LONGITUDINAL STUDY OF FIRST-TIME FRESHMEN USING DATA MINING

Ashutosh R. Nandeshwar

Dissertation proposal 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 Dissertation
in
Industrial Engineering

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

Department of Industrial and Management Systems Engineering

Morgantown, 2009

I dedicate this dissertation to two people, who have shaped my life, and I would not be here writing these words without these two.

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.

Abstract

LONGITUDINAL STUDY OF FIRST-TIME FRESHMEN USING DATA MINING

Ashutosh R. Nandeshwar

In 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. However, previous studies focused more on the social and psychological aspects rather than the data itself, and presented theoretical models on the student retention problem. 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 is to build and analyze data mining models using historical data to predict "high-risk" first-time freshmen students, who are likely to dropout, using data mining.

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, whereas, data mining techniques work well with these conditions. The objective of this research is to study student retention problem using Weka, open-source data mining software, by selecting attributes using feature subset selection (FSS), developing models (trees, rules, bayes, and function based), evaluating models using quartile charts and win-loss tables.

Acknowledgements

Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

Abbreviation	Description	Definition
ANN	Artificial Neural Networks	page 15
CRISP-DM	CRoss Industry Standard Process for Data Mining	page 7
CART	Classification and Regression Tree	page 17
EFC	Expected Family Contribution	page 47
ERP	Enterprise Resource Planning	page 4
ETL	Extract, Transform, and Load	page 5
GMDH	Group Method of Data Handling	page 50
HSGPA	High School GPA	page 48

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). 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.

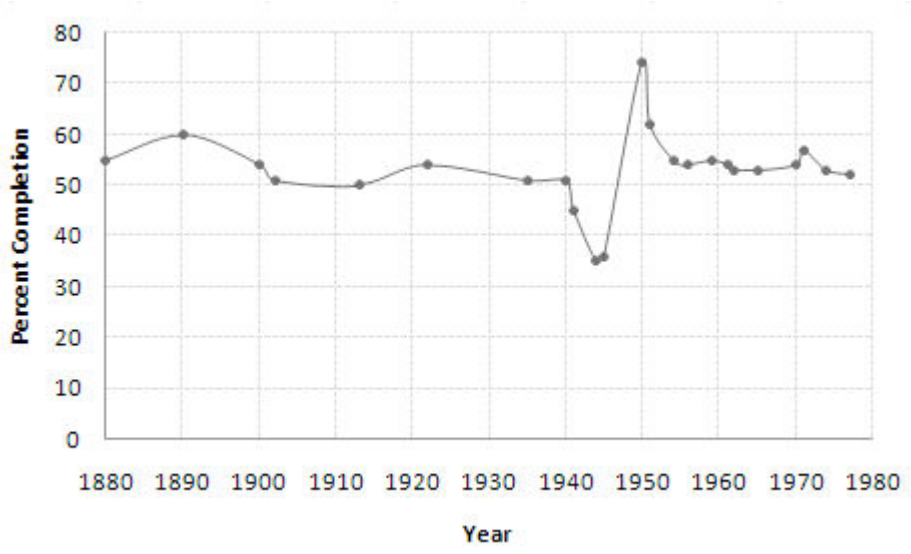
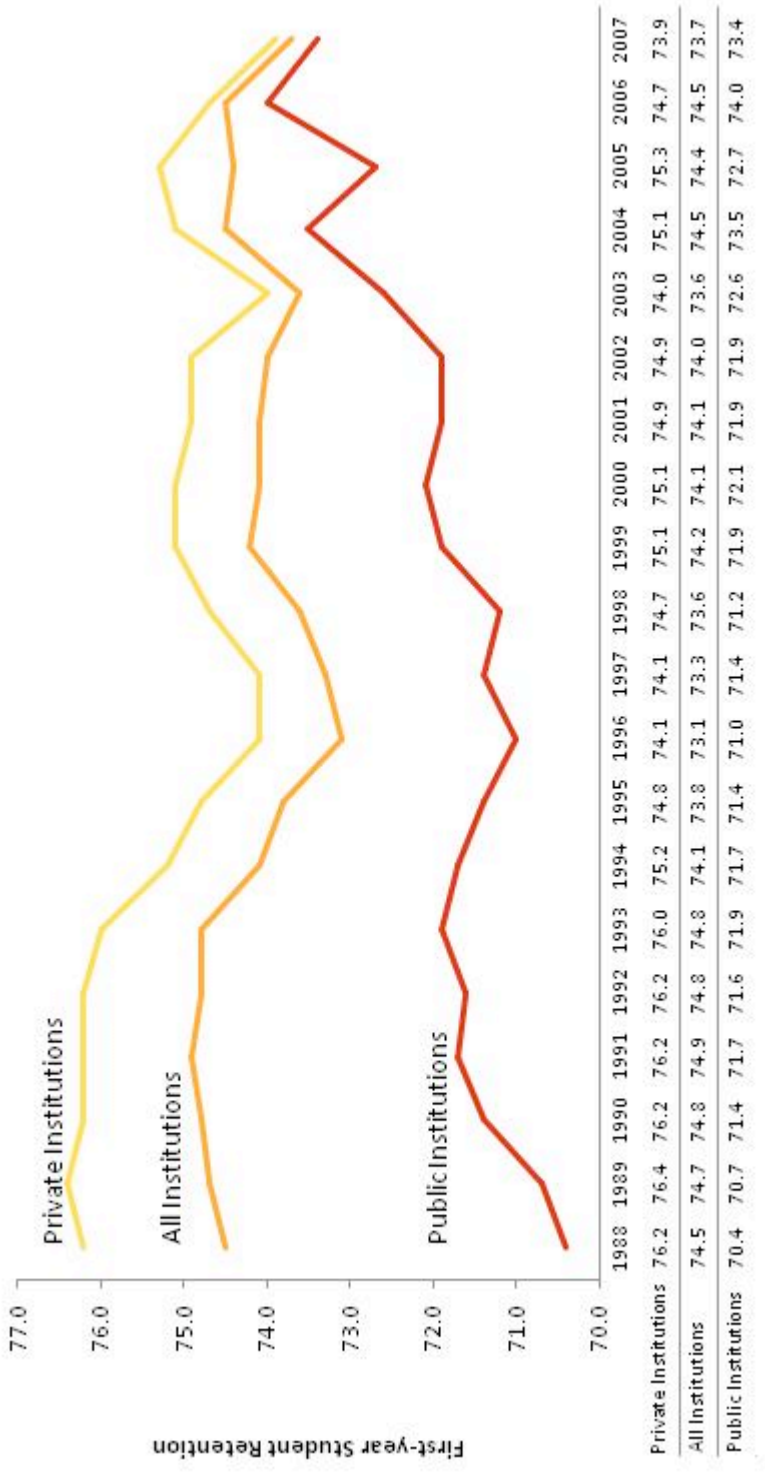


Figure 1.1: BA Degree Completion Rates for the period 1880 to 1980, where Percent Completion is the Number of BAs Divided by the Number of First-time Degree Enrollment Four Years Earlier (Tinto, 1982)



	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)

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.

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 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.

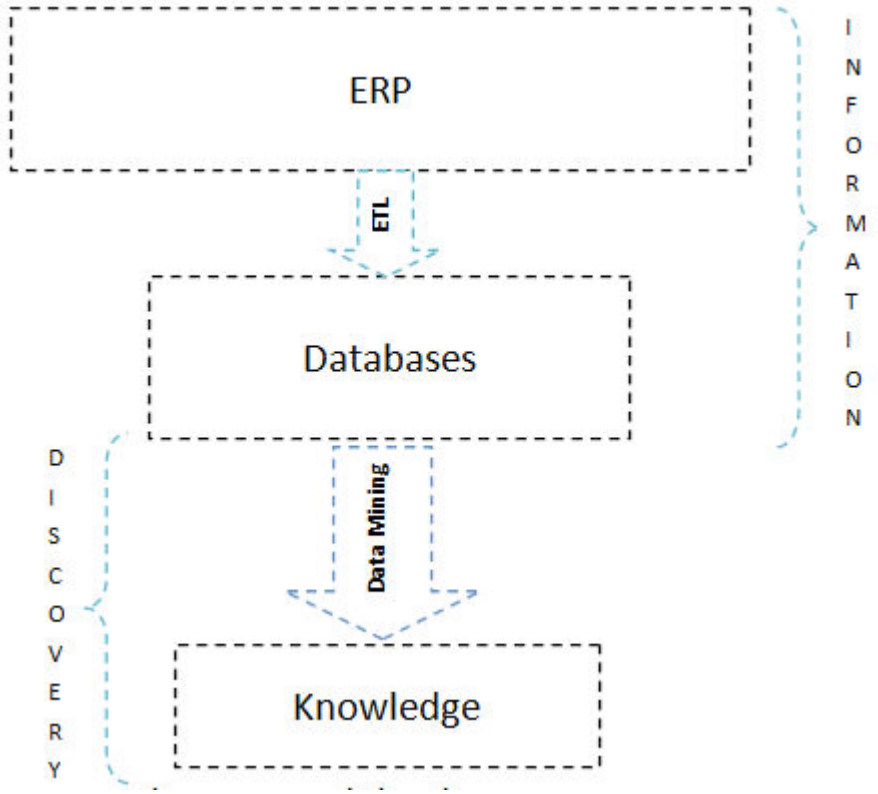


Figure 1.3: Data to Knowledge

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 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.

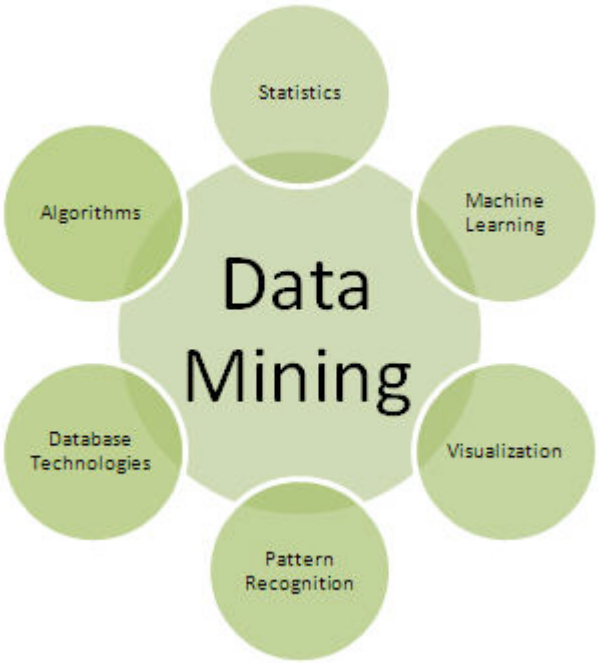


Figure 1.4: Data Mining-Confluence of Multiple Disciplines

> 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.

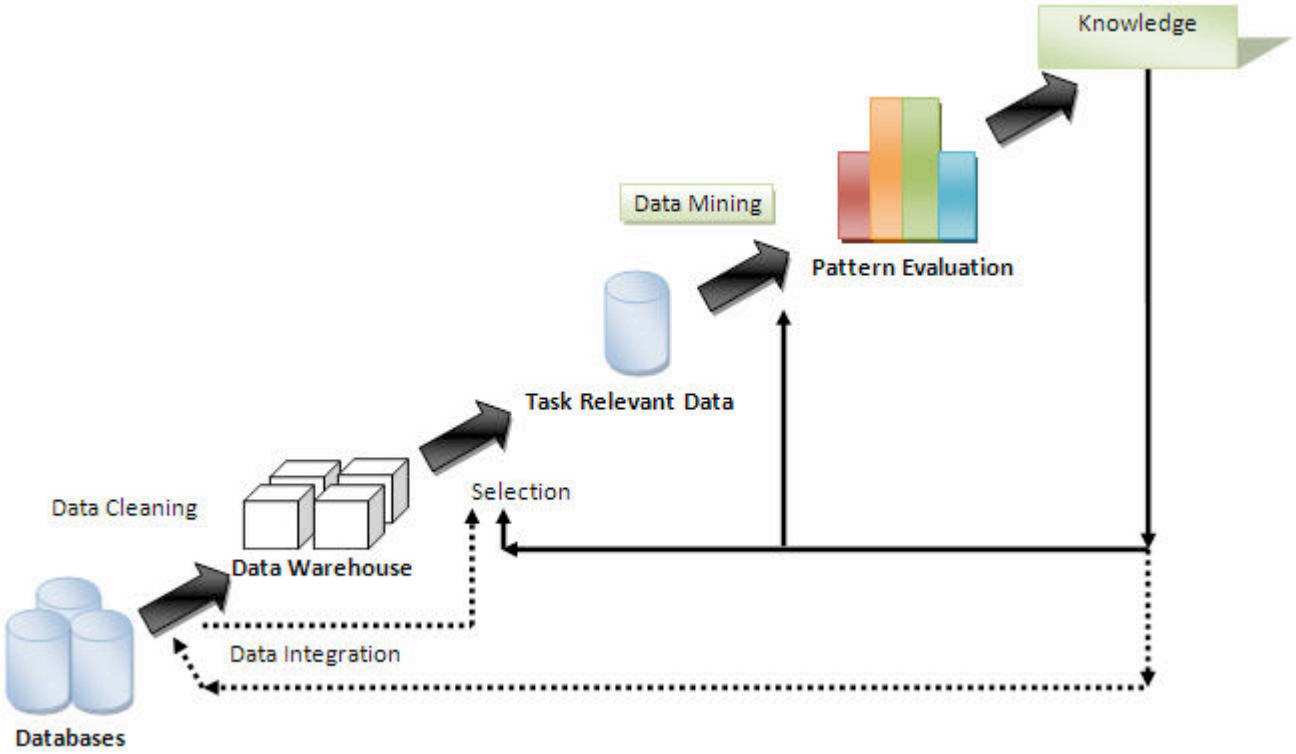


Figure 1.5: Knowledge Discovery Process

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

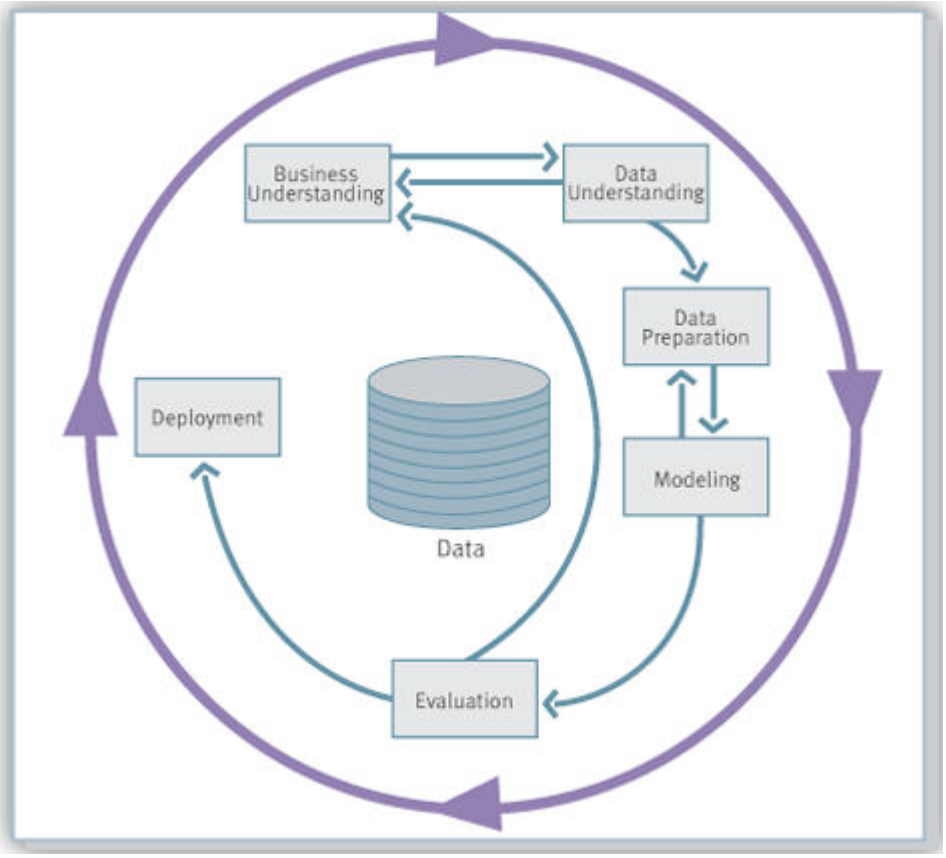


Figure 1.6: CRISP-DM Model Version 1.0

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 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.

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

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. (2007) explained the trade-off between efficiency and explanation of the models: "sometimes the explanatory power must be decreased in order to increase

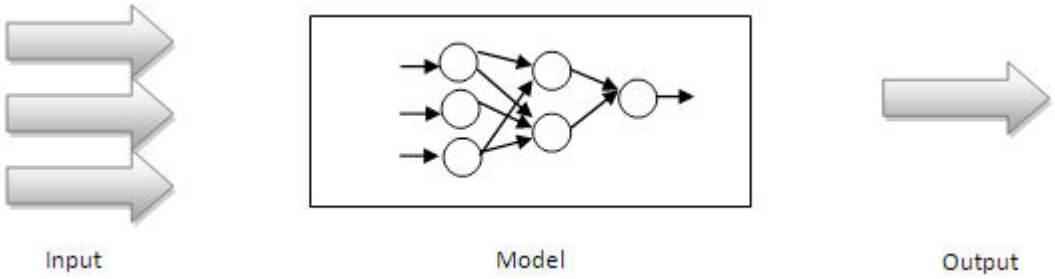


Figure 1.7: Modeling Process

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).

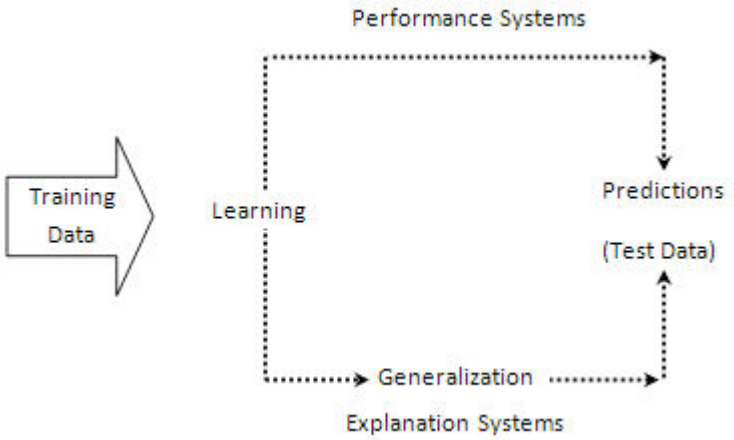


Figure 1.8: Performance vs. Explanation Systems

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, accu-

		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

racy 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

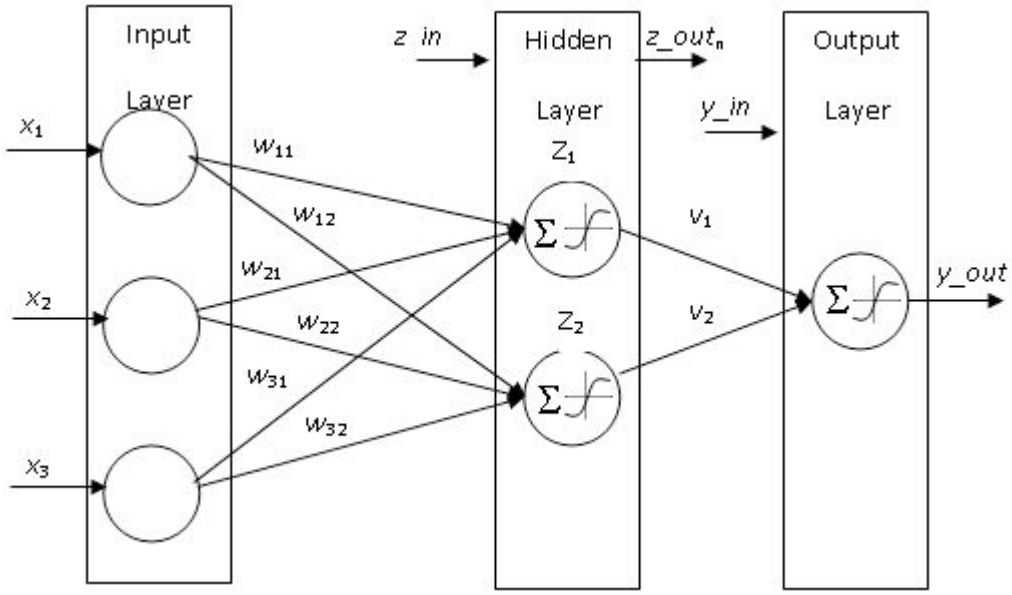


Figure 1.10: Feed-forward Network with 3-2-1 Architecture

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 (1.9) 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.

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

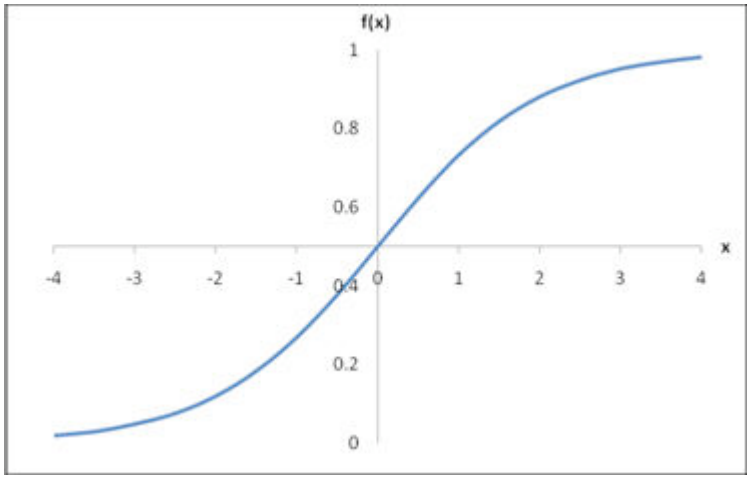


Figure 1.11: Sigmoid or Logistic Activation Function

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

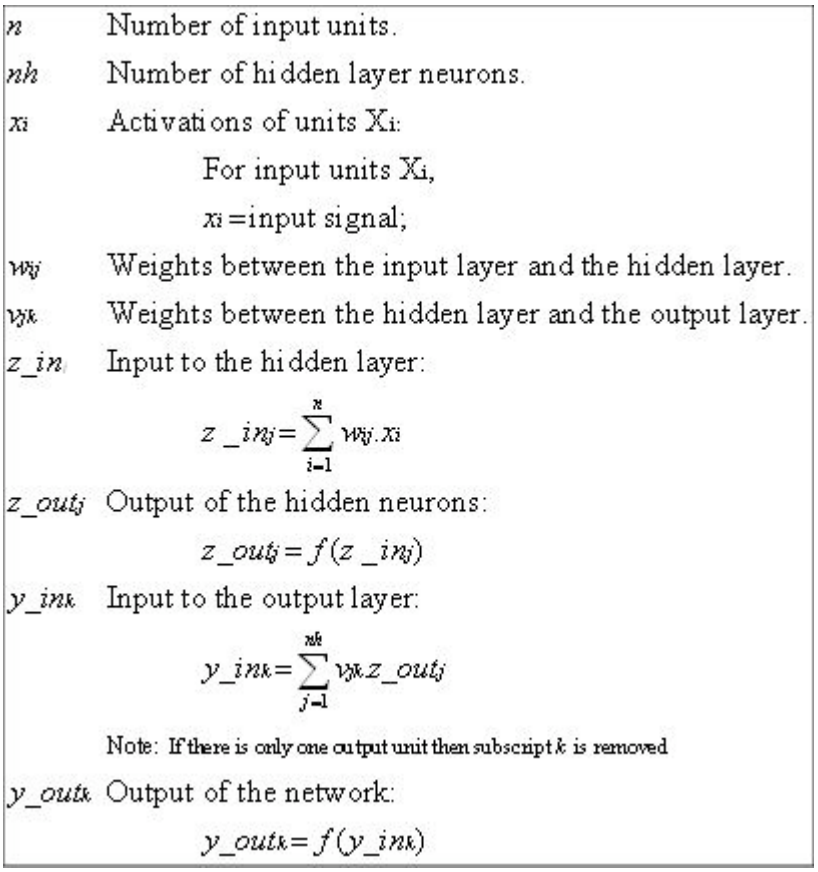


Figure 1.12: Feed-forward Network with one Hidden Layer

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, 1986, 1996). The Gini index is given by Equation 1.14, and the information gain is given by Equation 1.15.

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

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

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).

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

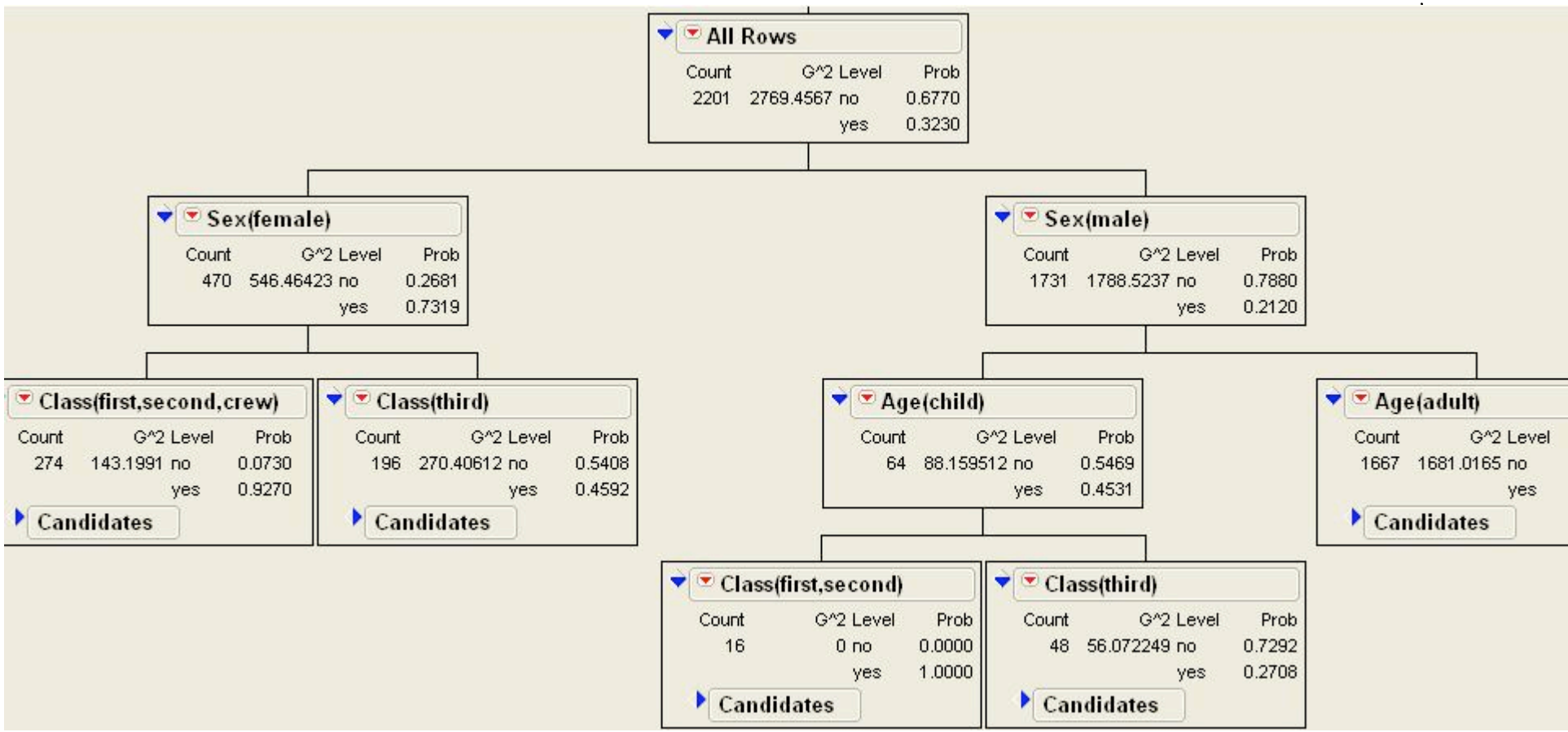


Figure 1.13: Construction of Decision Tree by JMP

```
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

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) commented, "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. (2007) suggested, use of ensemble techniques, such as, discretization, cross-validation, and feature subset selection, can produce high performance and good explanation models. 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. (2007).

1.4 Research Objectives

The major research objectives of this study are:

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 regres-

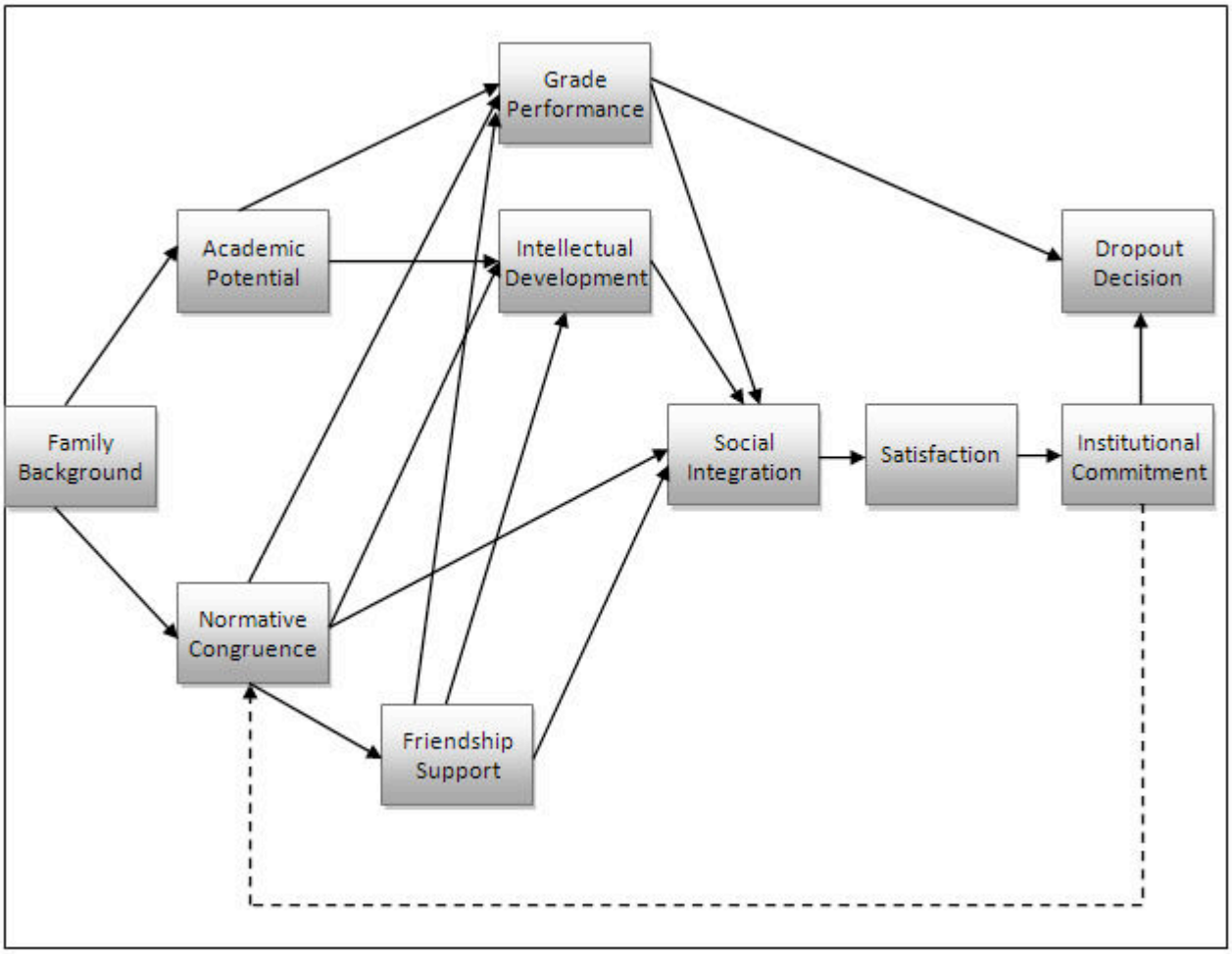


Figure 2.1: Spady's Theoretical Model (Spady, 1971)

sion 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

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

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 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.

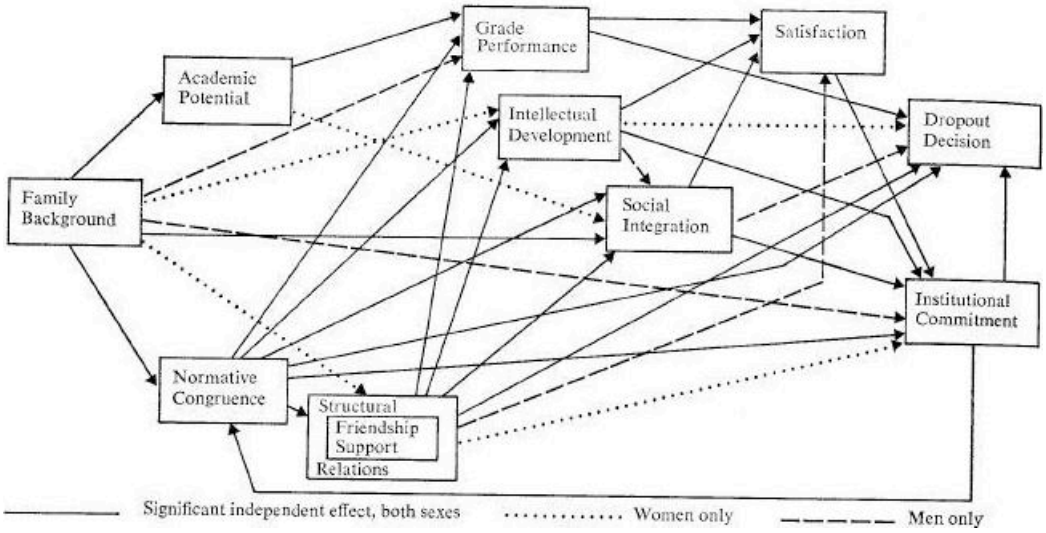


Figure 2.2: Spady's Revised Theoretical Model of the Undergraduate Dropout Process (Spady, 1971)

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

[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.)

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 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);

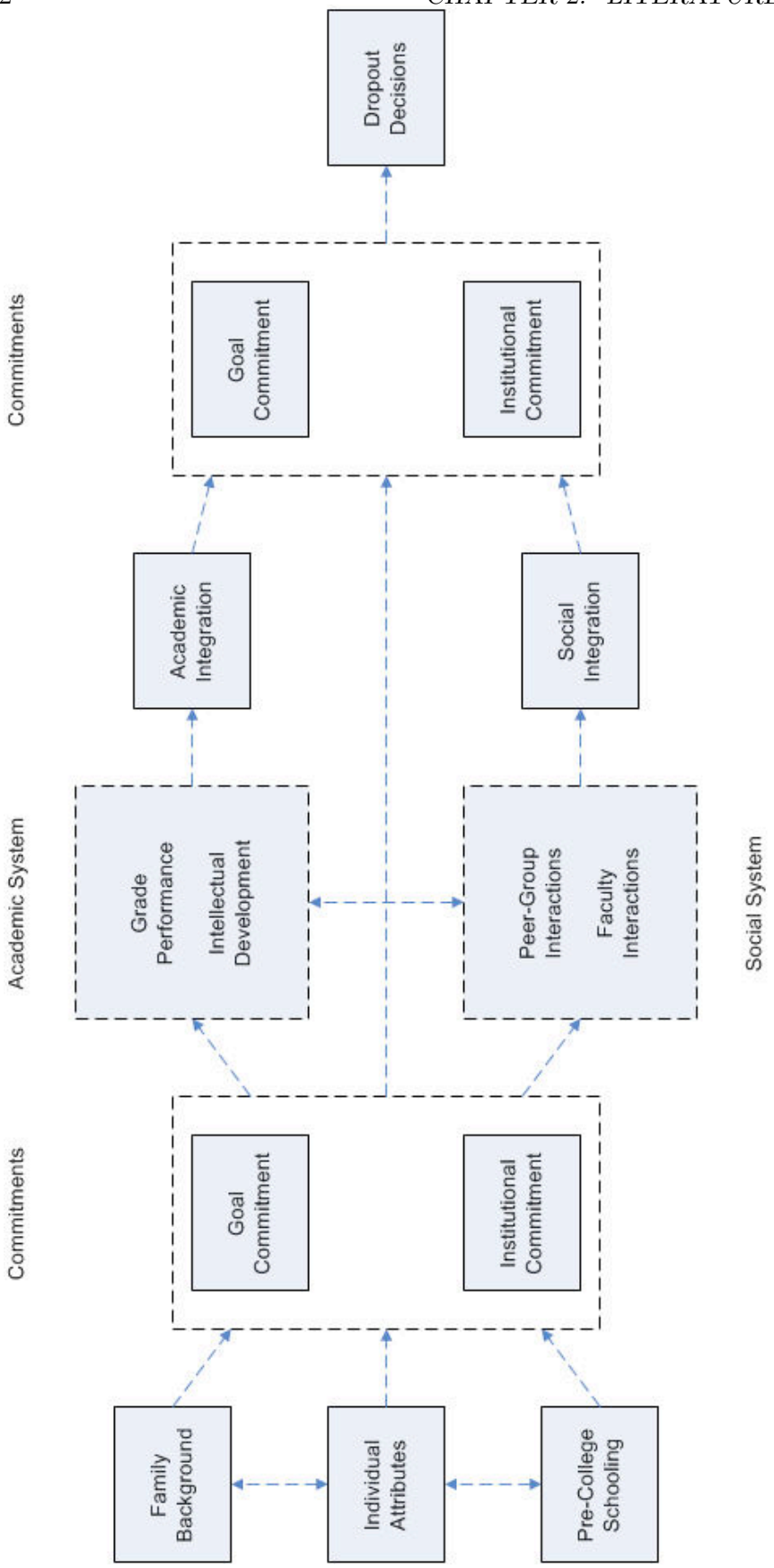


Figure 2.3: Tinto's Model of Student Dropouts (Tinto, 1975)

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

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.

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

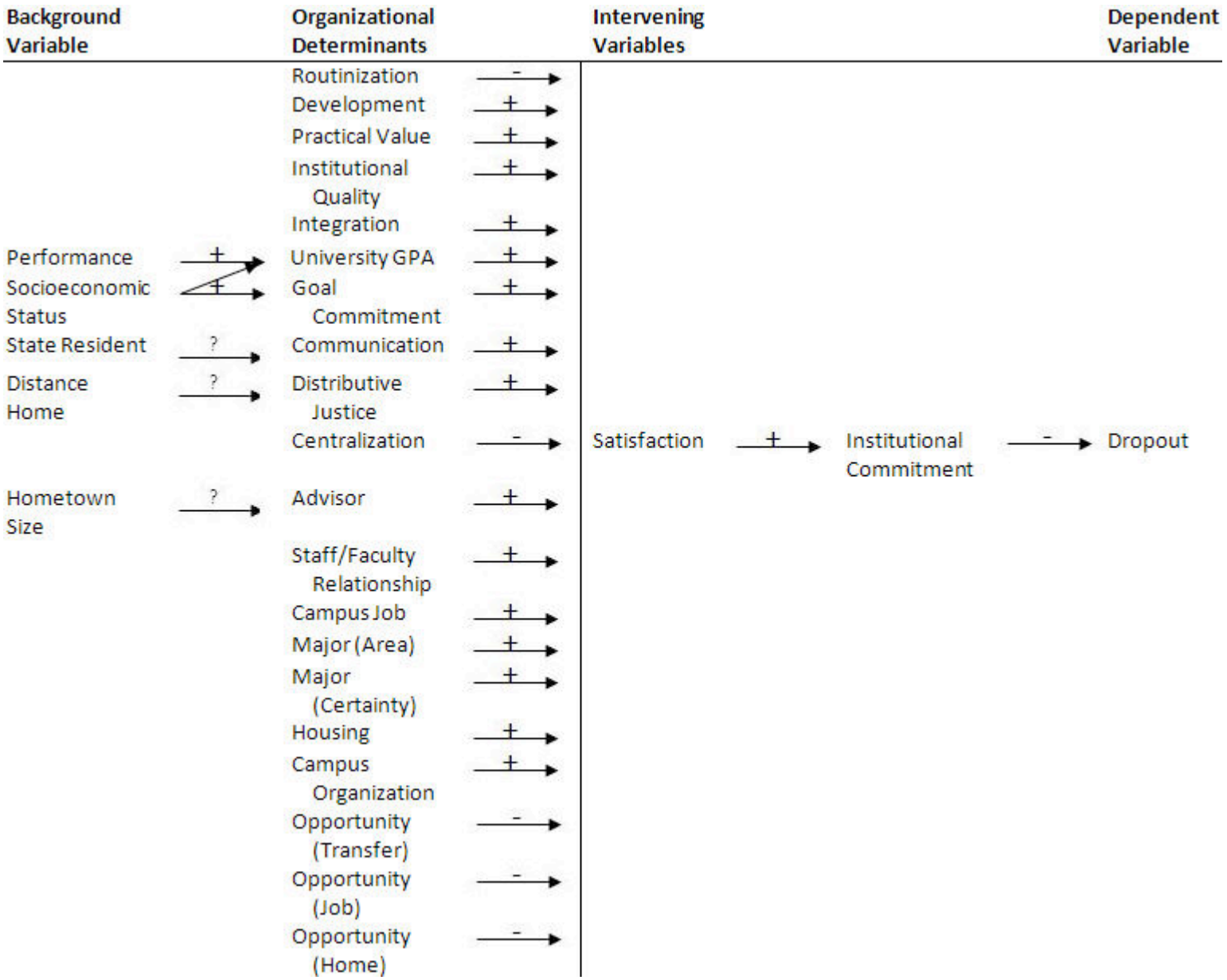


Figure 2.4: Bean's Casual Model of Student Dropout (Bean, 1980)

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)

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.

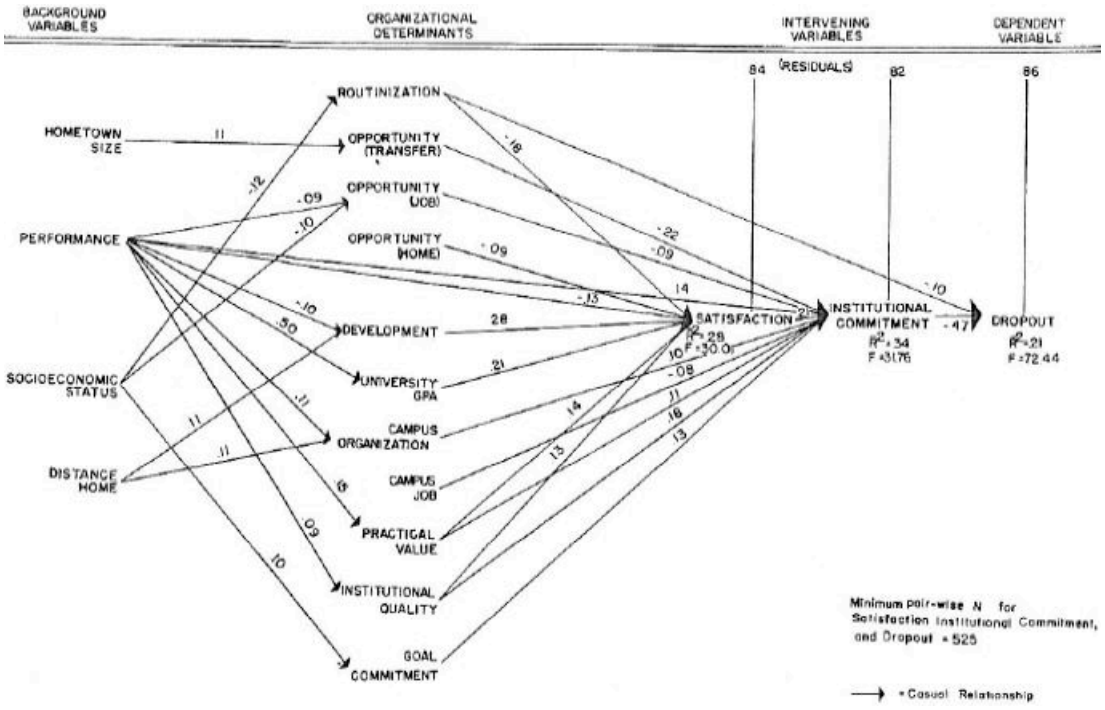


Figure 2.5: Bean's Path Model of Student Attrition for Women (Bean, 1980)

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

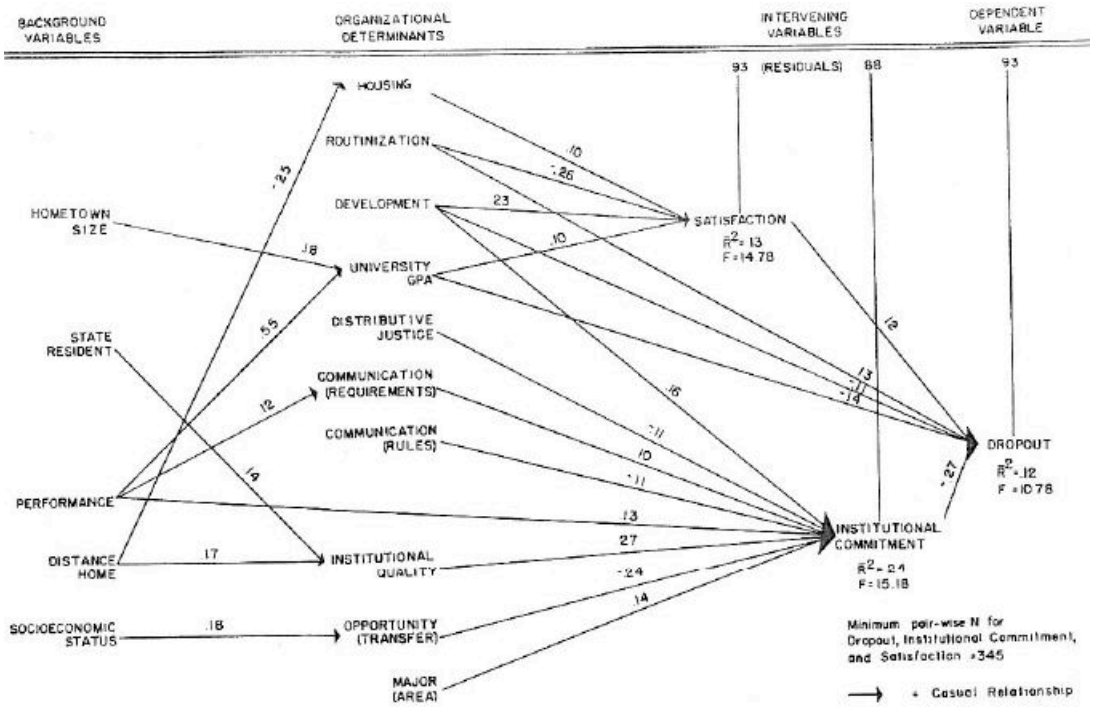


Figure 2.6: Bean's Path Model of Student Attrition for Men (Bean, 1980)

and $R^2 = 0.21$ for women) than the earlier models except 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.

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)

- 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.

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)

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.
- 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 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

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

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)

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.

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

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)

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 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; Sujitparapitaya, 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% prob-

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)

ability 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. 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

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

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 U.S. news college ranking data to find the factors that in-

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

fluenced 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 approxi-

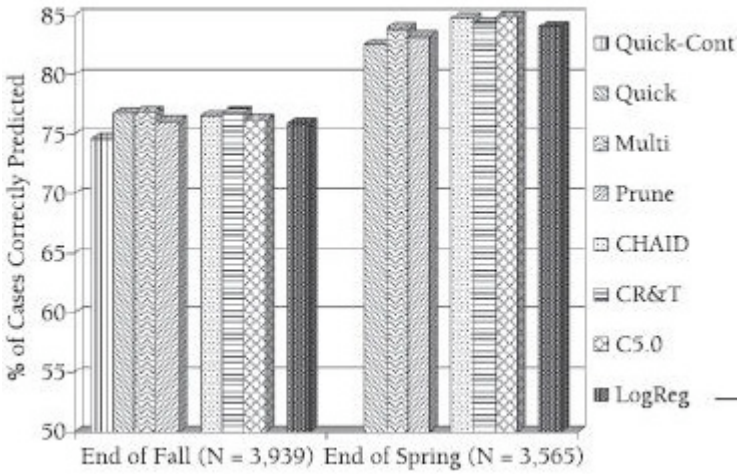
mately 33% for various dataset combinations. The authors used principal component analysis to reduce the number of variables from 56 to 14, however, reported that the results using the reduced datasets were "much worse" than the complete datasets.

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.

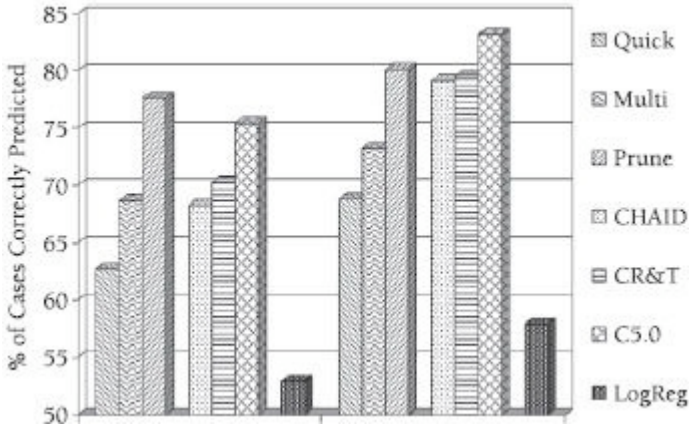
Using the National Student Clearinghouse (NSC) data, Sujitparapitaya (2006) differentiated between stopout, retained, and transfer students. The overall classification rates for the validation sets using logistic regression, neural networks, C5.0 were 80.7%, 84.4%, and 82.1% respectively. Herzog (2006) 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.

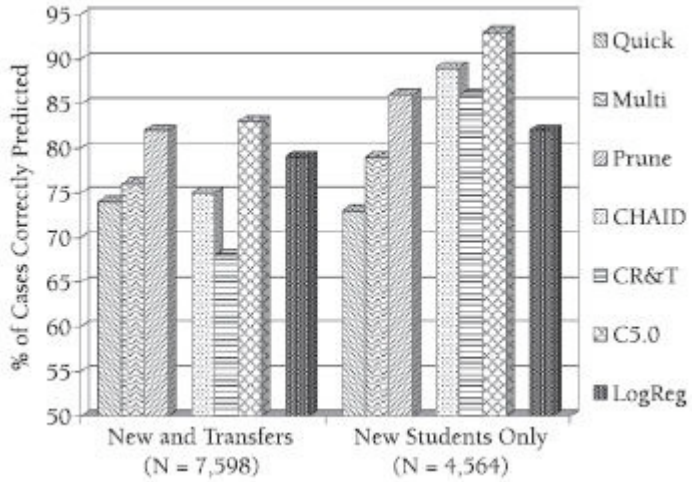
Yu et al. (2007) studied the data from Arizona State University using decision trees, and included variables, such as demographic, pre-college academic performance indicators, current curriculum, and academic achieve-



(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)

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)

ment. 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.

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 cus-

tomer 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.

2.5 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.

Although the use of data mining in the field of education is in 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.

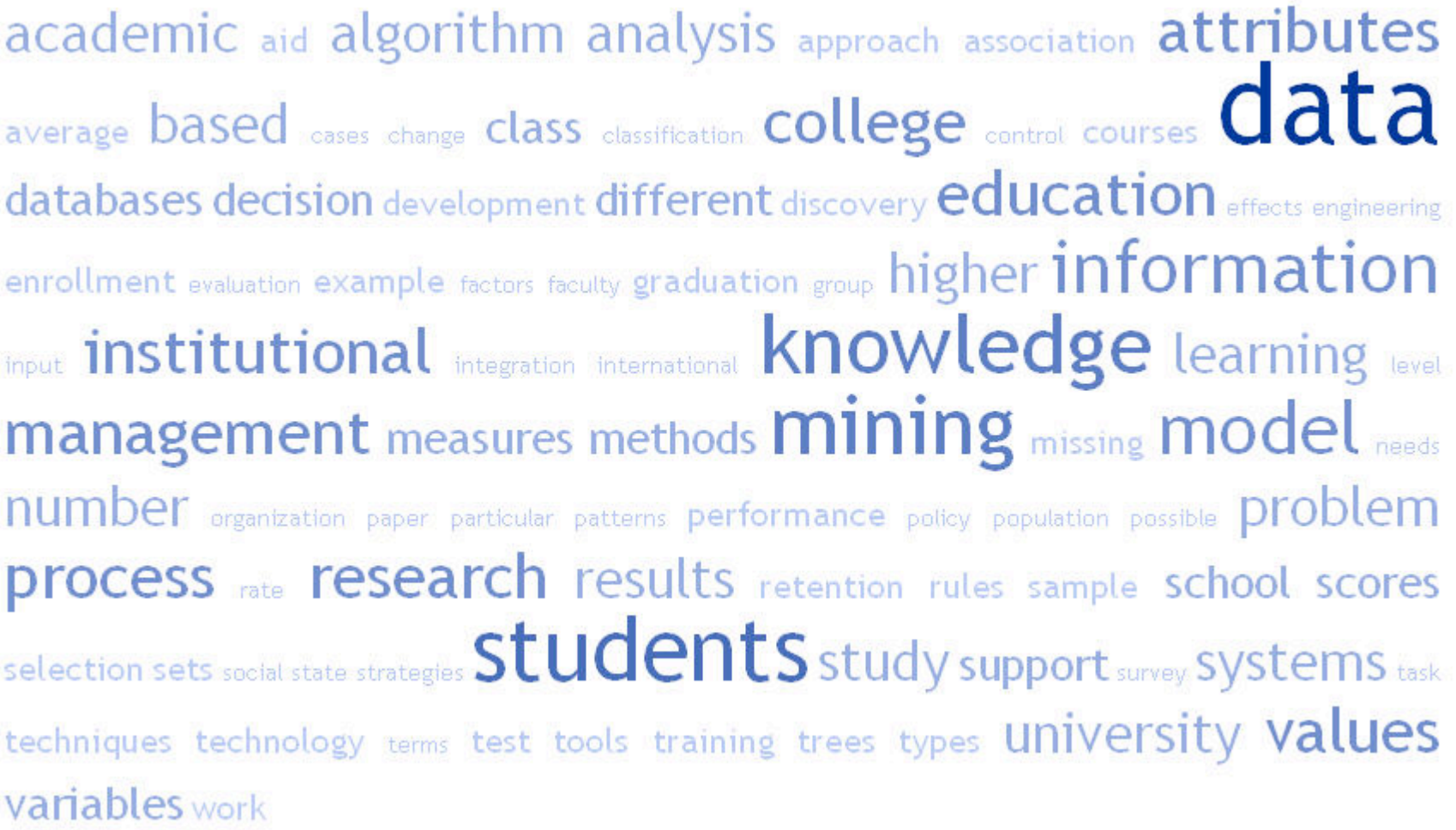


Figure 2.8: Tag Cloud of the Papers Studied in the Literature Review. (Bolder and bigger fonts represent high frequency terms)

Chapter 3

Methodology

> 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

The main objective of this study is to create high performance and explainable models using ensemble methods explained by Menzies et al. (2007). For this study, there are three main sources of data :

1. The Integrated Postsecondary Education Data System (IPEDS): IPEDS is the data collection initiative by National Center for Education Statistics (NCES). IPEDS collects data from all postsecondary education institutes on areas such as, institutional characteristics, graduation, enrollment and retention, faculty and staff, finances, and financial aid. These data would be used to study the effects of various institutional features on retention.

2. The institutional data warehouse: the study institution has collection of historical and current demographic and academic data. These

data would be used to find the effects of demographics and academic performance on retention

3. Freshmen survey data: the study institution conducts a survey, which is based on Tinto's framework, of new freshmen in their first semester; this survey asks questions on student's goals, career objectives, commitment, and other subjects. A well-known consulting firm administers and analyzes these surveys, and predicts scores in various key areas. These data would be used to find the effects of behavioral and motivational indicators on retention.

3.2 Method

All of the three datasets will be analyzed separately using CRISP-DM standards. After the business and data understanding, feature subset selection (FSS), wrappers and filters, and discretization techniques would be used to select variables for modeling. Using these selected variables and the ranks for these variables, these variables would be added sequentially to observe if adding variables to the models make any significant difference. Various combinations of datasets and classifiers would be tested and evaluated using cross-validation, win-loss tables, and quartile charts. The methodology of this research is depicted in Figure 3.1.

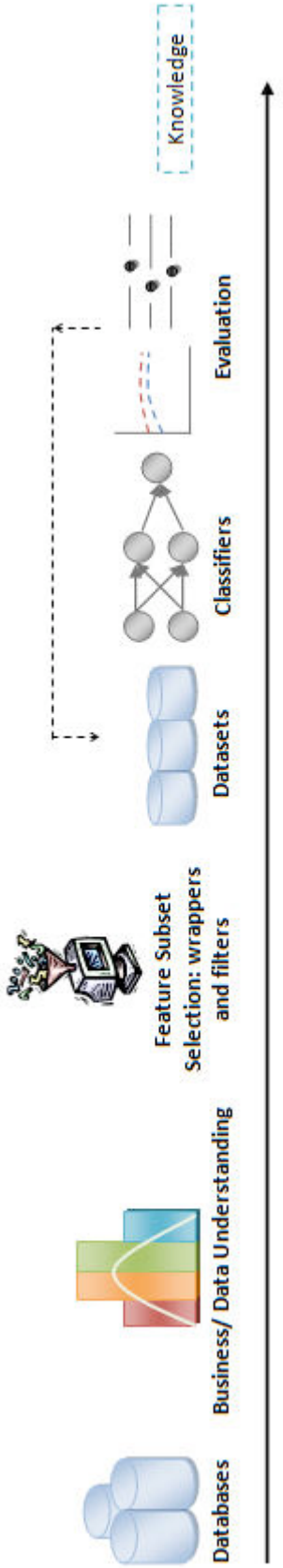


Figure 3.1: Methodology of this Research

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. 48, 50]

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. 48, 54, 56]

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. 48, 50]

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. 48, 52]

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. 22, 48, 53]

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. 31, 57]

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. 25, 30, 31, 34, 35, 36, 37, 38]

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. 47]

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. 25, 46]

S.E. Beitel. *Applying Artificial Intelligence Data Mining Tools to the Challenges of Program Evaluation*. Dissertation, University of Connecticut, 2005. [cited at p. 52]

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, 7, 9]

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. 56]

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. 56, 57]

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. 47]

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. 47]

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. 9, 25]

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. 25]

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. 56]

L. Chang. Applying data mining to predict college admissions yield: A case study. *New Directions for Institutional Research*, 2006(131), 2006. [cited at p. 22, 48]

N. Delavari and M. R. Beikzadeh. A new analysis model for data mining processes in higher educational systems, 2004. [cited at p. 48]

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. 48, 75, 81]

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. 22, 48, 56]

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. 47]

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. 42, 44]

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. 2, 48, 52]

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. 25, 30]

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. 56]

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. 48, 50]

L. Fausett. *Fundamentals of neural networks: architectures, algorithms, and applications.* Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1994. [cited at p. 16]

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. 21]

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. 45]

M. Gillespie and J. Noble. Factors affecting student persistence: A longitudinal study. Technical report, ACT Institute, 1992. [cited at p. 25, 42, 43, 45]

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. 48]

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. 24]

J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2006. [cited at p. 6]

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

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. 45, 46]

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. 22, 48, 54, 55]

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. 56]

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. 48, 51]

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. 48, 52]

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. 46]

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. 46, 47]

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. 47]

K.G. Jöreskog and D. Sörbom. *LISREL 7: A Guide to the Program and Applications*. SPSS, 1989. [cited at p. 41]

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. 5]

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. 47, 49]

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, 22, 48]

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. 48, 51]

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. 48, 53]

T. Menzies. Data mining class, 2006. http://menzies.us/cs591o. [cited at p. 9, 22]

T. Menzies, O. Mizuno, Y. Takagi, and T. Kikuno. Explanation vs performance in data mining: A case study with predicting runaway projects, 2007. http://menzies.us/pdf/07runaway.pdf. [cited at p. 9, 23, 59]

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. 48, 51]

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. 45]

A. Nandeshwar. Models for calculating confidence intervals for neural networks. Master's thesis, West Virginia University, 2006. [cited at p. 15]

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. 50]

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. 48, 50]

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. 13]

K. S. Ng and H. Liu. Customer retention via data mining. *Artificial Intelligence Review*, 14(6):569–590, 2000. [cited at p. 56, 57]

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. 57]

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. 48, 51]

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. 48, 50]

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. 25, 38]

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. 25, 38]

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, 1986. [cited at p. 17]

J. R. Quinlan. Improved use of continuous attributes in C4. 5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996. [cited at p. 17]

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. 56]

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. 56]

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. 48, 53]

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. 48, 53]

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. 2]

Jeffrey A. Schumann. *Data mining methodologies in educational organizations*. Dissertation, University of Connecticut, 2005. [cited at p. 50]

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. 56, 57]

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. 46]

W. G. Spady. Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange*, 1(1):64–85, 1970. [cited at p. 25, 30, 57]

W. G. Spady. Dropouts from higher education: Toward an empirical model. *Interchange*, 2(3):38–62, 1971. [cited at p. 25, 26, 30]

F.K. Stage. Motivation, Academic and Social Integration, and the Early Dropout. *American Educational Research Journal*, 26(3):385–402, 1989. [cited at p. 40, 41]

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. 22, 48, 53]

S. Sujitparapitaya. Considering student mobility in retention outcomes. *New Directions for Institutional Research*, 2006(131), 2006. [cited at p. 48, 54]

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. 48, 54]

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. 38, 39]

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. 48, 52]

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. 4, 23, 25, 30, 31, 32, 38, 57]

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. 23, 25]

J.P. Vandamme. Predicting Academic Performance by Data Mining Methods. *Education Economics*, 15(4):405–419, 2007. [cited at p. 2, 48, 51]

W. R. Veitch. Identifying characteristics of high school dropouts: Data mining with a decision tree model, 2004. [cited at p. 48, 53]

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. 42]

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, 18, 20]

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. 20]

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. 20]

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. 48, 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. 53]

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

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 previous male and female students in test score	Associating student personal information (gender, race, age, marital status, nationality) with test score	Association
		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			

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 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		
		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

Table A.1 continued on next page

Main Process	Sub-Process	Knowledge	Enhanced or New Process Trough Data Mining	Data Mining Function
		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
		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

Table A.1 continued on next page

Main Process	**Sub-Process**	**Knowledge**	**Enhanced or New Process Trough Data Mining**	**Data Mining Function**
	Lecturer Performance	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
		Cluster of various student characteristics	Predicting behavior of population cluster	Prediction; Clustering
			Clustering to offer comprehensive characteristics analysis of student	Clustering

Table A.1 continued on next page

Main Process	**Sub-Process**	**Knowledge**	**Enhanced or New Process Trough Data Mining**	**Data Mining Function**
	Program Selection Counseling	The patterns of previous student who were good in a given program Association of student and the most appropriate program Classification of student to the existing programs	Associating student to the most appropriate program Classifying student to the most appropriate available program in the university	Association Classification

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

Index