levels of parents, and it is clear that retention percentage increases with the parent's education level.

As shown in Figure3.4, parent's household size had an effect on third-year retention. This trend is counter intuitive, as bigger household would suggest less parental attention and sharing resources.

Figure 3.5 shows the positive effect of high school GPA on the third-year retention percentage; and for the tax form type 3 and 4, the retention percentages are very high compared to the retention percentages for tax form type 1 and 2.

3.2 Building the Experiment

As discussed in Section 2.4.1, learning algorithms should perform at least better than the ZeroR learner, which simply returns the majority class. For this study, the lower bounds were (given in Table 3.3):

- For first year retention: 71.3%.
- For second year retention: 60.4%.

Category	Financial Aid	**C**ategory	Performance Indicators	**C**ategory	Faculty Type & Experience
Attribute	**D**escription	**A**ttribute	**D**escription	**A**ttribute	**D**escription
FinAidAwardType_G	Financial aid amount of grants	ACT_COMP	ACT comprehensive score (old)	FacExpLT1Cnt	Count of courses taught by faculty [CCTF] w/ less than 1 yr experience
FinAidAwardType_J	Financial aid amount of jobs	ACT_ENGL	ACT english score (old)	FacExpLT1Ratio	Ratio of courses taught by faculty [RCTF] w/ less than 1 yr experience to the total courses
FinAidAwardType_L	Financial aid amount of loans	ACT_MATH	ACT math score (old)	FacExp1to5Cnt	CCTF w/ experience between 1 & 5
FinAidAwardType_S	Financial aid amount of scholarship	ACT1_COMP	ACT comprehensive score (new)	FacExp1to5Ratio	RCTF w/ experience between 1 & 5 to the total courses
FinAidAwardType_W	Financial aid amount of waiver	ACT1_ENGL	ACT english score (new)	FacExp6to10Cnt	CCTF w/ experiencAe between 6 & 10
FinAidDEPENDENCY	Dependency status	ACT1_MATH	ACT math score (new)	FacExp6to10Ratio	RCTF w/ experience between 6 & 10 to the total courses
FinAidFATHER_ED	Father's education level	ACTEQUIV	ACT equivalent of the sat score	FacExp11to15Cnt	CCTF w/ experience between 11 & 15
FinAidFATHER_WAG	Father's income	MaxACT	Max of ACT score and ACT equivalent	FacExp11to15Ratio	RCTF w/ experience between 11 & 15 to the total courses
FinAidMOTHER_ED	Mother's education level	COMP_READ	Compass read score	FacExp16to20Cnt	CCTF w/ experience between 16 & 20
FinAidMOTHER_WAG	Mother's income	COMP_WRITE	Compass write score	FacExp16to20Ratio	RCTF w/ experience between 16 & 20 to the total courses
FinAidOfferedInd	Financial aid offered indicator	SAT_TOT	SAT total score	FacExp21to25Cnt	CCTF w/ experience between 21 & 25
FinAidPARENT_AGI	Parent's adjusted gross income	SAT_VERB	SAT verbal score	FacExp21to25Ratio	RCTF w/ experience between 21 & 25 to the total courses
FinAidPARENT_HOU	Parents's household size	HS_CODE	High school code	FacExp25to30Cnt	CCTF w/ experience between 25 & 30
FinAidPARENT_MAR	Parents's martial status	HS_GPA	High school gpa	FacExp25to30Ratio	RCTF w/ experience between 25 & 30 to the total courses
FinAidPARENT_TAX	Parents's tax form type	HS_PERCENT	High school percentile	FacExpGT31Cnt	CCTF w/ experience greater than 31 years
FinAidSPOUSE_WAG	Spouse's wages	HS_RANK	High school rank	FacExpGT31Ratio	RCTF w/ experience greater than 31 years to the total courses
FinAidSTUDENT_AG	Student's adjusted gross income	HS_SIZE	High school class size	NoTenureFacCnt	Count of courses taught by no rank faculty
FinAidSTUDENT_HO	Student's household size	RankHSGPA	Percentile of hs gpa among all freshmen	NoTenureFacRatio	Ratio of courses taught by no rank faculty to the total courses
FinAidSTUDENT_MA	Student's martial status	RankMaxACT	Percentile of max ACT among all freshmen	NTTFacCnt	Count of courses taught by ntt faculty
FinAidSTUDENT_TA	Student's tax form type	ANTH18	Enrolled in anthropology course	NTTFacRatio	Ratio of courses taught by ntt faculty to the total courses
FinAidSTUDENT_WA	Student's wage	BSCI10	Enrolled in biological science course	TTFacCnt	Count of courses taught by tenured/tenure-track faculty
FirstGenInd	First generation indicator	CHEM10	Enrolled in chemistry course	TTFacRatio	Ratio of courses taught by tenured/tenure-track faculty to the total courses
TotalFinAidOffered	Total financial aid offered	ENG10	Enrolled in English course		
		ENG11	Enrolled in English course		
		GEOL11	Enrolled in geology course		
		LEST16	Enrolled in leisure studies course		
		MATH10	Enrolled in math 100 level course		
		MATH11	Enrolled in math 110 level course		
		MATH12	Enrolled in math 120 level course		
		MATH14	Enrolled in math 14 level course		
		PHY11	Enrolled in physics 11 level course		
		PEP15	Enrolled in physical ed 15 level course		

Table 3.4: List of Attributes by Stated Hypotheses

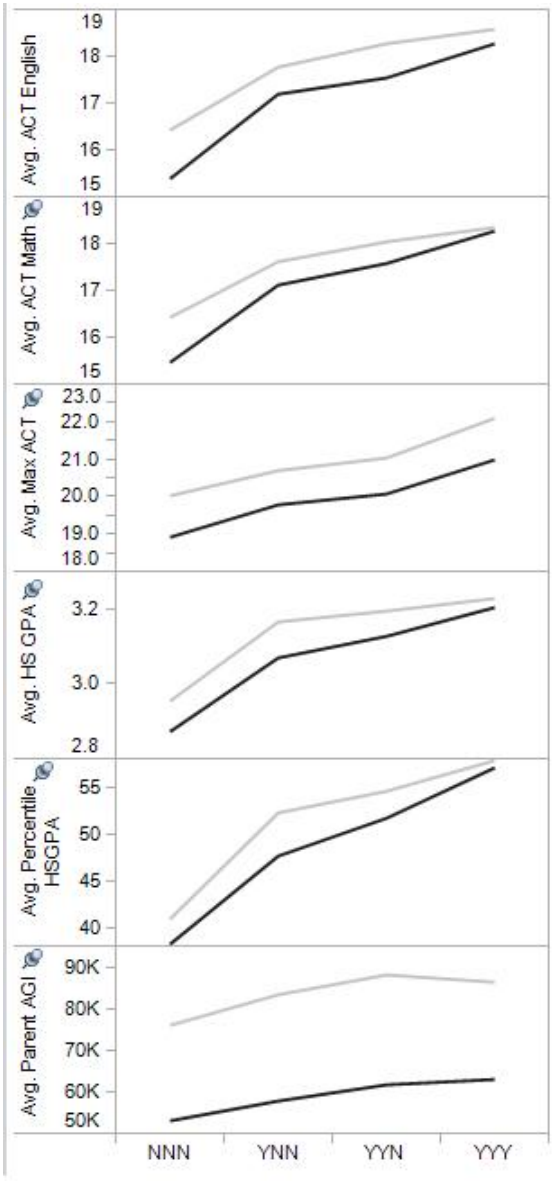


Figure 3.2: Retention over the years by Avg. ACT English score, Avg. ACT Math score, Avg. of Max of ACT and SAT scores, Avg. HS GPA, Avg. Percentile Rank of HS GPA amongst other freshmen, and Avg. Parent's AGI. Dark line represents first-generation student and lighter line represents non-first-generation student.

- For third year retention: 54.8%.

To find better performing learners the following approach was used:

- Remove spurious attributes using *feature subset selection*;

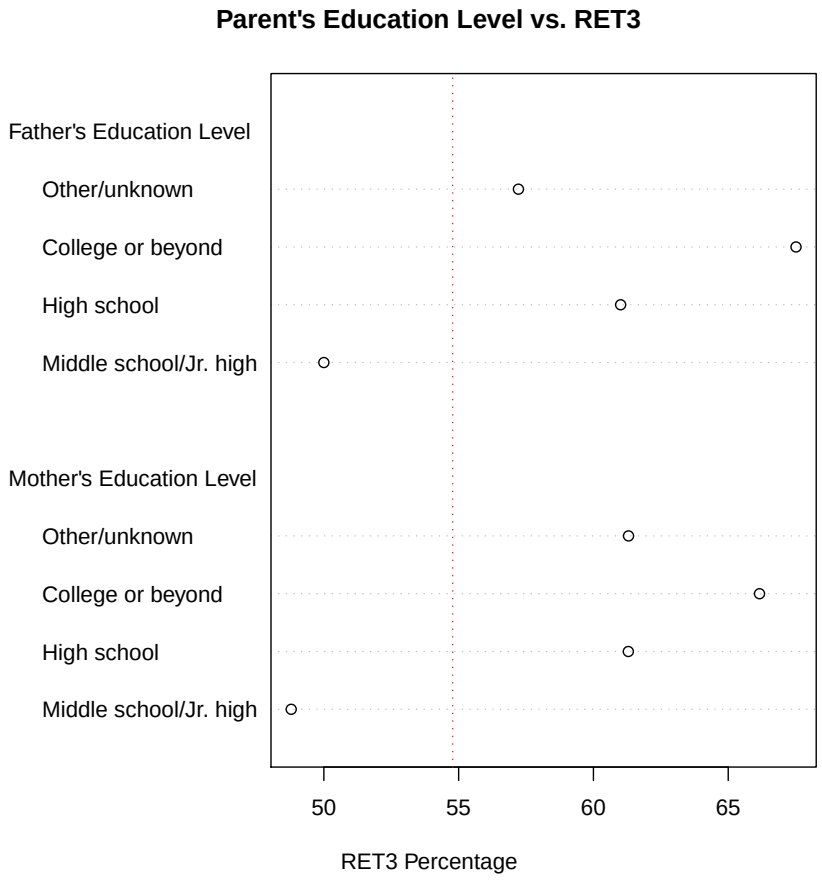


Figure 3.3: Parent's education level vs. RET3 percentage. Red dashed line represents the baseline RET3 percentage

- Explore *a large range of classifiers*;
- Assess the learned theories by their *variance*, as well as their *median* performance.
- Asses the learned theories by their *variance*, as well as their *median* performance.
- Study the *delta* of student factors *between* those who stay and those who are retained.

3.2.1 Feature Subset Selection

Table 3.4 lists a sample of the 103 attributes used in this study. To remove unnecessary attributes, which did not contribute to the prediction of retention, before applying any learners, *attribute selection* was explored.

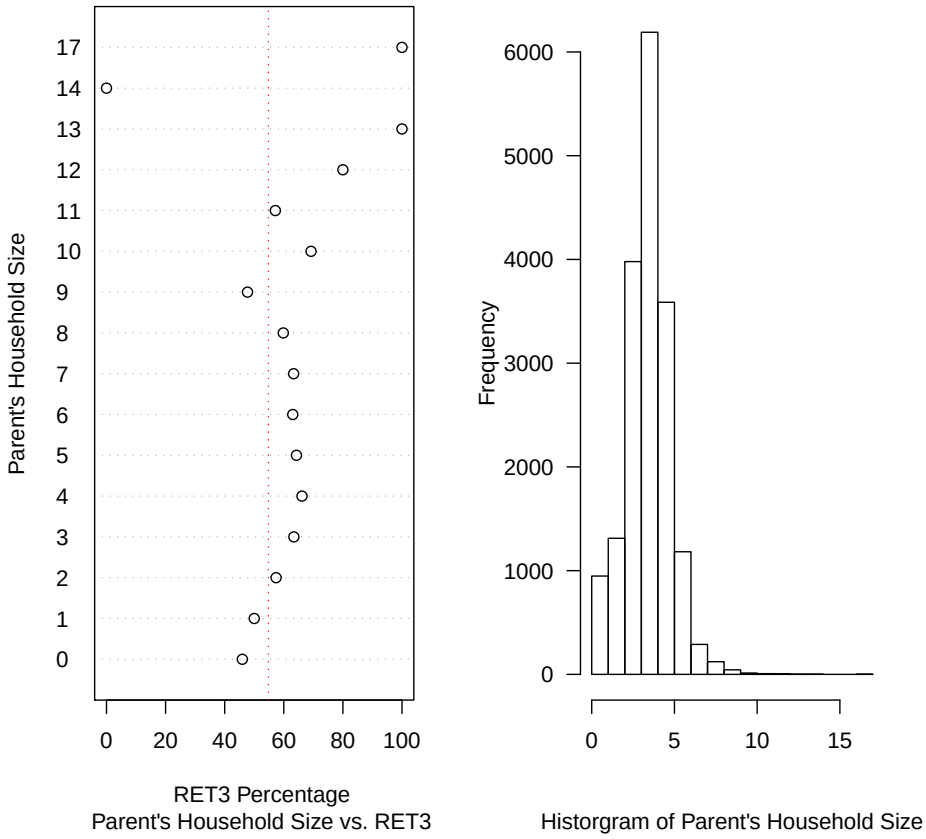


Figure 3.4: Parent's household size vs. RET3 percentage (left), and distribution of parent's household size (right). Red dashed line represents the baseline RET3 percentage. Note that the "0" household size denotes missing and auto-computed entries.

Selected attributes can comment on the hypotheses presented in 3.1.1, because if the attributes from these hypotheses are missing from the selected attributes, it can be concluded that these attributes do not add any value to the performance of the predictor.

In this experiment, 103 attributes were ranked from most informative to least informative. Theories were built using the top $n \in \{5, 10, .., 100, 103\}$ ranked attributes. Attributes were then discarded if adding them in did not improve the performance of the retention predictors.

The attributes were ranked using one of four methods: CFS, Information Gain, chi-squared, and One-R.

- *Correlation-based feature selection* constructs a matrix of feature to feature,

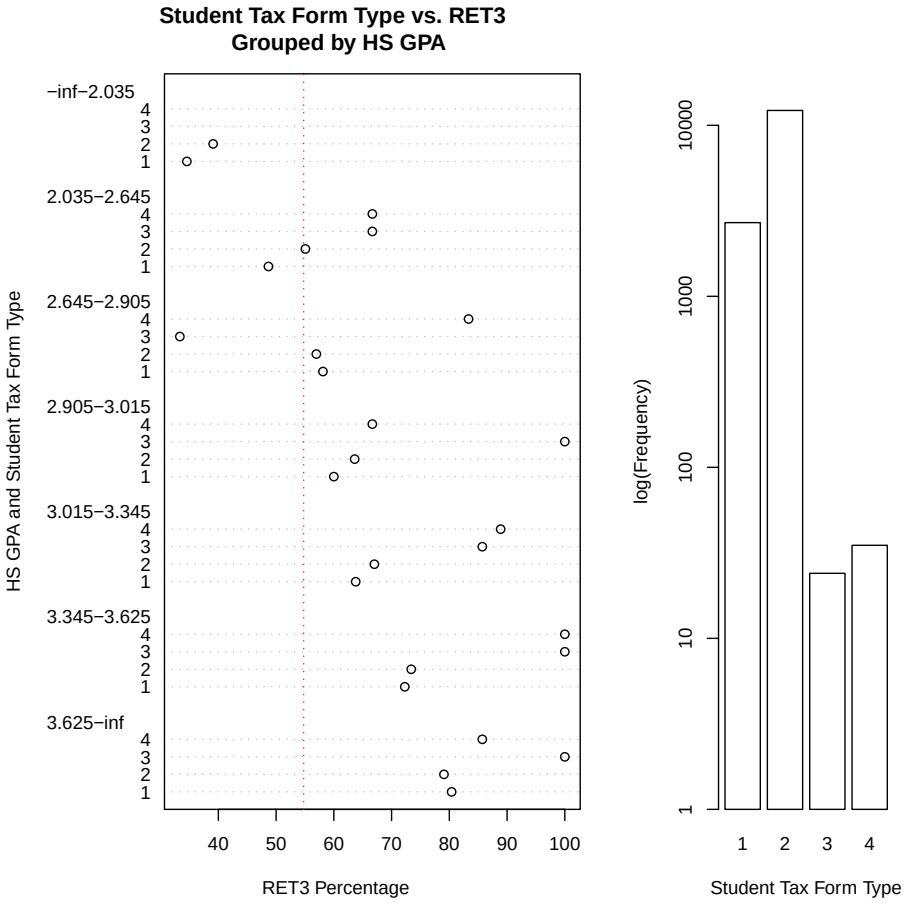


Figure 3.5: Student tax form type vs. RET3 percentage, grouped by high school GPA. Red dashed line represents the baseline RET3 percentage. On the right, a log-frequency histogram of student tax form type.

and feature-to-class correlations (Hall, 2000). CFS uses a best first search by expanding the best subsets until no improvement is made, in which case the search falls to the unexpanded subset having the next best evaluation until a subset expansion limit is met.

- *Information Gain* uses an information theory concept called *entropy*. Entropy measures the amount of uncertainty, or randomness, that is associated with a random variable. Thus, high entropy can be seen as a lack of purity in the data. Information gain, as described in Mitchell (1997) is an expected reduction of the entropy measure that occurs when splitting examples in the data using a particular attribute. Therefore an attribute that has a high purity (high information gain) is better at describing the data than the one that has a low purity. The resulting attributes are then ranked by

sorted their information gain scores in a descending order.

- The *chi-squared* statistic is used in statistical tests to determine how distributions of variables are different from one another (David Moore, 2006). Thus, the chi-squared statistic can evaluate an attribute's worth by calculating the value of this statistic with respect to a class. Attributes are then ranked based on this statistic.

- The *One-R* classifier, can be used to deliver top-ranking attributes. One-R constructs and scores rules using one attribute; feature selectors using One-R sort the attributes based on these scores.

3.2.2 Classifiers

As discussed in Section 1.2.4.1, classifiers are used to learn connections between independent features and the dependent feature (called the *class*) . Once these patterns are learned, outcomes can be predicted in new data by reflecting on the data that has already been examined. This study tried six different classifiers: One-R, C4.5, ADTrees, Naive Bayes, Bayes networks, and radial bias networks. These are some of the well-known and standard classifiers in the machine learning field, except for ADTrees.

One-R, described in Holte (1993), builds rules from the data by iteratively examining each value of an attribute and counting the frequency of each class for that attribute-value pair. An attribute-value is then assigned as the most frequently occurring class. Error rates of each of the rules are then calculated, and the best rules are ranked based on the lowest error rates.

A *radial basis function network* (RBFN) is an artificial neural network (ANN) that utilizes a radial basis function as an activation function (Bors). An ANN's activation function is used in order to offer non-linearity to the network. This is important for multi-layer networks containing many hidden layers, because their advantages lie in their ability to learn on non-linearly separable examples.

C4.5 (Quinlan, 1993) is an extension to the ID3 (Quinlan, 1986b) algorithm. A decision tree (shown in Figure 3.6) was constructed by first determining the best attribute to make as the root node of the tree (Mitchell, 1997). ID3 decides this root attribute by using one that best classifies training examples based upon the attribute's information gain (described above) (Quinlan, 1986b). Then, for each value of the attribute representing any node in the tree, the algorithm recursively builds child nodes based on how well another attribute from the data describes that specific branch of its parent node. The stopping criteria are either when the tree perfectly classifies all training examples, or until no attribute remains unused. C4.5 extends ID3 by making several improvements, such as the ability to operate on both continuous as well as discrete attributes, training data that

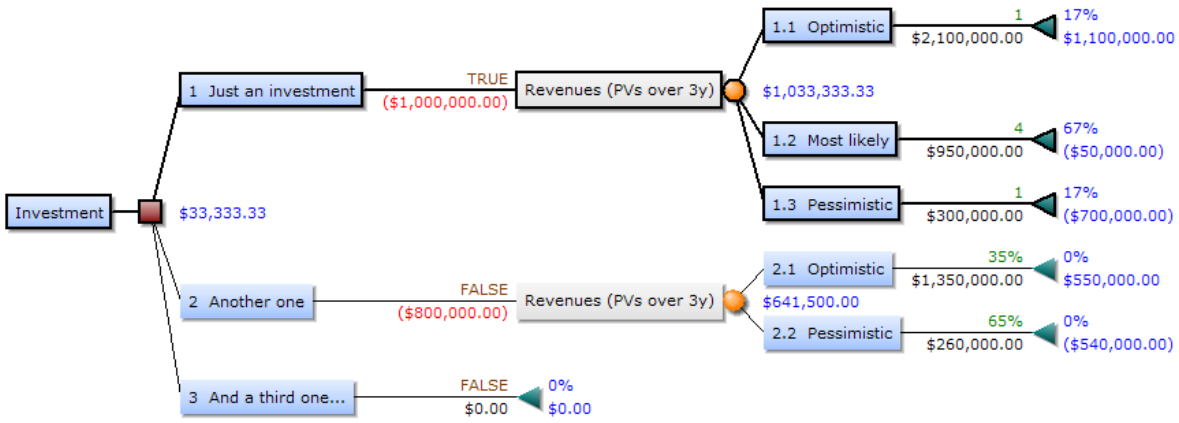


Figure 3.6: A decision tree consists of a root node and descending children nodes who denote decisions to make in the tree's structure. This tree, for example, was constructed in an attempt to optimize investment portfolios by minimizing budgets and maximizing pay-offs. The top-most branch represents the best selection in this example.

contains missing values for a given attribute(s), and employ pruning techniques on the resulting tree.

ADTrees are decision trees that contain both decision nodes, as well as prediction nodes (Freund and Mason, 1999). Decision nodes specify a condition, while prediction nodes contain only a number. Thus, as an example in the data follows paths in the ADTree, it only traverses branches whose decision nodes are true. The example is then classified by summing all prediction nodes that are encountered in this traversal. ADTrees, however, differ from binary classification trees, such as C4.5, where those trees only traverses a single path down the tree.

A *naive Bayes* classifier uses Bayes' theorem to classify training data. Bayes' theorem, as shown in Equation 3.1, determines the probability P of an event H occurring given an amount of evidence E. This classifier assumes feature independence; the algorithm examines features independently to contribute to probabilities, as opposed to the assumption that features depend on other features. Surprisingly, even though feature independence is an integral part of the classifier, it often outperforms many other learners (Rish; Domingos and Pazzani, 1997).

$$Pr(H|E) = \frac{Pr(E|H) * Pr(H)}{Pr(E)} \tag{3.1}$$

Bayesian networks, illustrated in Figure 3.7, are graphical models that use a directed acyclic graph (DAG) to represent probabilistic relationships between variables. As stated in Heckerman (1996), Bayesian networks have four important

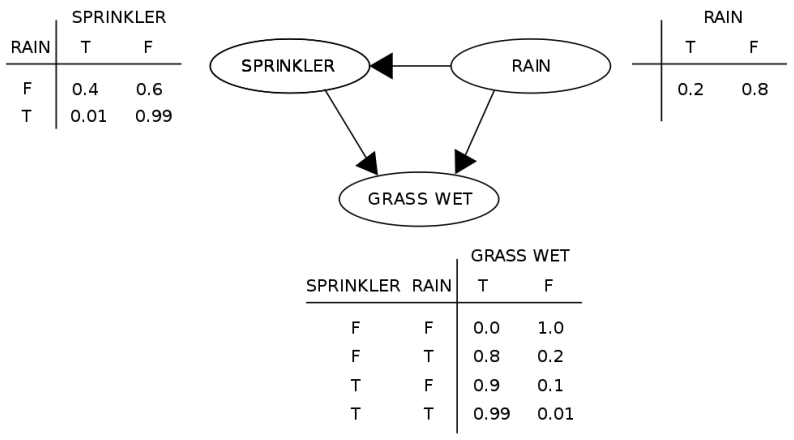


Figure 3.7: In this simple Bayesian network, the variable *Sprinkler* is dependent upon whether or not it's raining; the sprinkler is generally not turned on when it's raining. However, either event is able to cause the grass to become wet - if it's raining, or if the sprinkler is caused to turn on. Thus, Bayesian networks excel at investigating information relating to relationships between variables.

elements to offer:

1. Incomplete data sets can be handled well by Bayesian networks. Because the networks encode a correlation between input variables, if an input is not observed, it will not necessarily produce inaccurate predictions, as would other methods.

2. Causal relationships can be learned about via Bayesian networks. For instance, whether a certain action taken would produce a specific result and to what degree can be found.

3. Bayesian networks promote the amalgamation of data and domain knowledge by allowing for a straightforward encoding of causal prior knowledge, as well as the ability to encode causal relationship strength.

4. Bayesian networks avoid over fitting of data, as "smoothing" can be used in a way such that all data that is available can be used for training.

3.2.3 Cross-Validation

The value of different attributes can be assessed using equations one to four. In this experiment, a 5×5 cross-validation i.e. the data was partitioned five times into a test set consisting of $\frac{1}{5}$-th of the data and a training set of $\frac{4}{5}$-ths of the

```
For each run
    For each number of attributes bins
        For each FSS
           For each bin of cross-validation
              Divide data in train and test
              For each learner
                  Learn on train data and generate results on Test
              Loop
           Loop
         Loop
     Loop
Loop
```

Figure 3.8: Pseudocode of the experiment set-up for selecting the number of attributes

```
For each run
   For each bin of cross-validation
      Divide data in train and test
      For each learner
         Learn on train data and generate results on Test
      Loop
   Loop
Loop
```

Figure 3.9: Pseudocode of the experiment set-up for generating results once the dataset is reduced

data was performed. After the five rounds, the median values of recall and false alarm rates were recorded to study the variance in these performance figures.

Figure 3.8 shows the pseudocode of the experiment set-up for selecting the number of attributes, and Figure 3.9 shows the pseudocode of the experiment set-up for generating results once the dataset is reduced.

Results obtained from the set-up given in Figure 3.8 were used to perform variance analysis on the probability of detection (PD) and the probability of false alarm on various attribute group sizes (i.e. $n \in 5, 10, 15, .., 100, 103$). The dataset was then reduced to the number of attributes that performed the best in terms of PD, PF, and variance; this reduced dataset was again used to generate results using learners and cross-validation.

3.2.4 Contrast Set Learning

After determining the subset of the attributes that best predict for student retention, a *contrast set study* was conducted . Contrast set learners such as TAR3 (Menzies and Hu, 2007) seek attribute ranges that are most *different* in various outcomes .

One way to read these contrast sets are as *treatments* that promise if action X was applied to a domain, then this would favor outcome X over outcome Y. In this study, TAR3 was used two ways:

1. Used TAR3 to find which treatments were selected the most for retention;
2. Used TAR3 in the opposite direction to find the treatments that most selected for students leaving university.

The first use of TAR3 found the actions that encouraged retention and second use of TAR3 found the actions that increased drop-out.

Chapter 4

Results

> I have had my results for a long time: but I do not yet know how I am to arrive at them.
>
> Carl Friedrich Gauss

4.1 Analysis of Experimental Results

4.1.1 Evaluation Metrics

The evaluation metrics used in this experiment were standard data mining performance measures of a method. They were:

- Probability of detection (PD);
- Probability of false alarm (PF);
- And variance PD and PF seen over the cross-validation study.

It was critical to study the variance in PD and PF values, as some classifiers can obtain very high PDs in some runs but low PDs in some other runs. Thus, the results of such classifiers are inconsistent to generate a reliable theory. For this study, all the PD values and PF values were rejected if the variance was greater than $\pm 25\%$.

The PDs, PFs, and variances statistics were collected over 1500 experiments, which were repeated 20 times to check for conclusion stability. In total, the number of experiments was:

$$5 * 5 * 4 * 6 * 3 * 20 = 36,000$$

experiments; i.e. 5×5 cross-validation using four feature subset selectors and 6 different learners, for the three data sets of three years of retention. This was repeated 20 times using the top $n \in 5, 10, 15, .., 100, 103$ attributes as found by the feature selector.

4.1.2 Visualizing the Results

Figures 4.1, 4.2, and 4.3 show the PD and PF median results of first, second and third year retention against the variances of these values. Each point in these figures represents a combination of the number of attributes selected, the feature subset selector used to select the attributes, and the classifier used to train on the sub-selected data. The color of each point shows the number of attributes used for that particular combination representing that point.

The horizontal line segmenting these graphs was a baseline reference to the existing retention rates in the data. For example, approximately 70% of the students returned after the first year, hence, the baseline for PD graph of the first-year data was set at 70. To have good predictability in the learners, the learners should perform better than the baseline. As illustrated in the figures 4.1, 4.2, and 4.3, the median probabilities of detection of retention values for the first year were lower than the baseline, and therefore using these methods, first-year retention could not accurately be predicted. Although the median probabilities of detection of retention values for the second year were better than the baseline, these results were marginally better than the baseline. Third year PD values however successfully exceeded the baseline and were studied in detail.

Figure 4.4 shows how various learners performed on the third-year retention data for all attribute ranges; this figure shows that AdTrees and Bayes net had relatively high PDs and low PFs compared to other learners over all runs (twenty-five cross-validation runs $\times$ twenty attribute ranges $[n \in 5, 10, 15, .., 100, 103] \times$ four reducers $\times$ seven learners (including ZeroR)).

4.1.3 First Results

After rejecting all results with (1) a PD lower than the ZeroR limit; (2) a PD variance greater than $\pm 25\%$; and (3) a PF higher than 25%, it was found that there were no good predictors for Year 1 or Year 2 retention. This was the first major finding for this research: *it is very difficult to predict for lower year retention*; this was demonstrated in the literature review as well.

Figure 4.5 shows the distribution of PDs and PFs for the third-year retention across all attribute ranges using all learners and reducers. The distribution shows tremendous variation in the performance. Therefore, the results with variance greater than $\pm 25\%$ were pruned.

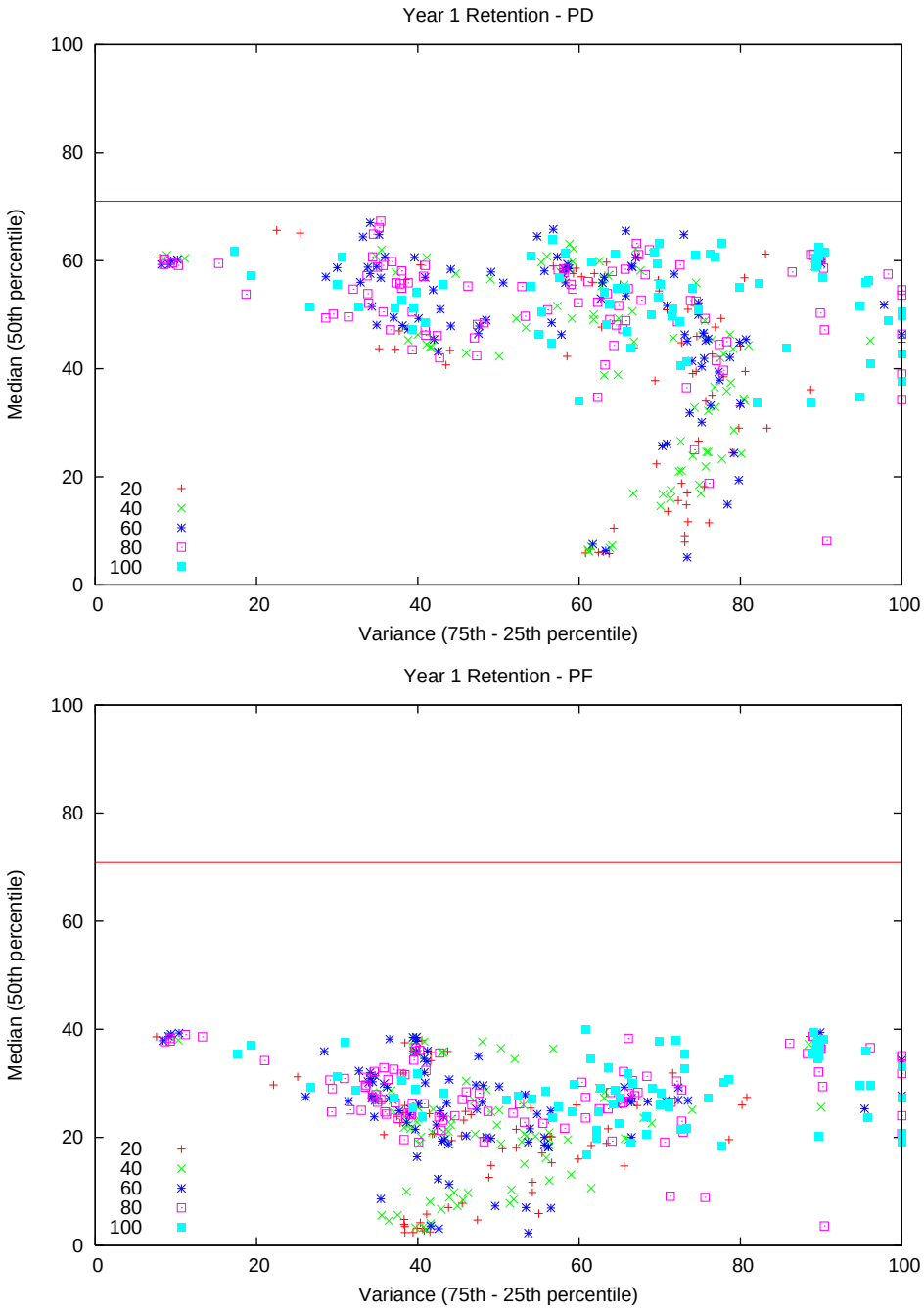


Figure 4.1: Probability of Detection (PD) and Probability of False Alarm (PF) with variances for first year retention.

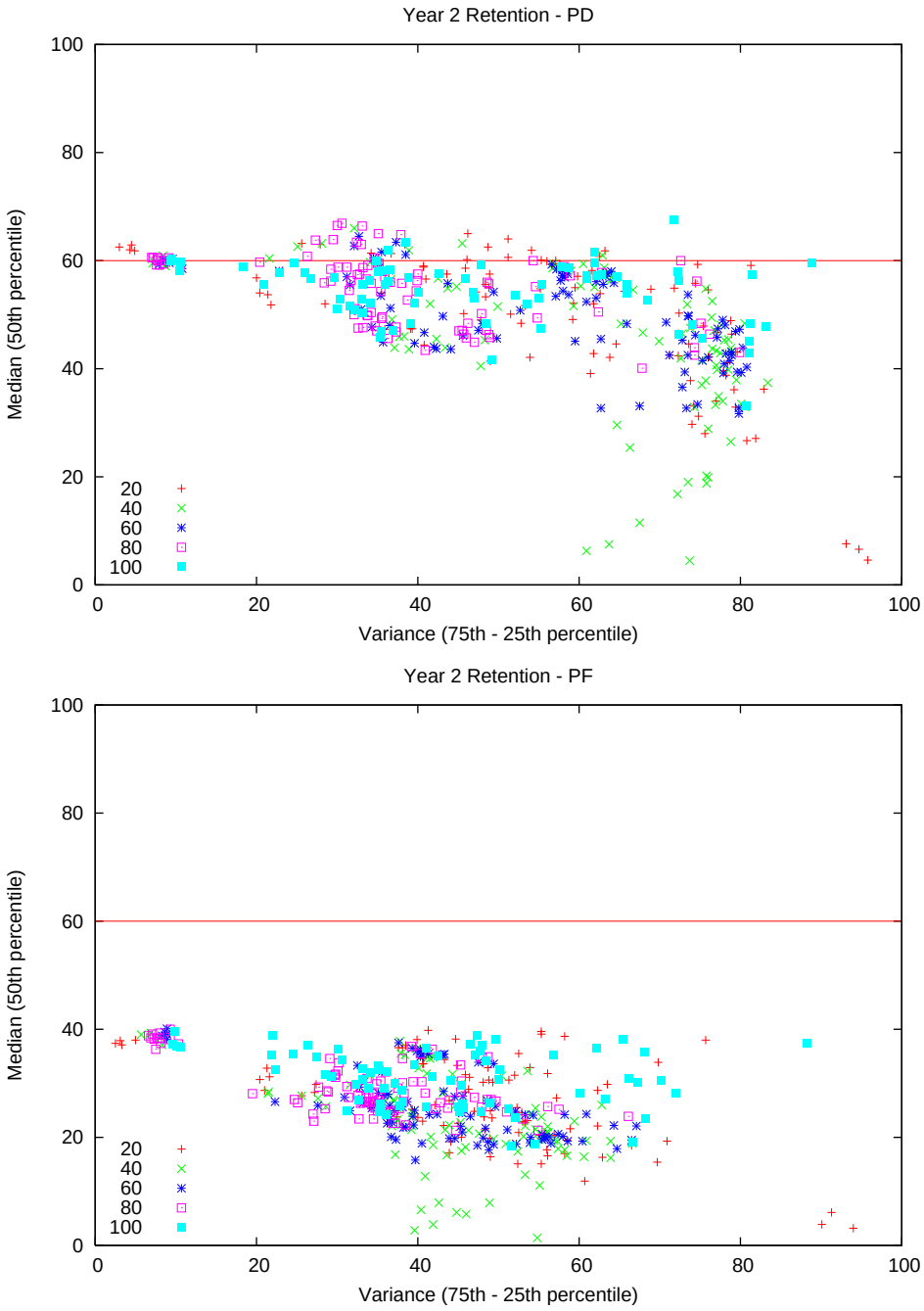


Figure 4.2: Probability of Detection (PD) and Probability of False Alarm (PF) with variances for second year retention.

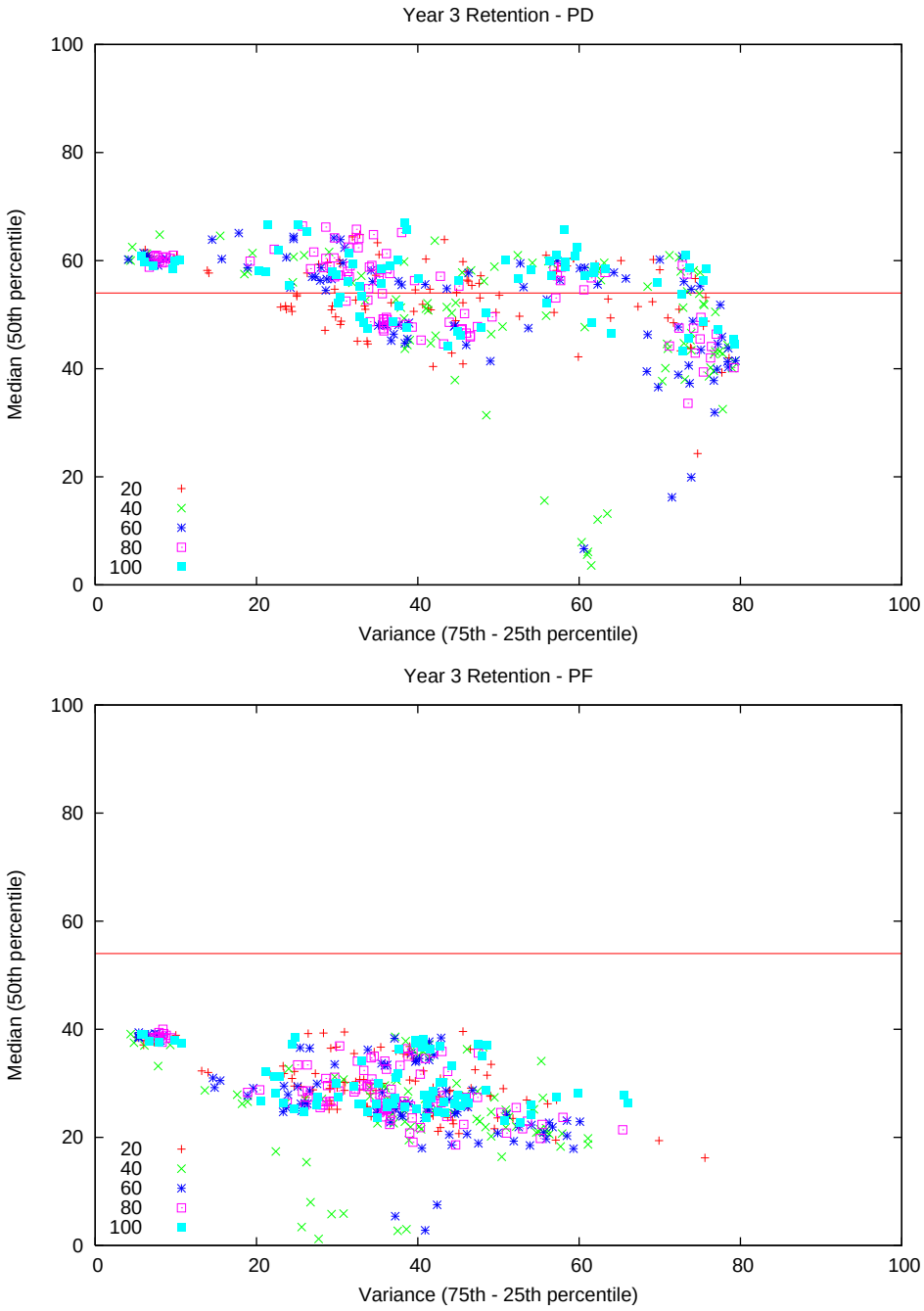


Figure 4.3: Probability of Detection (PD) and Probability of False Alarm (PF) with variances for third year retention.

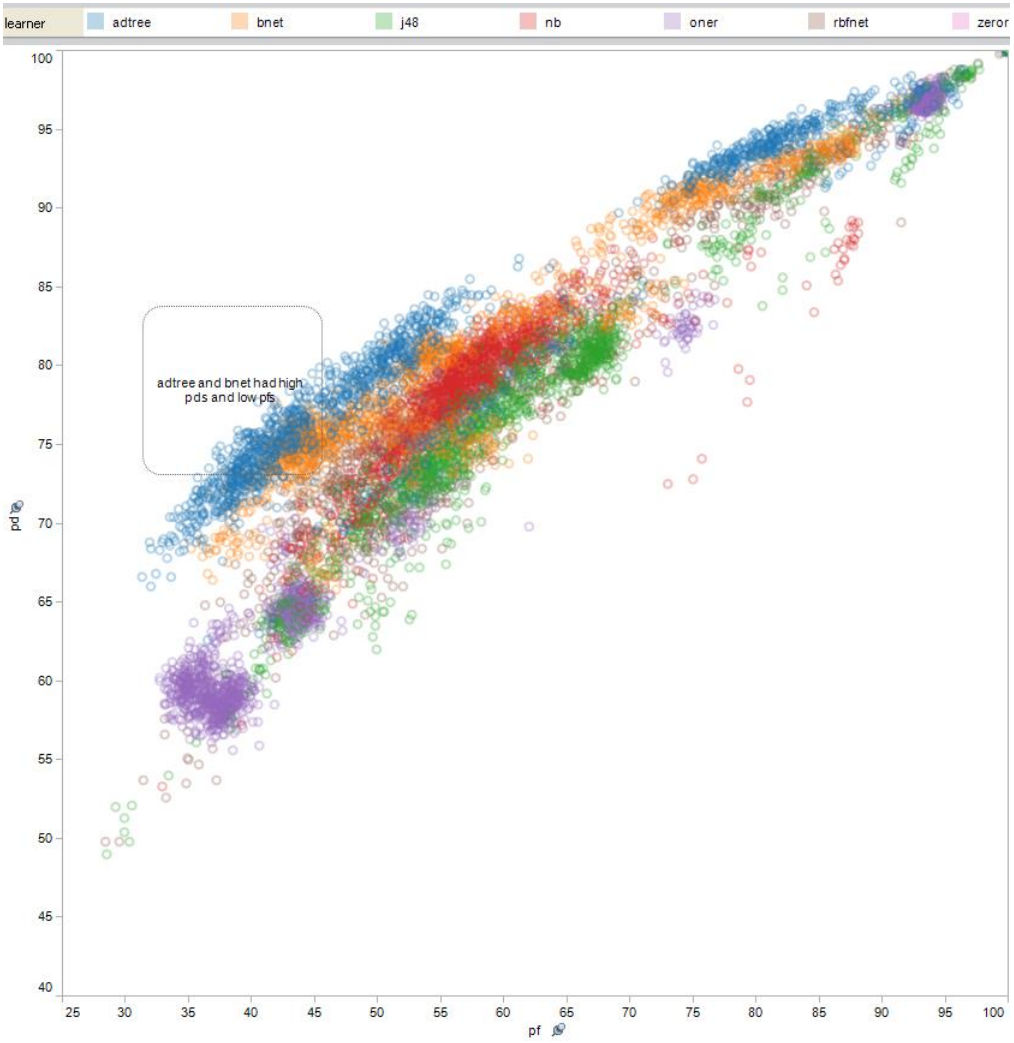


Figure 4.4: Probability of Detection (PD) and Probability of False Alarm (PF) performance of learners for all attribute ranges.

For the rest of this study, only third year retention was studied. Studying third year retention was based on these factors:

- Although first-year success is critical for higher education institutions, graduation from the university is rewarding for the university and the student both. In addition, if the goal is to provide a complete university education for a student, then predicting survival till second year is less interesting than lasting till third year.
- Third year retention implies second and first year retention; it drives the graduation rates as well.

Figure 4.6 shows the distribution of PDs and PFs for attribute range between

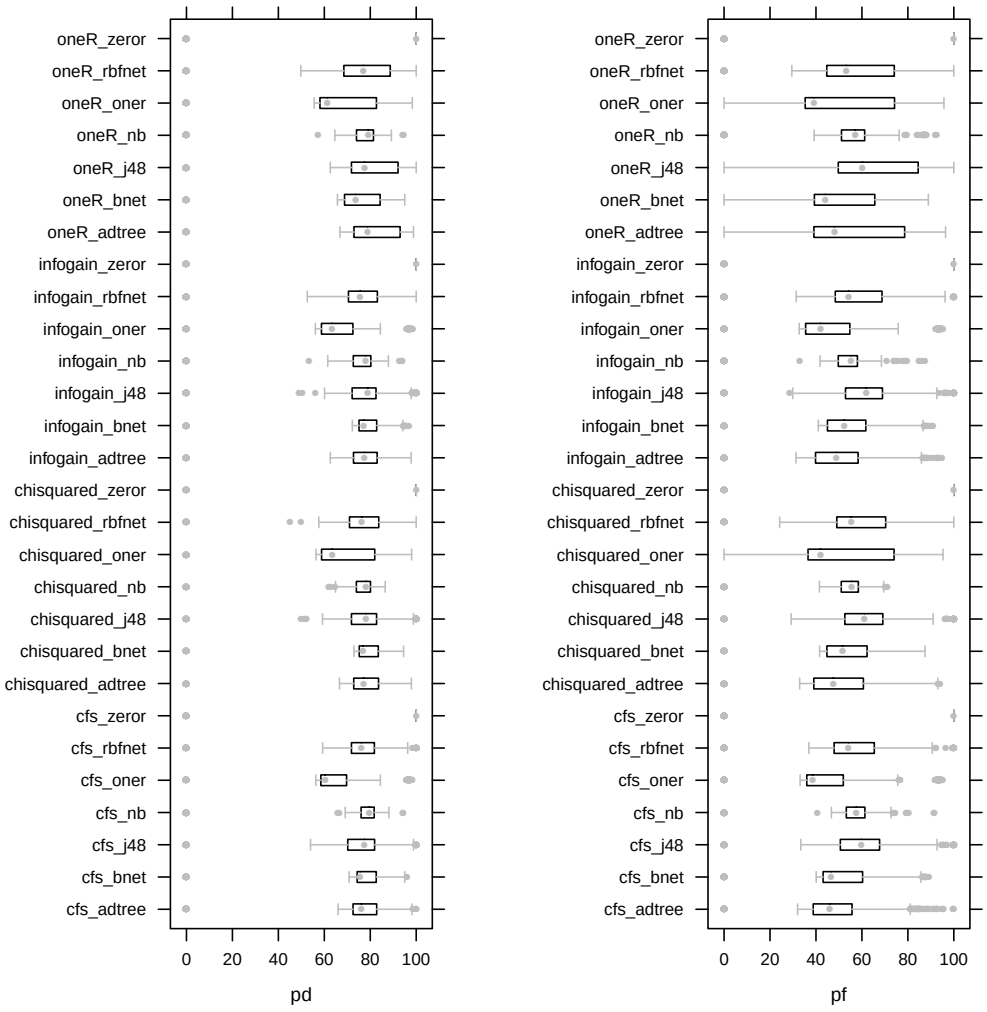


Figure 4.5: Probability of Detection (PD) and Probability of False Alarm (PF) Distribution for all attribute ranges, learners, and reducers

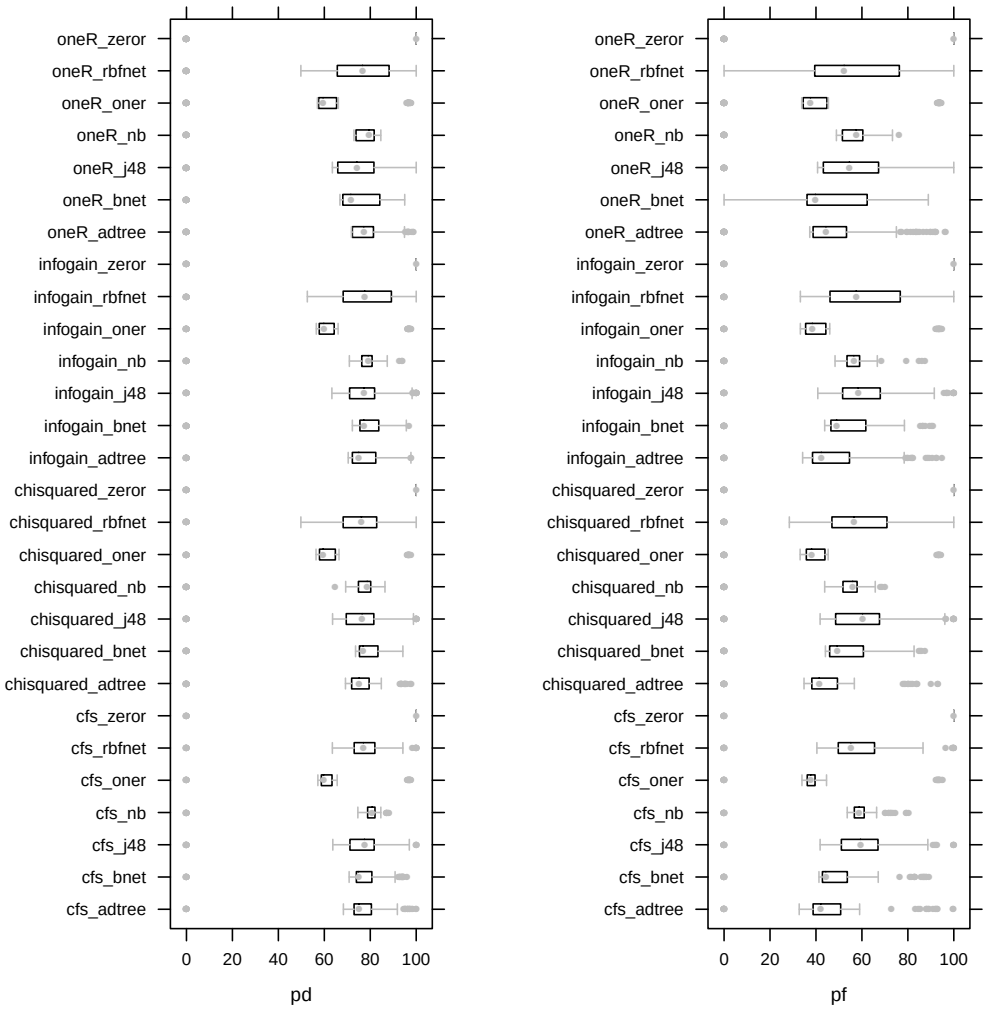


Figure 4.6: Probability of Detection (PD) and Probability of False Alarm (PF) Distribution for attribute range between 30 and 50, learners, and reducers

Rank	Number of Attributes	FSS	Classifier
61	30	oneR	bnet
61	50	cfs	adtree
57	50	oneR	adtree
56	30	oneR	adtree
55	30	cfs	adtree
52	50	oneR	bnet
51	30	infogain	adtree
51	30	cfs	bnet
48	50	infogain	adtree

Figure 4.7: The top ten ranking treatments for third year retention. Ranks represent how many times a particular treatment wins over all other treatments in the experiment.

30 and 50 before pruning the results by variance of PD. Some methods had low variance and some had high variance. To find the best performing method in terms of high PDs, win-loss table was created using Mann-Whitney test .

4.1.4 Ranking with the Mann-Whitney Test

After pruning results with low PD, high PF, or high PD variance, remaining results were ranked via a Mann-Whitney test with 95% confidence . The ranks were determined by counting how many times a combination won compared to another combinations. The method that won the most number of times was then given the highest rank; treatments that won with same PD values were given identical ranks. The table in Figure 4.7 shows the top ten ranking combinations based on a PD performance measure.

Figure 4.8 shows the plot of PDs vs. PFs for the final selected combinations of attribute ranges, learners, and reducers; this figure shows that combinations with high PDs also had high PFs and vice-versa. The region with reasonably high-PDs ($\geq$ 68 and $\leq$ 76) and with reasonably low-PDs ($\geq$ 35 and $\leq$ 46) was considered a "sweet-spot."

As the results achieved using 30 or 50 attributes were similar, Occam's Razor was applied and studied the top 30 attributes found to be best for oneR/bnet combination. Performance measures obtained by OneR feature selector and Bayes net classifier are shown in Figure 4.2. Figure 4.1 shows these top 30 attributes with their ranges and the probability of return after three years.

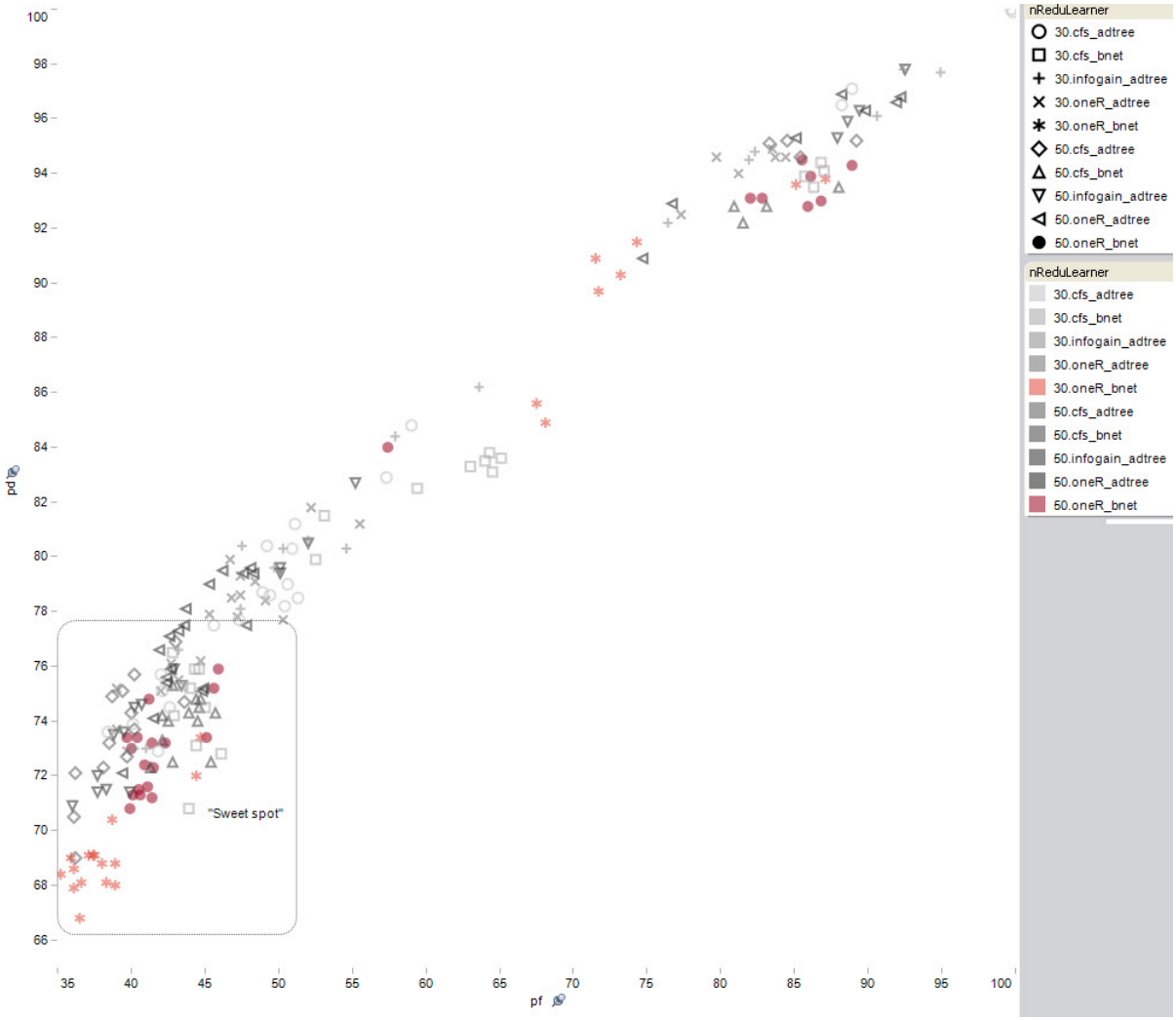


Figure 4.8: Probability of Detection (PD) vs. Probability of False Alarm (PF) for selected learners, reducers, and attribute ranges. Points marked by asterisks show the results obtained using 30 attributes and One-R as the reducer and Bayes net as the learner.

Attribute	Description	Value	Instances	$P(RET3) = Y$ 0% 20% 40% 60% 80% 100%
FinAidSTUDENT_TA	Student's Tax Form Type	4	35	
		3	24	
		2	12,215	
		1	2,697	
FinAidMOTHER_ED	Mother's Education Level	3	7,710	
		4	814	
		2	8,792	
		1	289	
FinAidSTUDENT_MA	Student's Marital Status	M	386	
		U	17,254	
		S	24	
FinAidFATHER_ED	Father's Education Level	3	7,502	
		2	8,461	
		4	1,136	
		1	436	
FinAidDEPENDENCY	Student's Dependency Status	I	2,523	
		D	15,154	
FirstGenInd	First Generation Student	N	10,370	
		Y	7,311	
FinAidPARENT_TAX	Parent's Tax Form Type	4	24	
		1	13,101	
		2	3,126	
		3	16	
FinAidSTUDENT_AG	Student's Adjusted Gross Income	4829.5-7915.5	4,152	
		3335.5-4829.5	2,780	
		16713.5-inf	1,022	
		1894.5-3335.5	2,540	
		-inf-1894.5	2,106	
FinAidSTUDENT_WA	Student's Wage	7850.5-9958	1,752	
		4092.5-7850.5	5,622	
		1.5-999.5	2,057	
		1903.5-4092.5	4,176	
		-inf-1.5	1,721	
HS_GPA	High School GPA	3.015-3.345	5,769	
		2.905-3.015	1,990	
		2.645-2.905	4,541	
		2.035-2.645	4,758	
		-inf-2.035	545	
FinAidPARENT_MAR	Parent's Marital Status		464	
		W	394	
		M	11,328	
		S	3,127	
		U	637	
PercentileRankHSGPA	Percentile Of Hs Gpa Among Freshmen Cohort	45.75-59.65	3,660	
		33.7-45.75	3,165	
		15.35-33.7	4,803	
		2.35-15.35	3,479	
		-inf-2.35	637	
FinAidPARENT_AGI	Parent's Adjusted Gross Income	96636-inf	3,751	
		58550.5-96636	6,045	
		18376.5-58550	5,598	
		-inf-18376.5	1,167	
FinAidFATHER_WAG	Father's Income	52366-inf	5,873	
		-inf-52366	9,459	
FinAidMOTHER_WAG	Mother's Income	42957-inf	3,148	
		-inf-42957	13,063	
HS_PERCENT	High School Percentile	80.5-inf	5,838	
		60.5-80.5	6,980	
		43.5-60.5	5,624	
		-inf-43.5	7,774	
MaxACT	Max Of ACT Score And ACT Equivalent	23.5-inf	6,952	
		19.5-23.5	10,044	
		15.5-19.5	7,001	
		-inf-15.5	2,219	
~~PercentileRankMaxACT~~	Percentile Of Max ACT ~~Among Freshmen~~ Cohort	71.35-inf	6,658	

Table ?? continued on next page

Attribute	Description	Value	Instances	$P(RET3) = Y$ 0% 20% 40% 60% 80% 100%
		30.55-71.35	10,281	
		8.35-30.55	6,763	
		-inf-8.35	2,514	
CUR_ERLHRS	Total Enrolled Hours	14.5-18.5	14,523	
		13.5-14.5	6,964	
		10.5-13.5	4,016	
		-inf-10.5	532	
		18.5-inf	181	
ACT1_COMP	ACT Comprehensive Score (new)	23.5-inf	5,669	
		19.5-23.5	8,667	
		17.5-19.5	4,043	
		-inf-17.5	7,837	
ACT1_MATH	ACT Math Score (new)	22.5-inf	7,082	
		19.5-22.5	4,767	
		16.5-19.5	6,611	
		-inf-16.5	7,756	
ACT1_ENGL	ACT English Score (new)	24.5-inf	4,676	
		19.5-24.5	8,271	
		16.5-19.5	4,877	
		-inf-16.5	8,392	
AGE	Age of Student at Matriculation	-inf-19.5	24,826	
		19.5-inf	1,390	
ENG10	Enrolled in English Courses	N	24,407	
		Y	1,809	
LIVEONCAMP	On-Campus Indicator	Y	20,087	
		N	6,129	
ADMIT_MAJ	Admit Major	ADV	38	
		AERN	433	
		AEDG	208	
COMP_WRITE	Compass Writing Score	-inf-9.5	3,780	
		73.5-inf	13,887	
		49.5-73.5	5,299	
		9.5-49.5	3,250	
TotalClasses	Total Number of Enrolled Classes	5.5-inf	15,021	
		4.5-5.5	10,237	
		-inf-4.5	958	

Table 4.1: Top 30 attributes with values. Only five attribute values with at least 10 records are shown.

After selecting the best combination of FSS (oneR) and classifier (Bayes Network) based on Mann-Whitney test rankings, we found that attributes given in Table 4.3 are critical to third-year persistence. Out of these 30 attributes, top ten attributes described student's family background and family's economic condition, and the most selected attribute was the student's tax form type, which came from the FAFSA submission and had these values :

1. IRS 1040
2. IRS 1040A, 1040EZ
3. A foreign tax return
4. A tax return with Puerto Rico, another U.S. territory or a Freely Associated State

A person is eligible to file 1040A or 1040EZ if he or she makes less than $100,000, does not itemizes deductions, does not claim dependents, etc. As shown

in Figure 3.5, there is a positive correlation between tax form type 2 and third-year retention for lower high school GPA ranges with the exception of the range: 2.645 to 2.905. Third-year retention percentages are significantly higher for the students who (or their parents) have filed a foreign tax return (type 3) or a U.S. territory tax return (type 4) than those who have filed U.S tax return (type 1 or 2).

Second attribute in the list was the parent's household size , which had a positive correlation with third-year retention percentage as shown in Figure 3.4 along with the distribution of the parent's household size. The sample size was low for student's with large number of people in the household, therefore, retention percentages in such cases is meaningless.

As previous research has concluded that parent's education level plays an important role in student's dropout decision (Spady, 1970; Tinto, 1975; Bean, 1979), Figure 3.3 shows that chances of student's persistence are higher if the parent's education level is higher. If the parents did attend college and beyond, father's education level has greater impact than mother's education level on student's persistence.

As shown in the Table 4.1, student's marital status does play a role in persistence, especially if the student is separated (denoted by S in the table). Out of 24 students, who indicated in FAFSA as separated , only four students persisted till the third year. Students income (FinAidSTUDENT_WA) also affect their persistence; students with wages in the range of $7850.5-$9958 had the highest percentages of return (close to 80%).

Figure 4.3 shows the ranges which, in isolation, had a retention probability greater than the ZeroR limit for (for third year, that ZeroR limit was 55%). Figure 4.3 also shows the "hypothesis" group of an attribute. According to this table, some key findings were:

- The ranges shown at the top of the table were most predictive for third year retention. "Financial Aid" attributes appeared the most at the top (Figure 4.3).
- Attributes related to student "Performance" were rarer in the list.
- None of the attribute ranges included the "Faculty Type and Experience" attributes of Figure 4.3.

This analysis led to these conclusions:

- Using experienced instructors or tenured faculty was *not* predictive for third year retention.
- Issues relating to financial aid (e.g. income, education level) dominated over student performance.

Class Value	Baseline	Probability of Detection (PD)	Probability of False Alarm (PF)	Precision	Accuracy
Y	54.78%	70.0%	34.9%	70.8%	67.8%
N	45.22%	65.1%	30.0%	64.2%	67.8%

Table 4.2: Performance Measures obtained for RET3 using OneR as the FSS and Bayes Net as the classifier.

4.1.5 Ranking with Contrast Set Learning

As these conclusions could be argued that Figure 4.3 only discusses the effect of attribute ranges in *isolation*, it is possible that combination of factors might lead to different conclusions . The TAR3 treatment learner was used to test this possibility . TAR3 learner was set to build at the maximum 10 rules (i.e. ten combinations of attribute ranges) from the 30 attributes selected by the best learning combination of Figure 4.3; however, TAR3 never found combinations larger than three ranges and this max size of 10 ranges was much larger than necessary.

4.2 Results

The stated data mining techniques were unable to significantly improve the classification rates for first-year and second-year retention prediction over the baseline, but achieved approximately 20% higher probability of detection for third-year retention over the baseline. As it is possible to predict third-year retention probability with high accuracy, based only on the first-year, beginning of term data, this result is significant in student persistence research.

Figure 4.3 lists the rankings of all attribute ranges which, in isolation, predicted for third year retention at a probability higher than the ZeroR limit (55%), and were supported by good number of records. Top six attributes affecting third-year retention were from financial aid hypothesis: student's wages, parent's adjusted gross income, student's adjusted gross income, mother's income, father's income, and high school percentile. Of those students who reported their wages, students who made between 7,850 and 9,958 had a 79% retention. Similar rules were found for parent's income and adjusted gross income. It means that the students with stronger financial support usually stay in college than the students with weaker financial support.

After these top six attributes, high school percentile of 81 or greater was an important attribute with 69% of students returning after three years. Some other "performance" attributes were ACT scores and ranks. This supports the

argument that scores do have some predictability of student retention.

TAR3 results, given in Figure 4.9, produced simple theories (treatments) that combined ranges of various attributes that maximized the student retention. For example, the student retention was very high for students with the AGI in the range from $7,000 to $724,724 and father's wages were in the range from $56,289 to $999,999. One more interesting theory with high retention was where father's education level was 3 (college) and student's rank amongst the freshmen cohort was between 66.3 and 98.4.

Treatments that predicted student drop-out were based on the total number of classes student was enrolled, English 10000, an introductory college writing and supplemental instruction class, and on-campus living. Students who took less than five class, enrolled in the English 10000 class, and did not live on-campus were at high risk of drop-out. Chart on the bottom of Figure 4.9 shows the retention percentage of each treatment. For example, students enrolled in English 10000 had a 40% retention in their third year.

Key findings were:

- Student's and parent's income capacity and levels affected student retention. Third-year retention was higher for the students with high income than the students with low income. According to treatment 1, approximately 82% of students who had at least $7,000 AGI and their fathers' income was at least $56,289 returned after three years. Similarly, according to treatment 5, approximately 79% of students who made at least $5,383 and their parents' AGI was at least $87,744 returned after three years.

- Students with better high school performance amongst their peers had higher chances of retention. According to treatment 2, approximately 81% of students who had at least $7,000 AGI and had high school percentile of 72 and better returned after three years. Approximately 79% students who had at least 3.34 HS GPA and whose parents had an AGI of at least $84,744 stayed after three years, given in treatment 4.

- ACT scores, rank of these scores amongst peers, and COMPASS scores affected student retention. Students with higher scores and rank had higher chances of retention. According to treatment 3, approximately 80% of students who had at least $7,000 AGI and had ACT math score of 21 or better returned after three years. Similarly, 77% of students who had at least 23 in ACT composite (or SAT equivalent) and had an income of at least $5,383 and less than $561,500 returned after three years, given in treatment 6.

- Parent's education level had a positive effect on student retention. Students whose parents did not attend college had a lower retention compared to

students whose parents did attend college. As given in treatments 7 and 10, a student was highly likely (77%) to return after three years: (7) if the mother of that student attended college, the student had a ACT composite score of 22 or better, the parents' AGI was at least $84,744; (10) if the father of that student attended college and the student's percentile rank amongst other freshmen in the cohort was at least 66.3.

- Enrolling in fewer classes (less than five), enrolling in English 10000 (an introductory college writing class), and living off-campus had a negative effect on student retention, as given in treatments 11, 12, and 13. It is important to note that enrolling in that English course itself was not a predictor of non-retention, but the sample of the students that attended this class were at high-risk of dropping out.

4.2.1 Strategic Actions

This study provides insights in student retention domain using beginning of term data. These insights can be used to design effective policies and strategic actions, such as:

- Most of the attributes were related to socio-economic levels and capacities of students and their parents; however, this cannot be controlled while admitting students, but better support programs and calculated financial-aid packaging for students with lower economic capacities can be created.

- First-year students should be encouraged to live on-campus by providing some incentives, as on-campus students have higher chances of retention.

- Special guidance and supplemental instruction in writing and reading should be provided to first-generation students. In addition, parents of first-year generation students have considerably low-incomes than the parents of non-first-generation students, and according to the results of this study, income of parents is a critical factor in student retention even if the students had similar academic performance.

- Students are placed in the supplemental instruction classes, such as English 10000, based on their COMPASS and ACT scores. As these students' scores indicated lack of academic preparedness in some areas, academic advisers correctly place students in such classes; however, if the students fail or perform poorly in such classes, it leaves a lasting impression and sets the students to for future drop-out, even after three years. Therefore, it is paramount that advisers not only place students in supplemental instruction classes, but also ensure the success of students in these classes and improve the skills that the students lack. Out of all classes considered in

#	Treatment
1	7000 $\leq$ FinAidSTUDENT_AG $<$ 724,724 and 56,289 $\leq$ FinAidFATHER_WAG $<$ 999,999
2	7,000 $\leq$ FinAidSTUDENT_AG $<$ 724,724 and HS_PERCENT $\geq$ 72
3	7,000 $\leq$ FinAidSTUDENT_AG $<$ 724,724 and 21 $\leq$ ACT1_MATH $<$ 36
4	84,744 $\leq$ FinAidPARENT_AGI $<$ 999,999 and HS_GPA $\geq$ 3.34
5	84,744 $\leq$ FinAidPARENT_AGI $<$ 999,999 and 5383 $\leq$ FinAidSTUDENT_WA $<$ 561,500
6	23 $\leq$ MaxACT $<$ 35 and 5383 $\leq$ FinAidSTUDENT_WA $<$ 561,500
7	22 $\leq$ ACT1_COMP $<$ 35 and 84,744 $\leq$ FinAidPARENT_AGI $<$ 999,999 and FinAidMOTHER_ED=3
8	5383 $\leq$ FinAidSTUDENT_WA $<$ 561,500 and 21 $\leq$ ACT1_MATH $<$ 36
9	HS_GPA $\geq$ 3.34 and 32,570 $\leq$ FinAidMOTHER_WAG $<$ 533,395
10	FinAidFATHER_ED=3 and 66.3 $\leq$ PercentileRankHSGPA $<$ 98.4
11	1 $\leq$ TotalClass$\leq$ 5
12	ENG10=Y
13	LIVE.ON.CAMP=N

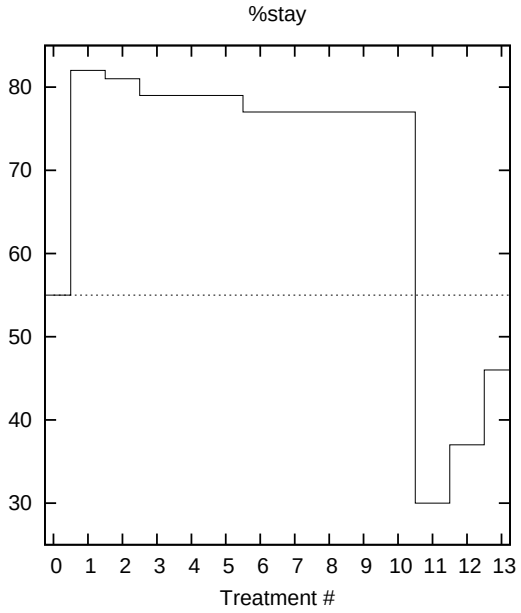


Figure 4.9: Treatments 1 to 10 are the top ten treatments found by this analysis that increases the third year retention rates. Treatments 11,12,13 are the worst three treatments found by this analysis that *most decrease* the third year retention rates. The effects of each treatment, is shown on the bottom plot.

this study, English seemed to have the greatest impact. Intuitive as it may be, to succeed in college, students need good writing and reading skills.

#	$P(Ret3\|X)$ (percent)	Support (#students)	Hypothesis	X (Feature = Range) Feature	Range
1	79	1,752	Financial Aid	Student's Wage	7850 to 9958
2	73	3,751	Financial Aid	Parent's Adjusted Gross Income	96636 to inf
3	71	4,152	Financial Aid	Student's Adjusted Gross Income	4830 to 7916
4	71	3,148	Financial Aid	Mother's Income	42957 to inf
5	70	5,873	Financial Aid	Father's Income	52366 to inf
6	69	5,622	Financial Aid	Student's Wage	4093 to 7851
7	69	5,838	Performance	High School Percentile	81 to inf
8	68	2,523	Financial Aid	Student's Dependency Status	I
9	68	7,502	Financial Aid	Father's Education Level	3
10	67	6,045	Financial Aid	Parent's Adjusted Gross Income	58551 to 96636
11	66	12,215	Financial Aid	Student's Tax Form	2
12	66	7,710	Financial Aid	Mother's Education Level	3
13	66	2,057	Financial Aid	Student's Wage	1.5 to 1000
14	66	10,370	Financial Aid	First Generation Student	N
15	65	7,082	Performance	ACT Math Score (new)	23 to inf
16	65	2,780	Financial Aid	Student's Adjusted Gross Income	3336 to 4830
17	65	4,676	Performance	ACT English Score (new)	25 to inf
18	65	5,669	Performance	ACT Comprehensive Score (new)	24 to inf
19	65	13,101	Financial Aid	Parent's Tax Form	1
20	65	11,328	Financial Aid	Parent's Marital Status	M
21	64	6,658	Performance	Percentile Of Max ACT Among Freshmen	71 to inf
22	64	6,952	Performance	Max Of ACT Score And ACT Equivalent	24 to inf
23	64	2,697	Financial Aid	Student's Tax Form	1
24	63	17,254	Financial Aid	Student's Marital Status	U
25	63	13,063	Financial Aid	Mother's Wages	-inf to 42957
26	63	3,126	Financial Aid	Parent's Tax Form	2
27	63	1,022	Financial Aid	Student's Adjusted Gross Income	16714 to inf
28	62	15,154	Financial Aid	Dependency	D
29	62	9,459	Financial Aid	Father's Income	-inf to 52366
30	61	8,792	Financial Aid	Mother's Education Level	2
31	61	8,461	Financial Aid	Father's Education Level	2
32	61	4,176	Financial Aid	Student's Wage	1904 to 4093
33	60	14,523		Total Enrolled Hours	15 to 19
34	60	2,540	Financial Aid	Student's Adjusted Gross Income	1895 to 3336
35	59	3,780	Performance	Compass Writing Score	-inf to 10
36	59	8,271	Performance	ACT English Score (new)	20 to 25
37	59	7,311		FirstGenInd	Y
38	59	6,980	Performance	HS_PERCENT	61 to 81
39	59	15,021	Performance	Total Number of Enrolled Classes	6 to inf
40	59	5,598	Financial Aid	Parent's Adjusted Gross Income	1838 to 58551
41	58	3,127	Financial Aid	Parent's Marital Status	S
42	58	8,667	Performance	ACT Composite	20 to 24
43	58	4,767	Performance	ACT Math Score (new)	20 to 23
44	58	13,887	Performance	Compass Writing Score	74 to inf
45	58	5,769	Performance	High School GPA	3.02 to 3.4
46	58	10,281	Performance	RankMaxACT	31 to 71
47	58	10,044	Performance	MaxACT	20 to 24
48	57	20,087		On-Campus Indicator	Y
49	57	11,36	Financial Aid	Father's Education Level	4
50	56	24,826		Age of Student at Matriculation	-inf to 19.5
51	56	24,407	Performance	Enrolled in English Courses	N
52	56	3,660	Performance	Percentile Of Hs Gpa Among Freshmen	46 to 60

Table 4.3: Ranking all attribute ranges which, in isolation, predict for third year retention at a probability higher than the ZeroR limit (55%). From the above, the strongest predictor for third year retention is a student's wage (at 79%). On the other hand, the bottom line of this table says that the percentile of a student amongst their Freshmen cohort is little better than ZeroR (at 56%).

Chapter 5

Conclusions

> Glaciers in the Himalaya are receding faster than in any other part of the world and, if the present rate continues, the likelihood of them disappearing by the year 2035 and perhaps sooner is very high. Not!!!
>
> IPCC

5.1 Summary of Research

After an extensive review, gaps in the present research were identified in the area of student retention, especially using data mining. It was found: (a) Existing techniques did not perform better than the baseline; (b) Researchers did not use ensemble techniques of discretization, feature subset selection, learning over various learners, and cross-validation; (c) In-depth data mining experiments were missing.

To address these questions, first-time freshmen (beginning of the term) data was analyzed using discretization, feature subset selection, learning over various learners, and cross-validation. These techniques were repeated for three different datasets: first-year retention, second-year retention, and third-year retention. Although these techniques could not predict first or second year retention with significantly higher accuracies than the baseline, these techniques obtained probability of detection approximately 15% higher for the class value of Y and 20% higher for the class value of N than the baseline percentages for third-year retention, based on the first-year beginning of the term data.

In the studied literature, no studies with such a significant improvement over the baseline for the third-year retention were found. In addition, if policies are designed to improve third-year retention rate (using this predictive model), not only will they improve first and second year retention rates, but also the six-year graduation rates.

For the studied institution, family background and family's social-economic status are critical for student's third-year persistence. Using feature subset selection methods, it was found that the attributes from the "financial aid" hypothesis were selected the most as predictors of retention, and although the attributes from the "performance" hypothesis were selected, their predictability, in isolation, was lesser than the attributes from the "financial aid" hypothesis. None of the attributes from the "faculty tenure and experience" were selected by the feature subset selectors.

These results could very well be true only for the studied institution; however, if the approach detailed in this study is followed, other institutions can find top performing classifier and important attributes. Recommended practice: (a) data discretization; (b) feature subset selection with cross-validation and evaluation the performance over various learners; (c) treatment learners, such as TAR3 to find succinct strategic actions in complex data.

5.2 Contributions of Research

As these types of extensive experiments are missing from the literature, this research presents a framework to study the student retention problem using data mining. According to literature review conducted in this research, this is the first time that the data from the first-year beginning of the term of new freshmen was used to predict third-year retention. Results of this study are significant in that the factors found for student persistence are mostly from "financial aid" attribute groups, and none of the faculty tenure and experience level attributes contributed to the predictability of the model.

The results of this study indicate that outreach programs that guide the students, especially first-generation students and students from lower socio-economic status, to college success are needed. Information and resources that would help students, who otherwise would lack such knowledge, to enroll in right classes, to study for the classes, and to succeed in the classes.

Another use of data mining in higher education is to test current practices and programs. Using the results obtained in this study, the studied institution can test whether it is using its already stretched resources efficiently. The treatments obtained in this study show the most predictive factors towards third-year retention, therefore, any student success or advising program that falls out of the scope of these rules is less likely to increase student persistence for third-year. In

addition, these rules, when updated with time, can be used to validate current hypotheses and assist administrators make data-driven and informed decisions.

5.3 Future Work

The data mining framework used in this study (and the results obtained using this framework) should serve as a platform for future studies. Some recommendations out of this study are:

- The dollar amounts of the financial-aid data for previous terms should be adjusted for inflation, as these changes would produce theories based on today's dollar amount.
- As it was found that first-year retention prediction is very difficult using the academic and financial aid data, new sources of data (such as exit interviews) should be tested whether they increase the explainability of the problem.
- Although the data sample used in this study was large, a study on larger scale *such as data from multiple universities* can be conducted to observe any geographical effects.
- As financial aid attributes were significant in this study, a longitudinal study of financial aid awards by year can be conducted using data mining to find whether certain award amounts are critical for persistence in a specific year. Although the financial aid award data was grouped by the award types, these data were not broken by type of the financial need i.e. need or merit based. This grouping will help analyzing the effects of need-based financial aid awards.
- A detailed study can be conducted on a group of students with similar characteristics (HS GPA, test scores, family background) to observe institutional effects contributing to retention of students.
- Using text mining techniques, studies can be conducted on qualitative data *such as Facebook data* that represents students' behavior, relationships, and integration with the university .

Bibliography

ACT. ACT National Collegiate Retention and Persistence to Degree Rates, 2007. http://www.act.org/research/policymakers/reports/retain.html. [cited at p. 2, 3]

C.M. Antons and E.N. Maltz. Expanding the role of institutional research at small private universities: A case study in enrollment management using data mining. *New Directions for Institutional Research*, 2006(131):69, 2006. [cited at p. 44, 46]

R. H. Atwell, W. Ding, M. Ehasz, S. Johnson, and M. Wang. Using data mining techniques to predict student development and retention. In *Proceedings of the National Symposium on Student Retention*, 2006. [cited at p. 46, 50, 54, 55, 57]

B.L. Bailey. Let the data talk: Developing models to explain IPEDS graduation rates. *New Directions for Institutional Research*, 2006(131):101–11515, 2006. [cited at p. 44, 46]

Bruce D. Baker and Craig E. Richards. A comparison of conventional linear regression methods and neural networks for forecasting educational spending. *Economics of Education Review*, 18(4):405–415, 1999. [cited at p. 46, 48]

K. Barker, T. Trafalis, and T. R. Rhoads. Learning from student data. *Systems and Information Engineering Design Symposium*, pages 79–86, 2004. [cited at p. 20, 44, 49]

J. P. Bean. Path Analysis: The Development of a Suitable Methodology for the Study of Student Attrition. Paper presented at the Annual Meeting of the American Educational Research Association,San Francisco, California, 1979. [cited at p. 30, 55, 89]

J. P. Bean. Dropouts and turnover: The synthesis and test of a causal model of student attrition. *Research in Higher Education*, 12(2):155–187, 1980. [cited at p. 23, 28, 30, 32, 33, 34, 35, 36, 57]

M.J. Beeson and R.D. Wessel. The impact of working on campus on the academic persistence of freshmen. *Journal of Student Financial Aid*, 32(2):37–45, 2002. [cited at p. 44]

C. Beil, C. A. Reisen, M. C. Zea, and R. C. CapIan. A longitudinal study of the effects of academic and social integration and commitment on retention. *NASPA Journal*, 37 (1), 1999. [cited at p. 23, 42]

S.E. Beitel. *Applying Artificial Intelligence Data Mining Tools to the Challenges of Program Evaluation.* Dissertation, University of Connecticut, 2005. [cited at p. 48]

M. J. Berry and G. Linoff. *Data Mining Techniques: For Marketing, Sales, and Customer Support.* John Wiley & Sons, Inc. New York, NY, USA, 1997. [cited at p. 4, 6, 8]

I. Bhandari, E. Colet, J. Parker, Z. Pines, R. Pratap, and K. Ramanujam. Advanced scout: Data mining and knowledge discovery in NBA data. *Data Mining and Knowledge Discovery*, 1(1):121–125, 1997. [cited at p. 52]

L. Bin, S. Peiji, and L. Juan. Customer churn prediction based on the decision tree in personal handyphone system service. *Service Systems and Service Management, 2007 International Conference on*, pages 1–5, 2007. [cited at p. 52, 53]

Adrian Bors. *Introduction of the Radial Basis Function (RBF) Networks.* [cited at p. 72]

A. Braunstein, M. McGrath, and D. Pescatrice. Measuring the impact of income and financial aid offers on college enrollment decisions. *Research in Higher Education*, 40 (3):247–259, 1999. [cited at p. 43, 44]

M. J. Bresciani and L. Carson. A study of undergraduate persistence by unmet need and percentage of gift aid. *NASPA Journal*, 40(1):104–123, 2002. [cited at p. 43, 44, 57]

P. T. Brinkman and C. McIntyre. Methods and techniques of enrollment forecasting. *New Directions for Institutional Research*, 1997(93):67–80, 1997. [cited at p. 8, 23]

V. Brunsden, M. Davies, M. Shevlin, and M. Bracken. Why do HE students drop out? a test of tinto's model. *Journal of Further and Higher Education*, 24(3):301–310, 2000. [cited at p. 23]

K.C.C. Chan, Au Wai-Ho, and B. Choi. Mining fuzzy rules in a donor database for direct marketing by a charitable organization. In *First IEEE International Conference on Cognitive Informatics*, pages 239–46, Calgary, Alta., Canada, 2002. IEEE Comput. Soc. [cited at p. 52]

L. Chang. Applying data mining to predict college admissions yield: A case study. *New Directions for Institutional Research*, 2006(131), 2006. [cited at p. 21, 44, 46]

William Notz David Moore. *Statistics: concepts and controversies.* 2006. [cited at p. 72]

N. Delavari and M. R. Beikzadeh. A new analysis model for data mining processes in higher educational systems, 2004. [cited at p. 46]

N. Delavari, M.R. Beikzadeh, and S. Phon-Amnuaisuk. Application of enhanced analysis model for data mining processes in higher educational system. *ITHET 6^{th}Annual International Conference* , pages 7–9, July 2005. [cited at p. 46, 109, 114]

C. DeLong, P. M. Radcliffe, and L. S. Gorny. Recruiting for retention: Using data mining and machine learning to leverage the admissions process for improved freshman retention. In *Proceedings of the National Symposium on Student Retention*, 2007. [cited at p. 20, 46, 52, 55, 57]

Department of Education. Integrated Postsecondary Education Data System (IPEDS), 2010. http://nces.ed.gov/ipeds/datacenter/. [cited at p. 65]

S. L. DesJardins, D. A. Ahlburg, and B. P. McCall. A temporal investigation of factors related to timely degree completion. *The Journal of Higher Education*, 73(5):555–581, 2002. [cited at p. 44, 65]

E. L. Dey and A. W. Astin. Statistical alternatives for studying college student retention: A comparative analysis of logit, probit, and linear regression. *Research in Higher Education*, 34(5):569–581, 1993. [cited at p. 39, 41, 57]

Pedro Domingos and Michael J. Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning*, 29(2-3):103–130, 1997. URL citeseer.ist.psu.edu/domingos97optimality.html. [cited at p. 73]

M. J. Druzdzel and C. Glymour. Application of the TETRAD II program to the study of student retention in u.s. colleges. In *Working notes of the AAAI-94 Workshop on Knowledge Discovery in Databases (KDD-94)*, pages 419–430, Seattle, WA, 1994. [cited at p. 1, 44, 48]

E. Durkheim. *Suicide, a study in sociology; translated by John A. Spaulding and George Simpson. Edited, with an introd. by George Simpson.* Free Press, New York, 1951. 854661. [cited at p. 23, 28]

R.G. Ehrenberg and L. Zhang. Do Tenured and Tenure-Track Faculty Matter? *Journal of Human Resources*, 40(3):647, 2005. [cited at p. 65]

A. E. Eiben, T. J. Euverman, W. Kowalczyk, F. Slisser, A. Skowron, and S. K. Pal. Modelling customer retention with statistical techniques, rough data models and genetic programming. *Fuzzy Sets, Rough Sets and Decision Making Processes*, 1998. [cited at p. 52]

P.W. Eykamp. Using data mining to explore which students use advanced placement to reduce time to degree. *New Directions for Institutional Research*, 2006(131):83, 2006. [cited at p. 44, 46]

L. Fausett. *Fundamentals of neural networks: architectures, algorithms, and applications.* Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1994. [cited at p. 15]

U.M. Fayyad and K.B. Irani. On the handling of continuous-valued attributes in decision tree generation. *Machine Learning*, 8(1):87–102, 1992. [cited at p. 19, 66]

R. M. Felder, G. N. Felder, and E. J. Dietz. A longitudinal study of engineering student performance and retention. v. comparisons with traditionally-taught students. *Journal of Engineering Education*, 87(4):469–480, 1998. [cited at p. 42]

Yoav Freund and Llew Mason. The alternating decision tree learning algorithm. In *In Machine Learning: Proceedings of the Sixteenth International Conference*, pages 124–133. Morgan Kaufmann, 1999. [cited at p. 73]

M. Gillespie and J. Noble. Factors affecting student persistence: A longitudinal study. Technical report, ACT Institute, 1992. [cited at p. 23, 39, 40, 42]

J.G. Glynn, P.L. Sauer, and T.E. Miller. Signaling student retention with prematriculation data. *NASPA Journal*, 41(1):41–67, 2003. [cited at p. 53, 57]

J. M. B. Gonzlez and S. L. DesJardins. Artificial neural networks: A new approach to predicting application behavior. *Research in Higher Education*, 43(2):235–258, 2002. [cited at p. 44, 46]

M. Greenberg. How the GI bill changed higher education, June 18, 2004 2004. [cited at p. 1]

M. A. Hall and G. Holmes. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, 15(6): 1437–1447, 2003. [cited at p. 22]

M.A. Hall. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000): June 29-July 2, 2000, Stanford University*, page 359. Morgan Kaufmann, 2000. [cited at p. 71]

J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2006. [cited at p. 5]

D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, Cambridge, MA, 2001. [cited at p. 4]

David Heckerman. A tutorial on learning with bayesian networks. 1996. [cited at p. 73]

S. Herzog. Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. *Research in Higher Education*, 46(8): 883–928, 2005. [cited at p. 42, 43, 57, 65]

S. Herzog. Estimating student retention and degree-completion time: Decision trees and neural networks vis--vis regression. *New Directions for Institutional Research*, 131 (2006), 2006. [cited at p. 21, 46, 50, 51, 53, 57]

R.C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11:63, 1993. [cited at p. 72]

C. Hueglin and F. Vannotti. Data mining techniques to improve forecast accuracy in airline business. In *Conference on Knowledge Discovery and Data Mining*, pages 438–442. ACM Press New York, NY, USA, 2001. [cited at p. 52]

K. H. Im, T. H. Kim, S. Bae, and S. C. Park. Conceptual modeling with neural network for giftedness identification and education. In *Advances in Natural Computation*, volume 3611, page 530. Springer, 2005. [cited at p. 44, 47]

C. Intrasai and V. Avatchanakorn. Genetic data mining algorithm with academic planning application. In *IASTED International Conference on Applied Modeling and Simulation*, pages 286–129, Alberta, Canada, 1998. [cited at p. 46, 48]

T.T. Ishitani and S.L. DesJardins. A longitudinal investigation of dropout from college in the united states. *Journal of College Student Retention: Research, Theory and Practice*, 4(2):173–201, 2002. [cited at p. 42, 43]

T.T. Ishitani and K.G. Snider. Longitudinal effects of college preparation programs on college retention. Annual Forum of the Association for Institutional Research, 2004. [cited at p. 42, 43]

D. Jacoby. Effects of part-time faculty employment on community college graduation rates. *Journal of Higher Education*, 77(6):1081–1103, 2006. [cited at p. 66]

E. P. John. The impact of student aid on recruitment and retention: What the research indicates. *New Directions for Student Services*, pages 61–76, 2000. [cited at p. 43, 44, 65]

K.G. Jöreskog and D. Sörbom. *LISREL 7: A Guide to the Program and Applications*. SPSS, 1989. [cited at p. 36]

TA. Klein. A fresh look at market segments in higher education. *Planning for Higher Education*, 30(1):5, 2001. [cited at p. 1]

D. Kuonen. Data mining and statistics: What is the connection?, 2004. [cited at p. 4]

L. K. Lau. Institutional factors affecting student retention. *Education*, 124(1):126–137, 2003. [cited at p. 2]

V.A. Lotkowski, S.B. Robbins, and R.J. Noeth. The role of academic and non-academic factors in improving college retention. *ACT Office of Policy Research*, 2004. [cited at p. 44, 45, 65]

J. Luan and A. M. Serban. Data mining and its application in higher education. In *Knowledge Management: Building a Competitive Advantage in Higher Education: New Directions for Institutional Research*. Jossey-Bass, 2002. [cited at p. 4, 20, 46]

Y. Ma, B. Liu, C. K. Wong, P. S. Yu, and S. M. Lee. Targeting the right students using data mining. In *Conference on Knowledge Discovery and Data mining*, pages 457–464, Boston, Massachusetts, 2000. ACM Press New York, NY, USA. [cited at p. 44, 47]

S. Massa and P.P. Puliafito. An application of data mining to the problem of the university students' dropout using markov chains. In *Principles of Data Mining and Knowledge Discovery. Third European Conference, PKDD'99*, pages 51–60, Prague, Czech Republic, 1999. [cited at p. 44, 49]

T. Menzies. Data mining class, 2006. http://menzies.us/cs591o. [cited at p. 9, 20]

T. Menzies and Y. Hu. Just enough learning (of association rules): The TAR2 treatment learner. In *Artificial Intelligence Review*, 2007. Available from http://menzies.us/pdf/07tar2.pdf. [cited at p. 76]

T. Menzies, O. Mizuno, Y. Takagi, and T. Kikuno. Explanation vs performance in data mining: A case study with predicting runaway projects, 2007a. http://menzies.us/pdf/07runaway.pdf. [cited at p. 9, 21]

Tim Menzies, Alex Dekhtyar, Justin Distefano, and Jeremy Greenwald. Problems with precision. *IEEE Transactions on Software Engineering*, September 2007b. http://menzies.us/pdf/07precision.pdf. [cited at p. 54]

B. Minaei-Bidgoli, D.A. Kashy, G. Kortmeyer, and W.F. Punch. Predicting student performance: an application of data mining methods with an educational web-based system. In *33rd Annual Frontiers in Education*, pages T2A–13–18 Vol.1, Westminster, CO, USA, 2003. IEEE. [cited at p. 44, 48]

Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997. [cited at p. 71, 72]

P. A. Murtaugh, L. D. Burns, and J. Schuster. Predicting the retention of university students. *Research in Higher Education*, 40(3):355–371, 1999. [cited at p. 41, 42, 57, 65]

A. Nandeshwar. Models for calculating confidence intervals for neural networks. Master's thesis, West Virginia University, 2006. [cited at p. 14]

A. Nandeshwar and S. Chaudhari. Student enrollment prediction model using admissions data: A data mining approach. Las Vegas, NV, October 2007. SAS M2007. http://stat.wvu.edu/~anandesh/CS591o/Project4/. [cited at p. 46]

P. Naplava and N. Snorek. Modeling of student's quality by means of GMDH algorithms. In *Modelling and Simulation 2001. 15th European Simulation Multiconference 2001. ESM'2001*, pages 696–700, Prague, Czech Republic, 2001. [cited at p. 44, 47]

NCPPHE. Retention rates - first-time college freshmen returning their second year (ACT), 2007. [cited at p. 2]

J. Neter, W. Wasserman, and M. H. Kutner. *Applied linear regression models*. Irwin Homewood, Ill, 1989. [cited at p. 12]

K. S. Ng and H. Liu. Customer retention via data mining. *Artificial Intelligence Review*, 14(6):569–590, 2000. [cited at p. 52]

EWT Ngai, L. Xiu, and DCK Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems With Applications*, 2008. [cited at p. 53]

E.N. Ogor. Student academic performance monitoring and evaluation using data mining techniques. *Electronics, Robotics and Automotive Mechanics Conference, 2007. CERMA 2007*, pages 354–359, 2007. [cited at p. 44, 47]

Z.A. Pardos, N.T. Heffernan, B. Anderson, and C.L. Heffernan. Using fine grained skill models to fit student performance with bayesian networks. In *8th International Conference on Intelligent Tutoring Systems (ITS 2006)*, pages 5–12, Jhongli, Taiwan, 2006. [cited at p. 44, 47]

E. T. Pascarella and P. T. Terenzini. Interaction effects in spady and tinto's conceptual models of college attrition. *Sociology of Education*, 52(4):197–210, 1979. [cited at p. 23, 36]

E. T. Pascarella and P. T. Terenzini. Predicting freshman persistence and voluntary dropout decisions from a theoretical model. *The Journal of Higher Education*, 51(1): 60–75, 1980. [cited at p. 23, 36]

Kathleen Pittman. *Comparison of data mining techniques used to predict student retention*. PhD thesis, Nova Southeastern University, 2008. [cited at p. 52, 55, 57]

J.L. Price. *The study of turnover*. Iowa State University Press Ames, 1977. [cited at p. 34]

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986a. [cited at p. 15]

J. R. Quinlan. Improved use of continuous attributes in C4. 5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996. [cited at p. 15]

J. Ross Quinlan. *C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning)*. Morgan Kaufmann, 1 edition, January 1993. [cited at p. 72]

J. Ross Quinlan. *Induction of decision trees*. 1 edition, 1986b. [cited at p. 72]

Irina Rish. An empirical study of the naive bayes classifier. In *IJCAI-01 workshop on "Empirical Methods in AI"*. http://www.intellektik.informatik.tu-darmstadt.de/~tom/IJCAI01/Rish.pdf. [cited at p. 73]

B. P. Roe, H. J. Yang, J. Zhu, Y. Liu, I. Stancu, and G. McGregor. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 543(2-3):577–584, 2005. [cited at p. 52]

P. Rusmevichientong, S. Zhu, and D. Selinger. Identifying early buyers from purchase data. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 671–677, 2004. [cited at p. 52]

A. Salazar, J. Gosalbez, I. Bosch, R. Miralles, and L. Vergara. A case study of knowledge discovery on academic achievement, student desertion and student retention. *Information Technology: Research and Education, 2004. ITRE 2004. 2nd International Conference on*, pages 150–154, 2004. [cited at p. 44, 49]

A.P. Sanjeev and J.M. Zytkow. Discovering enrolment knowledge in university databases. In *First International Conference on Knowledge Discovery and Data Mining*, pages 246–51, Montreal, Que., Canada, 1995. [cited at p. 44, 49, 65]

A. Scalise, M. Besterfield-Sacre, L. Shuman, and H. Wolfe. First term probation: models for identifying high risk students. In *30th Annual Frontiers in Education Conference*, pages F1F/11–16 vol.1, Kansas City, MO, USA, 2000. Stripes Publishing. [cited at p. 1]

Jeffrey A. Schumann. *Data mining methodologies in educational organizations*. Dissertation, University of Connecticut, 2005. [cited at p. 47]

K. A. Smith, R. J. Willis, and M. Brooks. An analysis of customer retention and insurance claim patterns using data mining: a case study. *Journal of the Operational Research Society*, 51(5):532–541, 2000. [cited at p. 52]

K.G. Snider and M. Boston. Longitudinal Effects of College Preparation Programs on College Retention. Paper presented at the Annual Forum of the Association for Institutional Research (AIR), May 28-Jun 2 2004. [cited at p. 42]

W. G. Spady. Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, 1(1):64–85, 1970. [cited at p. 23, 28, 55, 89]

W. G. Spady. Dropouts from higher education: Toward an empirical model. *Interchange*, 2(3):38–62, 1971. [cited at p. 23, 24, 28, 57]

F.K. Stage. Motivation, Academic and Social Integration, and the Early Dropout. *American Educational Research Journal*, 26(3):385–402, 1989. [cited at p. 36, 38, 57]

D. L. Stewart and B. H. Levin. A model to marry recruitment and retention: A case study of prototype development in the new administration of justice program at blue ridge community college, 2001. [cited at p. 20, 44, 49]

S. Sujitparapitaya. Considering student mobility in retention outcomes. *New Directions for Institutional Research*, 2006(131), 2006. [cited at p. 44, 50, 57]

J. F. Superby, J. P. Vandamme, and N. Meskens. Determination of factors influencing the achievement of the first-year university students using data mining methods. In *8th International Conference on Intelligent Tutoring Systems (ITS 2006)*, pages 37–44, Jhongli, Taiwan, 2006. [cited at p. 44, 50]

P. T. Terenzini and E. T. Pascarella. Toward the validation of tinto's model of college student attrition: A review of recent studies. *Research in Higher Education*, 12(3): 271–282, 1980. [cited at p. 36, 37, 57]

E. H. Thomas and N. Galambos. What satisfies students? mining student-opinion data with regression and decision tree analysis. *Research in Higher Education*, 45(3):251–269, 2004. [cited at p. 46, 48]

C. Tillman and P. Burns. Presentation on First Year Experience. `http://www.valdosta.edu/~cgtillma/powerpoint.ppt`. [cited at p. 2]

V. Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1):89–125, 1975. [cited at p. 2, 21, 23, 28, 29, 36, 55, 89]

V. Tinto. Limits of Theory and Practice in Student Attrition. *The Journal of Higher Education*, 53(6):687–700, 1982. [cited at p. 2]

V. Tinto. Stages of student departure: Reflections on the longitudinal character of student leaving. *Journal of Higher Education*, 59(4):438–455, 1988. [cited at p. 21, 23]

V. Tinto. Research and Practice of Student Retention: What Next? *Journal of College Student Retention: Research, Theory and Practice*, 8(1):1–19, 2006. [cited at p. 21]

J.P. Vandamme. Predicting Academic Performance by Data Mining Methods. *Education Economics*, 15(4):405–419, 2007. [cited at p. 2, 44, 47]

W. R. Veitch. Identifying characteristics of high school dropouts: Data mining with a decision tree model, 2004. [cited at p. 44, 49]

G. Waugh, T. Micceri, and P. Takalkar. Using ethnicity, SAT/ACT scores, and high school GPA to predict retention and graduation rates, 1994. [cited at p. 41, 65]

I.H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, San Francisco, 2 edition, 2005. [cited at p. 4, 16, 18, 19, 53]

Y. Yang and G.I. Webb. Proportional k-interval discretization for naïve-Bayes classifiers. *Proceedings of the 12th European Conference on Machine Learning*, pages 564–575, 2001. [cited at p. 19]

Y. Yang and G.I. Webb. Weighted Proportional k-Interval Discretization for Naive-Bayes Classifiers. *Lecture Notes in Artificial Intelligence*, 2637:501–512, 2003. [cited at p. 19]

Chong Ho Yu, Samuel DiGangi, Angel Jannasch-Pennell, Wenjuo Lo, and Charles Kaprolet. A data-mining approach to differentiate predictors of retention between online and traditional students, 2007. [cited at p. 46, 52]

H. Zhang and X. Zhang. Comments on 'data mining static code attributes to learn defect predictors'. *IEEE Transactions on Software Engineering*, September 2007. [cited at p. 53, 54]

J.M. Żytkow and R. Zembowicz. Database exploration in search of regularities. *Journal of Intelligent Information Systems*, 2(1):39–81, 1993. [cited at p. 49]

Appendices

Appendix A

Data Mining in Education Model (Delavari et al., 2005)

Main Process	Sub-Process	Knowledge	Enhanced or New Process Trough Data Mining	Data Mining Function
Evaluation	Student Assessment	The patterns of previous student's learning outcome	Predicting student learning outcome	Prediction
		Prediction of learning outcome	Creating meaningful learning outcome typology in combination with their length of study	Clustering
		The pattern of previous student's successful or unsuccessful in a specific course.	Grouping students into groups of successful and unsuccessful in a specific course	Classification
		The success patterns of high achieved student in a course		
		The patterns of students who show weak test scores	Predicting likelihood of success	Prediction
		The characteristics pattern of high student achiever		
		The patterns of previous students which were likely to be good in a given major		
		The success patterns of previous similar student	Predicting likelihood of persistence	Prediction; Clustering
		Prediction of likelihood of persistence		
		The patterns of previous successful and unsuccessful graduates	Predicting percentage accuracy which student will or will not graduate	Prediction; Clustering
		Prediction of graduation rate	Predicting graduation rate in every trimester	Prediction
		The patterns of previous students who planned for dropping subject	Predicting situations to act before student plans to drop out	Prediction
		Prediction drop-out rate	Predicting drop-out rate in coming trimester	Prediction
		The patterns of previous students who planned for resource allocation	Predicting situations to act before student plan for resource allocation	Prediction
		The patterns of previous male and female students in test score	Associating student personal information (gender, race, age, marital status, nationality) with test score	Association

Table A.1 continued on next page

Main Process	Sub-Process	Knowledge	Enhanced or New Process Trough Data Mining	Data Mining Function
		Association of student personal information with test score		
		The success patterns of previous students who previously had transferred subjects	Predicting likelihood of transferability	Prediction
		Prediction of the likelihood of transferability		
		The patterns of previous students attendance in accordance with test scores	Associating student course taken information with their test score	Association
		Association of student attendance rate and test score		
		Association of student health information and test score	Associating student health information with test score	Association
	Lecturer Assessment	The characteristics patterns of previous lecturers which were more effective than others	Predicting most effective lecturers in a year in accordance with learning outcome	Prediction; Classification
		Previous lecturers patterns in accordance with students test score level	Associating lecturer training with their student test score	Association
		Association of lecturer training with student test score		
		The patterns of most cost-effective courses	Predicting courses which are most cost-effective	Prediction
		Cluster of most cost-effective courses to be offered together	Grouping the courses are to be offered together to be most cost-effective	Clustering
	Course Assessment	The patterns of courses who offered previously to different types of students	Classifying which courses or curriculum work best for which type of student over time	Classification
		Classification of course to various student		
		Association of course to various type of students	Associating the courses or curriculum with various type of student	Association
		The patterns of previous student test score associated with their gender, race, attendance and so on	Predicting factors which are most affected in test score	Prediction
		Prediction of factors most affected in test scores		

Table A.1 continued on next page

Main Process	Sub-Process	Knowledge	Enhanced or New Process Trough Data Mining	Data Mining Function
		The patterns of programs (courses) which produce greatest return and investment in terms of student learning in coming year	Predicting how many programs (courses) produce greatest return and investment in terms of student learning in coming year	Prediction
		Prediction of programs produce the greatest return in terms of student learning outcome		
	Industrial Training Assessment	The patterns of previous training course for different type of student	Classifying the most suitable training course for different type of student	Classification
		Classification of training course to various student		
		Association of training course with various type of student	Associating the training course with various type of student	Association
Registration	Student Course Registration	The patterns of previous students who were taking various subjects	Predicting what type of students are most likely to take part of subjects	Prediction
		Associations of student to the most appropriate subject	Associating student with various type of subject	Association
		Classification of student to the most appropriate subject	Classifying student to the most appropriate subject during their studies	Classification
Performance	Student Performance	Association of student performance with CGPA	Associating student performance with CGPA	Association
		Association of student performance with their academic attitude	Associating student performance with their academic attitude	Association
		Association of student performance with project mark	Associating student performance with project mark	Association
		Association of student performance with lecturer satisfaction	Associating student performance with lecturer satisfaction	Association
		Association of student performance with planned course	Associating student performance with planned course (Time table, sequence of courses)	Association
		Association of student attendance with class situation	Associating student attendance with class situation (Time, Venue)	Association
		Association of student course mark and time and venue of class situation	Associating student course mark with class situation (Time, Venue)	Association

Table A.1 continued on next page

Main Process	Sub-Process	Knowledge	Enhanced or New Process Trough Data Mining	Data Mining Function
	Lecturer Performance	Association of student location with time and venue of class situation	Associating student location with class situation (Time)	Association
		Association of student to the time and venue of various classes	Associating student location with class attendance	Association
		Classification of student to the time and venue of various classes	Classifying student to the most appropriate time and venue for various classes	Classification
		The success pattern of high performed student who are having low CGPA	Predicting likelihood of high performed student who are having low CGPA	Prediction
		The success pattern of high performed student which are having bad attitude	Predicting the likelihood of high performed student which are having bad attitude	Prediction
		The success pattern of good performed student with low lecturer satisfaction	Predicting the likelihood of good performed student with low lecturer satisfaction	Prediction
		Association of lecturer who are not teaching well with student test score	Associating lecturer who are not teaching well with student test score	Association
		Association of lecturer who cancel the class frequently with student test score	Associating lecturer who cancel the class frequently with student test score	Association
		Association of lecturer performance with their attitude	Associating lecturer performance with their attitude	Association
		Association of lecturer personal information with his/her performance	Associating lecturer personal information and his/her performance in the class	Association
		Association lecturer background and his/her performance	Associating lecturer performance with his/her background	Association
		Association lecturer with course background	Associating lecturer with course background	Association
Examination	Student Examination	Association of exam level with student mark	Associating exam level with student mark	Association
		Association of exam level with lecturer class performance	Associating exam level with lecturer class performance	Association
Counseling	Student Behavioral Consulting	The patterns of previous students in an academic environment	Predicting student problem behavior	Prediction

Table A.1 continued on next page

Main Process	**Sub-Process**	**Knowledge**	**Enhanced or New Process Trough Data Mining**	**Data Mining Function**
		Cluster of various student characteristics	Predicting behavior of population cluster	Prediction; Clustering
			Clustering to offer comprehensive characteristics analysis of student	Clustering
	Program Selection	The patterns of previous student who were good in a given program	Associating student to the most appropriate program	Association
	Counseling	Association of student and the most appropriate program	Classifying student to the most appropriate available program in the university	Classification
		Classification of student to the existing programs		

Table A.1: Main Components of the Data Mining for Education Model (Delavari et al., 2005)

Index