CLINICAL DATA ANALYSIS FOR PREDICTION OF URINARY INCONTINENCE

by

SUZAN ARSLANTURK

A dissertation submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN
COMPUTER SCIENCE AND INFORMATICS

2015

Oakland University
Rochester, Michigan

Doctoral Advisory Committee:

Mohammad-Reza Siadat, Ph.D., Chair
Theophilus Ogunyemi, Ph.D.
Guangzhi Qu, Ph.D.
Ishwar Sethi, Ph.D.

© Copyright by Suzan Arslanturk, 2015
All rights reserved

Dedicated to my parents

ACKNOWLEDGMENTS

First of all, I would like to express my deepest gratitude to my advisor, Dr. Mohammad-Reza Siadat, for his excellent support, encouragement and guidance. I would like to thank Dr. Theophilus Ogunyemi for his great support and guiding my research for the past several years. I would also like to thank my committee members Dr. Ishwar Sethi and Dr. Guangzhi Qu who have contributed to this dissertation with numerous suggestions. Special thanks to Dr. Beth Zou who was willing to participate in my final defense.

I am also deeply thankful to my wonderful family for their faith in me, always been there to support me and encourage me with their best wishes. My research would not have been possible without their help.

Suzan Arslanturk

ABSTRACT

CLINICAL DATA ANALYSIS FOR
PREDICTION OF URINARY INCONTINENCE

by

Suzan Arslanturk

Advisor: Mohammad-Reza Siadat, Ph.D.

It is common for clinical data in survey trials to be incomplete and inconsistent for several reasons. One objective of this study is to identify and eliminate inconsistent data as an important data mining preprocessing step. We define three types of incomplete data: missing data due to skip pattern (SPMD), undetermined missing data (UMD) and genuine missing data (GMD). Identifying the type of missing data is another important objective as all missing data types cannot be treated the same. This goal cannot be achieved manually on large data of complex surveys since each subject should be processed individually. Experiments are conducted on a longitudinal questionnaire (MESA). MESA dataset was collected between 1983-1990 to create a set of questions that can reliably predict future Urinary Incontinence (UI). The analyses are accomplished in a mathematical framework by exploiting graph theoretic structure inherent in the questionnaire. An undirected graph is built using mutually inconsistent responses as well as its complement. The responses not in the largest maximal clique of complement graph are considered inconsistent. Further, all potential paths in questionnaire's graph are considered, based on responses of subjects, to identify each type of incomplete data. Once SPMD is determined, MESA data is stratified to divide the data into stratums with potentially different UI risk factors. Rough set imputation is applied, on the GMD portion of the incomplete data. ReliefF attribute selection technique and logistic regression is used to determine the potential predictive factors with their corresponding prediction probabilities forming the continence index on the preprocessed MESA data. The incomplete data analysis results show 15.4% *GMD*, 9.8% *SPMD*, 12.9% *UMD* and 0.021% inconsistent data. Proposed preprocessing methods are prerequisites for any data mining of clinical survey data. The predictive index can be applied for immediate screening and for predicting future urinary incontinence in older woman of comparable demographics.

TABLE OF CONTENTS

TABLE OF CONTENTS- Continued

TABLE OF CONTENTS- Continued

LIST OF TABLES

LIST OF TABLES- Continued

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACG	Adjusted Clinical Groups
CPS	Current Population Survey
DI	Deterministic Imputation
FEG	Failed Edit Graph
GMD	Genuine Missing Data
KNN	K-Nearest Neighbor
MAR	Missing At Random
MCAR	Missing Completely Random
MDS	Minimal Deletion Set
MESA	Medical Epidemiological and Social Aspect of Aging
MSLP	Multiple Sub Linear Path
NIA	National Institute of Aging
NMAR	Not Missing At Random
SI	Random Imputation
SOM	Self-organization Maps
SPMD	Missing Data Due to Skip Pattern
SSLP	Single Sub Linear Path
UI	Urinary Incontinence
UMD	Undetermined Missing Data
WEKA	Waikato Environment for Knowledge Analysis

CHAPTER ONE

INTRODUCTION

1.1. Urinary Incontinence

Urinary incontinence (UI) is a condition in which involuntary urine leakage is demonstrable. UI is commonly seen on older women 60 years and older and has tremendous social and economic costs. It is one of the chronic health conditions that have the greatest effect on a woman's health related quality of life. One in three adult women in the United States suffers from UI. Estimates vary, but in general twice as many women suffer from UI than men. One meta-analysis reported ranges of UI prevalence from 4.5% to 44% (mean 23.5%) in women and 4.6% to 24% (mean 14.5%) in men.

UI is not just a medical problem; it also has adverse social and psychological effects on sufferers and their families by contributing to social isolation and depression. UI also leads to dependency and is a significant factor in nursing home admissions. The economic burden on individuals, families and communities is considerable. Since average life expectancy is increasing and population estimates project that the percentage of women over 65 years old will continue to grow through 2030, the negative implications of UI are likely to increase.

There are three types of urinary incontinence: stress, urge and mixed incontinence. Stress incontinence is due essentially to insufficient strength of the pelvic floor muscles to prevent the passage of urine. It can be caused by coughing, sneezing or movements that put pressure into bladder. Urge incontinence is involuntary loss of urine occurring from no apparent reason while suddenly feeling the need or urge to urinate. This may occur because of damage to nerves of the bladder, the nervous system, or muscles themselves. Mixed incontinence, on the other hand, is a combination of both stress incontinence and urge incontinence. Incontinence varies in degree of severity from several drops to complete bladder emptying. It may occur daily, or many times a day, or only occasionally, perhaps once a month. It may be fairly predictable (low grade stress incontinence) or totally unpredictable (urge incontinence).

A scientifically developed and tested predictive UI index would help identify women who are most likely to develop UI and permit widespread prevention or early treatment. Already identified risk factors include aging, onset during pregnancy tract symptoms, parity, higher body mass index, and functional and cognitive impairment. However, those risk factors were extracted to predict post prostatectomy continence in men after catheter removal by evaluating clinical parameters. This was the only study reported in the literature.

Studies with prospective or longitudinal designs are required to establish the temporal ordering between risk factors and the onset of UI. Furthermore, longitudinal study is needed to determine the role of natural history or medical interventions or factors in inducing UI remission.

1.2. MESA Dataset

The Medical Epidemiological and Social Aspects of Aging (MESA) epidemiology study was conducted with National Institutes of Aging (NIA) funding from

1983-1990. This longitudinal population based study consisted of three detailed household interviews at 1-2 year intervals. The baseline of Medical, Epidemiological and Social Aspects of Aging (MESA) collected in 1983, is a questionnaire containing 825 questions and 1956 respondents, of which the majority, 1154 are female. The female are seniors age 60 years and older. (1,099 subjects age 60-69 years, 589 age 70-79 years, and 268 age 80 and up; 59% women; 91% white). The respondents were interviewed for approximately two hours at home at baseline (1983-1984 interviews) and then re-interviewed at 1-2 year intervals. Re-interview response rates were 69% and 72% in those subjects that were still living. The respondents are interviewed on a variety of health related questions that may play a role in the prevalence of urinary incontinence (UI). Although, the survey focused on the epidemiology of UI, many other attributes were also assessed including medical history, mobility, cognitive function, current health, and quality of life.

One challenge of MESA data is that, it has a considerable amount of incomplete data (37.1%). The incomplete data rates for the first, second and third follow-up are, 65.1%, 43.5%, 64.1%, respectively. There is also, noise and multi-colinearity in the dataset, in which the percentages are unknown at this point.

1.3. Problem Statement

The first step is to preprocess MESA by identifying each type of incomplete data and treating them separately. Next, determining the risk factors that play an active role in the prevalence of UI and generating a predictive index that helps to more readily identify women who are most likely to develop UI are the major steps of this study.

1.4 Related Work

A common challenge of analyzing clinical survey data is to deal with incomplete and inconsistent data. A variety of methods have been developed to enable such analysis of survey data to deal with incomplete data (also referred as missing values) such as imputation, partial deletion, interpolation (Junninen et al. 2004) and maximum likelihood estimation (Beale et al. 1975). Imputation is the most commonly used among the above methods (Heijden et al. 2006; Zhang et al. 2006; Penny et al. 2006). Besides these techniques, an important approach to deal with missing data is to distinguish between different types of missing values so that each group can be treated differently. Another important challenge is to determine and eliminate the inconsistent data, which is considered as noise. Once the inconsistencies are eliminated and the incomplete data are distinguished into subcategories the data is prepared for further analysis with better representation and quality.

We define the types of inconsistent and incomplete data here starting with the definition of inconsistent data. The answer to a branching question determines which alternative set of following questions to be presented to a respondent. When more than one alternative set of questions are answered, it causes inconsistent data. Therefore, inconsistent data occur when subjects do not follow the questionnaire instructions by answering questions that they were not supposed to answer (i.e. questions that they were supposed to skip). Answers given by a non-smoker to an exclusive set of questions that are specific to smoker respondents are an example of inconsistent data.

Data can be incomplete for several reasons: A question can be left unanswered because it is not applicable to a subject. For example, a non-smoking respondent is not supposed to answer a question about the 'number of cigarettes smoked per day' since the questionnaire instructs such a person to skip this question. We refer to this type of missing data that cause data incompleteness as missing data caused by skip patterns (*SPMD*). Most data analysts handle *SPMD*s by recoding the data. If we have a y/n question on smoking and a subsequent question of number of cigarettes smoked per day for smokers, recoding would consist of imputing zero for non-smokers in the latter item. The approach described in this study alleviates the labor of manually detecting and recoding each SPMD. A question that is applicable to a subject can be left unanswered out of negligence, discomfort or other reasons. We refer to this type of missing data as genuine missing data (*GMD*). If a branching question is not answered along with any of the alternative set of following questions it causes undetermined missing data (*UMD*). For example, leaving female surgery related questions as well as all the alternative questions entirely unanswered along with their branching question that asks whether the person had a female surgery causes *UMD*.

As mentioned before, one common approach to resolve incomplete data is to use imputation techniques. There have been several imputation techniques in the literature for filling the missing information in a dataset. One way is to ignore all the entries that have missing values and only focus on a subset of the data with known entries. Another way is to replace the missing entries with a statistical model, usually distributional assumptions such as multinomial normal distribution or iterated linear regression. However, the critical assumption here is that, there is a linear relationship among the variables. Even if there is a non-linear relationship, a different model can still be modelled, but the main point here is to decide what type of a model to choose. A badly chosen model can even show worse results than the ignorance strategy. The basic idea in imputation should be to estimate the missing values by minimizing a loss function.

Penny et al. has studied three different imputation methods to confront missing data in a trauma injury dataset (Penny et al. 2006). The data is collected at a UK hospital and the injury severity related measure that is associated with patient death is missing for 12% of the patients. Three different imputation methods (hot-deck imputation, predictive model based imputation and propensity score imputation) are used. The imputed datasets are classified by artificial neural network and logistic regression. Results show that the complete case analysis (no imputation) shows slightly more accurate results (0.89) than the imputation methods (hot-deck imputation: 0.86, predictive model based imputation: 0.86, propensity score imputation: 0.85). This paper has used different imputation techniques to compare the effectiveness of the imputation methods with complete case analysis. However, they did not consider different types of incomplete data (*UMD, SPMD, GMD*) in their analysis.

Another possible solution to impute the incomplete data is to automatically trace the generated graph for each subject and impute the incomplete data of a particular subject with values of another subject whose pattern shows the highest similarity with that subject's pattern (similarity based imputation). In similarity based imputation, a similarity model between subjects with incomplete data (missing data) is generated, constructing a similarity matrix of subjects and the nearest undifferentiated subject sets of each subject to impute the missing data iteratively. This approach will improve the prediction performance on small sample sizes.

Imputation methods based on rough set theory have been proposed in the literature and shown to be effective. In rough set theory, data is stored in an information table. Let $I = (U, A)$ be an information system, where U be a set of objects and A be a non-empty set of attributes such that $a: U \rightarrow V_a$. With any $P \subseteq A$ there is an equivalence relation $IND(P)$: $IND(P) = \{(x, y) \in U^2 | \forall a \in P, a(x) = a(y)\}$. Let $X \subseteq U$ represent the attribute subset P; and an arbitrary set of objects comprising X, and we wish to express this subset of objects using the equivalence classes induced by attribute subset P. Since X cannot be expressed exactly (because there may cases when some objects will be included and/or excluded) the lower bounds $\underline{P}X = \{x | [x]_P \subseteq X\}$ and upper bounds. $\overline{P}X = \{x | [x]_P \cap X \neq \emptyset\}$ of X need to be defined.

Hu et al. have used a rough set theory based approach in order to impute the missing data. The data is represented in a “condition” → “decision” format where the “condition” are the attributes and the “decision” is the class label (Hu et al. 2014). They proposed an approach where the table is rearranged in such a way where the attribute with the missing values always becomes the decision attribute, and decision rules can be deducted from the attributes excluding the object with the missing attribute. Then the decision rules deducted can be used to replace the missing values. They have introduced the concept of roughness of rearrangement. The roughness of rearrangement on concept Y can be defined as the following: $\beta(\mathcal{T}_Y) = \frac{|\overline{A}(Y) - \underline{A}(Y)|}{\overline{A}(Y)}$. Once the roughness of rearrangements are all calculated and sorted in the descending order they end up with a list of attributes a_{k_1}….a_{k_n}. The roughness of an attribute determines the relationship between the attributes and the decision. If the relationship is strong, the decision rules derived from the rough set theory can be used to determine the value of the missing items; otherwise the missing values can be imputed by using other approaches. Optimal logical flow indicates the logical relationships among the attributes in an information table under a group of selected concepts. The last attribute in the ordered list, a_{k_n}, which yields the smallest roughness value, is the optimal logic attribute, i.e. the attribute with the strongest attribute decision relation. If the optimal logic attribute's roughness is less than a predetermined threshold value, then decision rules can be derived and the missing values can be imputed based on the decision rules. Otherwise, any other traditional method can be used.

Erden et al. have used rough set theory in order to calculate the accuracy of the imputation and to determine decision rules in a real life data from the US government website including the traffic accidents that took place in the USA in 2011 according to their occurrence reason in order to discover useful knowledge from the dataset (Erden et al. 2014). The dataset contains a collection of data about people involved in car accidents with fatalities, the final injuries, and alcohol/drugs tests. The decision parameter is chosen to be the fatalities in the accidents. The condition variable, fatalities, is discretized taking either "1" or "2" as "1" denoting 1 person and "2" denoting 2 or more people. The Expectation Maximization algorithm is used to impute the data. The lower and upper approximations for decision variables $d = 1$ and $d = 2$ are calculated. The accuracy, α, of the approximations are calculated $(\alpha(d,m) = \underline{A}(X)/\overline{A}(X), where\ m = 1\ or\ 2)$ after imputation. We have also used a rough set based approach in order to impute the incomplete data in our dataset. However, the imputation is not applied on the entire set of incomplete data, but only on the *GMDs* in order to minimize the misinformation imputation may lead.

Heijden et al. used a clinical data of 398 subjects that have missing values to evaluate different imputation techniques (missing indicator method, single imputation of unconditional and conditional mean, and multiple imputations). Their aim is to diagnose the presence or absence of pulmonary embolism (Heijden et al. 2006). By using multivariable logistic regression analysis, a diagnostic prediction model is trained. Finally, the area under the ROC curve for complete case analysis, missing-indicator method, single imputation of unconditional and conditional mean, and multiple imputations are 0.794, 0.813, 0.775, 0.792, and 0.787, respectively. In conclusion, the risk factors obtained from the data based on the complete case analysis were more biased than imputation techniques studied in this paper. Single imputation methods perform well because of the low overall number of missing values.

Zhang et al. has studied a clustering based imputation technique for data preprocessing (Zhang 2006). The dataset is first divided into clusters excluding the instances with missing values. Each instance with missing values is then assigned to the most similar cluster. The missing values are then replaced with values generated using kernel based methods: Deterministic Imputation (DI) and Random Imputation (SI). Simulation data is used to evaluate the effectiveness of the strategy under missing rates of

5%, 10% and 40%. The results show that in most cases filling the missing values by the proposed clustering strategy lead to lower error rate than without using it in a clustering task. The SI shows better results than DI when the dataset is divided into clusters. The missing data in this paper is MCAR. That is, if the probability of a missing observation does not depend on its measurement or on other observed or unobserved measurements then the observation is MCAR (Little and Rubin 1987). Note that there is no clear mapping between the two sets of definitions MCAR, MAR, NMAR, and GMD, UMD, SPMD. In the former the definitions are based on dependencies on observed or unobserved measurements. While in latter the definitions are based on the structure of questionnaire and in particular branching questions. This fundamental difference in the bases on which such definitions is formed prevents a clear mapping of one set to the other.

Zhong has experimented with health related individual level survey data that has incompleteness (Zhong 2009). The missing variables are imputed by concentration indices. Concentration indices quantify the degree of inequality across the distribution of a variable. Missing values on health variables are unlikely to be always MCAR. A possible solution to correct the introduced bias due to the concentration indices is proposed. The imputation results are discussed when the data is MCAR, MAR or NMAR. A case study and a simulation study are used. In addition to the type of missing data, the imputation technique being utilized determines the effectiveness of the procedure. Therefore, choosing the proper imputation technique based on the data and missing data models can result in an unbiased treatment of NMAR, MAR, MCAR. In our study, on the other hand, SPMD missing data type is not supposed to be imputed based on the instructions of the questionnaire. Hence, performing any imputation on SPMDs will introduce misinformation. In the case of UMD, the mutually exclusive paths of questions are unanswered along with the branching question itself. So, it is ambiguous which alternative set of questions (branch) is supposed to be answered by the respondent (GMD) and which set(s) is (are) supposed to be skipped (SPMD). Therefore, imputing all UMD questions results in imputing their corresponding SPMDs as well. However, if one can determine which alternative set of questions (branch) in a UMD is supposed to be answered (GMD) and which one(s) are supposed to be skipped (SPMD), then the SPMD portion should not be imputed. However, the GMD missing data type and GMD branch of UMD can be treated where an analyst may identify the type of missing data, whether MCAR, MAR or NMAR, to determine a statistically valid imputation technique.

Yuanyuan et al. have used a nearest neighbor imputation in wireless sensor networks (Yuanyuan 2014). Ambler et al. investigated a number of methods for imputing missing data to evaluate their effect on risk model estimation and the reliability of the predictions. A large national cardiac surgery database is used in this study (Ambler 2007). Jerez et al.(Jerez 2010) have evaluated the performance of several statistical and machine learning imputation methods to predict the recurrence in a real breast cancer dataset. The imputation techniques studied are mean, hot-deck and multiple imputation, and machine learning techniques, e.g., multi-layer perceptron (MLP), self-organization maps (SOM) and k-nearest neighbor (KNN). The imputation methods based on machine learning algorithms outperformed imputation statistical methods in the prediction of patient outcome. Ouzienko et al. has proposed an imputation technique on longitudinal social surveys. The experiments are constructed on synthetic and real life datasets with 20% to 60% of nodes missing (Ouzienko 2014). As mentioned before, SPMD type missing data are not supposed to be imputed. This important notion is not of a major concern in any of these studies.

Dillman et al. on the other hand, has considered *SPMD*, *GMD* and inconsistent data, but not *UMD*. They studied the skip pattern compliance in three test forms. They defined two terms: loop error and gap error. The loop error is defined as questions that were not skipped by the respondent when they were supposed to, and the gap error is defined as questions that were skipped by the respondent when they were not supposed to. The loop error corresponds to our inconsistent data and the gap error corresponds to our *GMD*. This study shows that there is some likelihood that getting perfect compliance that reduces loop error may increase the likelihood of making gap errors. Three test forms (Census Form, Arrow Form and Right Box Form) are tested and compared in terms of their error rates. The three test forms only vary with regard to how the skip pattern instructions are provided. The loop error rates for three test forms are 11.4%, 4.3%, 5.2% and the gap error rates for three test forms are 0.8%, 1.8% and 1.6%, respectively (Dillman et al. 1999). This work is one of the very few studies about skip patterns in the literature. Unlike our study, they focus on human factor analysis in order to make it easy for people to understand how to design a questionnaire with skip patterns such that it lowers the rate of error. However, they do not perform any analysis on determining the skip patterns.

Fagan and Greenberg's interesting report has inspired and given us the basis on which we have built a comprehensive mathematical framework with all required definitions, lemmas and theorems as well as their proofs (Fagan and Greenberg 1988). They focus on skip pattern analysis on questionnaires using graph theory. The missing values are grouped into undetermined data, skip patterns and missing data, which correspond to UMD, SPMD and GMD, respectively, in our study. Their report has not provided an accurate account on how to determine GMD, UMD and SPMD as it considers the questions (vertices of the questionnaire graph) and not the answers (edges of the graph). They have not applied their proposed method on real data nor have they proposed any algorithm to implement it. As a result, they have not performed complexity analysis and they have not validated their method using simulated data.

The UMD's or GMD's determined from data do not provide valuable information for the analyses without any imputation. On the other hand, the SPMD's can be used for data stratification. It is important to stratify the data into stratums which will provide the opportunity to evaluate each stratum separately.

In clinical survey data, it is common for subjects to have their own unique set of answers to the questionnaire. However, the subjects can be grouped into populations based on common answers to specific questions. It is important to divide the dataset into sub-populations based on these questions that have common answers for each population. This will help us to investigate heterogeneous results, or to answer specific questions about particular patient groups and to see whether and how risk factors vary across sub populations. This approach leads us to extract maximum amount of information from the data and gives the clinicians the possibility to apply different treatments for different groups of people. It is important based on what questions to divide (stratify) the population into groups, which we will refer as the stratifying factors throughout this study. This part of the study proposes a method that shows how to stratify a population based on a simulated study and a longitudinal clinical survey data. Here, we use the branching questions to stratify the population. The idea is to find the significant branching questions that divide the population into well represented groups.

There have been several studies in sub group analyses. Su et al. have focused on a comparative study where two or more treatments are compared and how the treatment effect varies across subgroups induced by covariates. Treatment effect can be defined as the amount of change in a condition or symptom because of receiving a treatment compared to not receiving the treatment. They have considered a binary treatment effect (0 or 1), a continuous output and a number of covariates where the components are of mixed types (categorical and continuous). They have used a tree-structured subgroup analyses algorithm since; the tree algorithm is a well-known tool for determining the interactions between the treatment and the covariates. Their goal in subgroup analysis is to find out whether there exist subgroups of individuals in which the treatment shows heterogeneous effects, and if so, how the treatment effect varies across them. By recursively partitioning the data into two subgroups that show the greatest heterogeneity in the treatment effect, they were able to optimize the subgroup analyses. They have used simulated studies to validate their approach. Also, they have used the Current Population Survey (*CPS*) database conducted by the U.S Census Bureau for the Bureau of Labor and

Statistics, in 2004. The CPS is a survey data of 60.000 households. The investigators were interested in specific subgroups of the working population where the pay gap between sexes is dominant. The questions in the survey were related to some demographic characteristics of the respondents, the employment status, hours worked and the income earned from their work. There were different covariates in the data such as gender, age, education, race, citizenship, tax status, etc. The results show that for most of the subgroups that constitute the majority of the population, women are paid significantly less than men. Also, the wage disparity between men and women varies with the industry, occupation and age. In our study, on the other hand, instead of recursively partitioning the data into sub populations, we are using the branching questions leading the skip patterns to occur as our stratifying factor.

Subgroup analyses are a highly subjective process since the subgroups themselves as well as the number of subgroups are determined by the investigator beforehand (Assmann et al. 2000). It is important to determine which specific subgroup to use in the experiment. The incorrect selection of the subgroups may cause unreliable results. Therefore, significance testing is a common approach in subgroup analyses. That is, testing the numerous plausible possibilities to see which subgroup performs better. However, this approach cannot be considered as an efficient way of splitting the data. We are utilizing the wisdom of experts embedded in the data through the questionnaire design processes when selecting the branching questions as stratifying factors.

There have been several studies on determining variables that are for understanding the underlying phenomena of interest. It is important to reduce the dimension of original data prior to any modeling. Different attribute selection techniques have been used in order to reduce the dimensionality and the computational complexity (Azhagusundari et al.). The attributes that are significant can be extracted, ranked and weights can be assigned to each attribute to compare the significance. Decision trees based on information gain techniques have been widely used in order to perform feature selection. Decision trees divide the population into subgroups recursively until the leaf nodes represents the class labels. However, in this study none of these techniques are used to determine the significant branching questions. The feature selection techniques cannot be directly used in order to determine the significant branching questions, since the class labels of individual subjects are irrelevant. Instead high support and confidence for prospective extracted rules from each stratum is of interest. That is, a population with mixed class labels in a stratum would be favorable as long as it lends itself to rules with high confidence factor and support. For instance, smoking could be a significant branching question if the rules applied to smokers are different than those applied to non-smokers. However, there could be populations with mixed class labels in smokers and non-smokers groups.

Risk stratification in clinical data is used to divide patients into different acuity levels and to determine a person's risk for suffering a particular condition and the need for preventive intervention. Haas et al. have used several risk stratification techniques to evaluate the performance in predicting healthcare utilization. They have studied 83 patients empanelled in 2009 and 2010 in a primary care practice. 7 different risk stratification techniques were used: Adjusted Clinical Groups (ACGs), Hierarchical

Condition Categories (HCCs), Elder Risk Assessment, Chronic Co morbidity Count, Charlson Co morbidity Index, and Minnesota Health Care Home Tiering and a combination of Minnesota Tiering and ERA. To predict the healthcare utilization and cost, historical data (data from 2009) have been used by a logistic regression model using demographic characteristics and diagnosis such as emergency department visits, hospitalizations, 30 day readmissions. The results show that ACG model outperforms the other risk stratification methods. They have studied data stratification based on the acuity of each patient and generated different results for each stratum. However, in our study the stratifying factor is unknown beforehand and needs to be determined from the existing branching questions by using statistical methods. We have used the contingency tables to divide the population based on branching questions. The Fisher's Exact Test is used in order to compare the significance of the selected branching questions.

Once the data is preprocessed, a predictive index can be constructed by starting with the detection of important attributes (risk factors). Attribute selection is a machine learning method that selects an optimal subset of attributes by eliminating the ones which contain less predictive information. Reducing the dimensionality of an attribute space improves the performance by diminishing the curse of dimensionality effect. Attribute selection also gains advantage from efficiency in terms of storage and computational costs. Also, the execution time spent for both training and testing phases will decrease.

The attribute selection methods are categorized into three different forms: filter, wrapper or embedded (Molina 2002). It is important to evaluate the existing methods and figure out which one performs better in certain situations. Some algorithms perform well on correlated data while others can handle noise or missing values depending on the nature of the data.

In filtering methods, attribute selection is a preprocessing step independent of the induction algorithm. Information gain algorithm determines the importance of each attribute by evaluating the uncertainty reduction, while ReliefF is sampling an instance and evaluating the difference between its nearest neighbors from both the same and the opposite class. Relevance scores are assigned to each attribute. Correlation based feature selection methods eliminate one of the less important attribute that correlates with another attribute (Hall 1998). In embedded methods, inducer has its own attribute selection algorithm embedded such as J48, a widely used decision tree algorithm. The occurrence of an attribute in a tree provides information about the importance of that particular attribute. Information gain and entropy reduction methods can be applied to each candidate attribute of the decision tree node to evaluate the importance of each attribute (Sugumaran 2006). There have been several studies on attribute selection methods. Hall and Holmes (Hall 1998) use the UCI dataset, a real world data set, which contains different data types such as categorical, continuous and multivariate. Six attribute selection methods are compared in terms of classification accuracy, reduction rate and speed. Molina et al., (Molina 2002) use a simulation based randomly created binary and nominal valued dataset. Different types of syntactic functions are applied to the dataset to generate the class labels. Since it is a fully controlled case, varying number of relevant, irrelevant and redundant attributes are placed in the dataset. However, the simulated dataset doesn’t lend itself to real longitudinal dataset of clinical trials and the functions used in this study do not care about any particular attribute when deciding about the resultant, which does not fit to what we usually see in real world biomedical datasets.

Rule extraction is a machine learning method that selects formal rules from a set of observations. Given a set of training examples where the class labels are known, the aim is to find out the classification rules that will help to predict the new instances. Agrawal et al. introduced association rules for discovering the relations between variables in large databases. Rule extraction gains advantage from efficiency in terms of computational costs and after extracting the proper rules from the dataset the time spent for classification of the new instances decrease.

The different rule extraction methods used in this study are Apriori, PART, Prism and Jrip. There have been several studies on rule induction methods. Pires and Branco use a simulation based dataset to compare the results of two multinomial classification rules in terms of their performances. One of those classification rules is the Bayesian approach while the other one is the likelihood measures.

In this research, on the other hand, several attribute selection and rule extraction algorithms are applied to a simulation based dataset with longitudinal trials in order to compare the performances of the algorithms in terms of different noise levels and different incomplete data levels and combination of both. Multicollinearity is added to the dataset to evaluate the most robust algorithms when dependencies between attributes are in question. Same attribute selection and rule extraction techniques are also applied to the combination of the longitudinal datasets. A simulation dataset is chosen since full control over the dataset is achieved.

Besides determining the most robust attribute selection and rule extraction algorithm, this experiment also helps us to determine up to what percentage of incomplete data, noise and multicollinearity can be handled.

Logistic regression is one of the most commonly used predictive modeling techniques. Wang et al. have used multilinear sparse logistic regression in order to predict the risk on clinical data (Wang 2014). We have used logistic regression in order to determine the potential predictive factors and to determine the predictive probability of each factor. The effectiveness of the potential predictive factors is determined by odd ratios and the predictive index performance is determined based on wald scores and confidence intervals.

CHAPTER TWO

METHOD - DATA PRE-PROCESSING

2.1 Analysis of Incomplete and Inconsistent Data

The next step of the proposed method is to further preprocess the incomplete and inconsistent data by converting the entire questionnaire into a directed acyclic graph $G(N, A)$ where, each question is represented by a vertex, $v \in N$, and each answer is represented by an edge $a \in A$. The questionnaires are usually designed so they can be directly converted to an acyclic graph. That is, each question is answered at most once. If this rule is not followed, one should be able to devise an acyclic equivalent of a cyclic graph. Therefore, in this study we assume a questionnaire with a corresponding acyclic graph.

There can be vertices with more than one child (branching questions) in G. The children of such vertices are mutually exclusive in the sense that the respondent should answer only one of them to insure only one path is visited. This corresponds to the case where the questionnaire instructs the respondent to skip some questions. The unvisited vertices based on the questionnaire's instruction (i.e., vertices out of the path) cause skip pattern missing data (*SPMD*). Refusing or neglecting to visit (answer) a vertex on a path, when it is not supposed to, causes genuine missing data (*GMD*). Refusing or neglecting to visit any of the alternative paths going through a branching vertex along with the branching vertex itself causes undetermined missing data (*UMD*). This type of missing data is called undetermined because there is no way to determine whether they are *GMD* or *SPMD*. When more than one child of a branching vertex is visited, it causes inconsistent data. All the answers causing data inconsistency have to be removed (leaving just one valid path) as if they were never visited. Once the answers causing data inconsistency are removed, the incompleteness caused by the removal can be considered as *SPMD*.

Figure 2.1 shows a subgraph generated from questions v_{17} through v_{26} in MESA. The answers (edges) are labeled as a_k. For example, a_{12} in Figure 2.1 represents the first set of answers to v_{22} that leads the respondent to v_{23}, and a_{11} represents the second set of answers to v_{22} that leads the respondent to v_{24}. Unlike this example, there could be more than two possible set of answers to a question as well as only one set. The latter means regardless of the answer to such a question, the respondent will always be led to only one following question. In this case, leaving a question unanswered causes *GMD*. The former means, the respondent will select one answer out of two or more possible set of answers. For v_{22}, one option is to answer 'a_{12}' and continue with v_{23}. The alternative option is to answer 'a_{11}', skip v_{23} and continue with v_{24}. Missing value caused by skipping v_{23} is referred to as *SPMD*. If none of the three vertices (v_{22}, v_{23} and v_{24}) is answered the type of the missing values corresponding to these vertices cannot be determined (*UMD*). If a subject, on the other hand, answers both questions v_{25a} and v_{25b}, those two answers become mutually inconsistent and one of the alternative answers has to be removed from that subject's answer set.

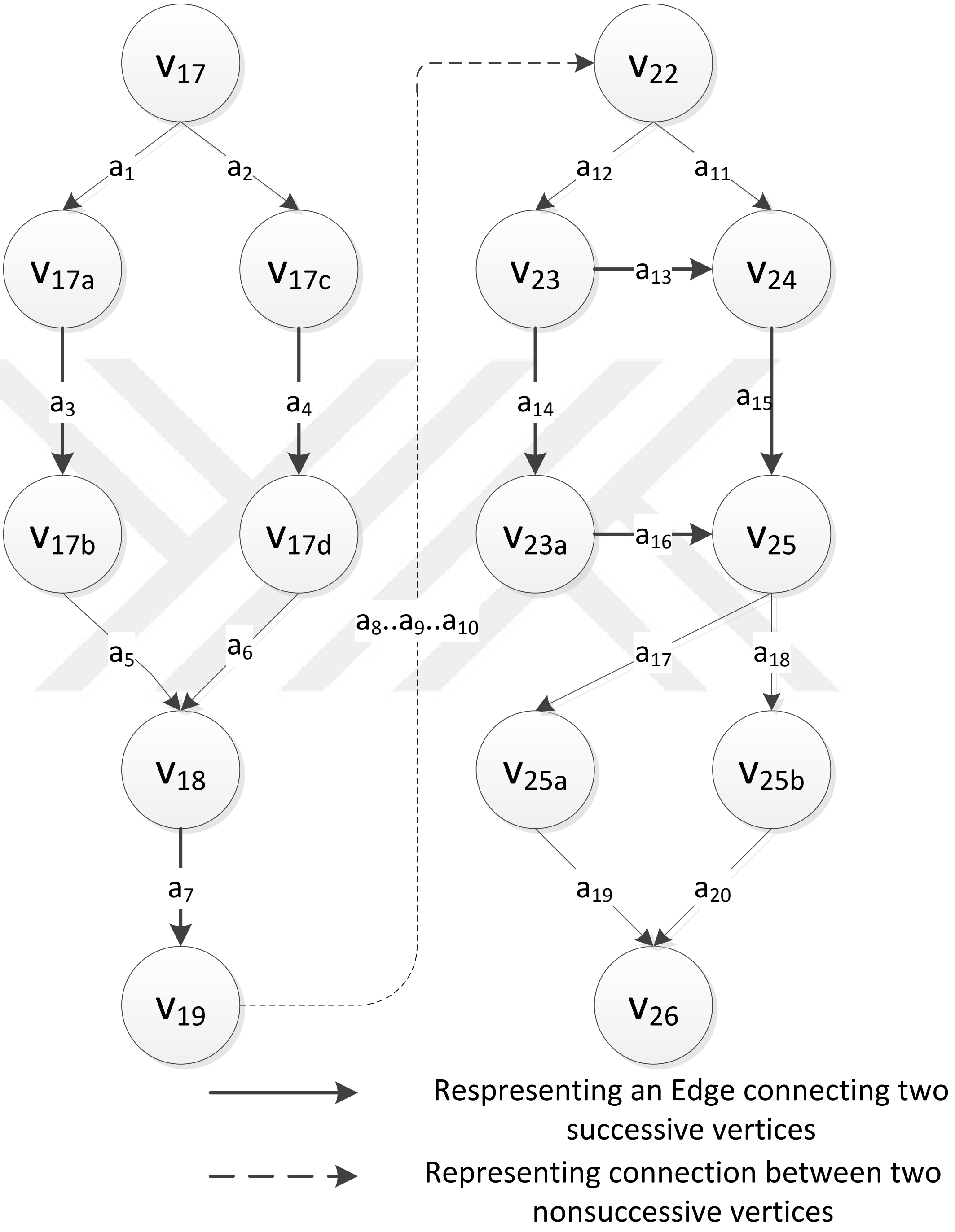

Figure 2.1- Subgraph of MESA with labeled responses

The answer set for each subject generates a different path on the graph. Thus the answer set that are incomplete or inconsistent are subject dependent. Therefore, incomplete and inconsistent data have to be extracted specific to each subject. Manual methods may show less than desired reliability especially when dealing with large amounts of data. This entails an automated method to be devised.

In this part of the study, we focus solely on the female population of the MESA baseline survey (HH1). There are 1154 women in MESA aged 60 and older. First, the inconsistent data are extracted from the dataset in order to minimize the noise. Then, our objective is to distinguish the incomplete data into *GMD*, *SPMD* and *UMD.* Table 2.1. shows the definitions of the basic terms being used throughout this study.

Table 2.1- Definition of basic terms

$\boldsymbol{Q}$	**Set of Consistent Questions Answered**
Q^*	Set of Questions Answered
R^*	Set of Responses of a Subject
$\boldsymbol{R}$	Set of Consistent Responses of a Subject
$\boldsymbol{P}P_i$	An augmented set of answers
$\cup\, PP_i$	Actual Path
$\boldsymbol{I}$	Inconsistent Data Vertices
$\boldsymbol{S}$	*SPMD* Vertices
$\boldsymbol{N}$	All Vertices
$\boldsymbol{M}$	*GMD* Vertices
$\boldsymbol{U}$	*UMD* Vertices
$\boldsymbol{LP}$	Linear Path

The first step of the analysis is to determine the inconsistent answers. After the inconsistent answers are detected and removed from the dataset, the remaining data will be consistent. The *SPMD*, the *GMD* and *UMD* analyses are performed on the consistent dataset.

2.1.1 Inconsistent Data Analysis

In this section, we present the mathematical framework of the method that determines inconsistent answers. We define two functions that return the corresponding questions of a given answer set where one returns the previous question, $P_b : A \rightarrow N$, and the other returns the next question $P_a : A \rightarrow N$. $P_a(\{a_k\})$ returns the question that follows the answer a_k. $P_b(\{a_k\})$ returns the question corresponding to answer a_k. One can observe the following:

$P_a(\{a_k\}) = \{v_j\}$ and $P_a^{-1}(\{v_j\}) = \{a_k\}$ where $a_k = (v_i, v_j) \in A$.

$P_b(\{a_k\}) = \{v_i\}$ and $P_b^{-1}(\{v_i\}) = \{a_k\}$ where $a_k = (v_i, v_j) \in A$.

Example I: Let's consider $R_s^* = \{a_2, a_5, a_4, a_6, a_8, a_9, a_{10}, a_{15}, a_{17}, a_{19}\}$ as the set of responses of s^{th} subject to the sub-questionnaire in Figure 2.1.

$Q_s^* = \{v_{17}, v_{17b}, v_{17c}, v_{17d}, v_{19}, v_{20}, v_{21}, v_{24}, v_{25}, v_{25a}\}$ contains the corresponding vertices of each edge in R_s^* , i.e., $Q_s^* = P_b(R_s^*)$.

Answering a questionnaire by a subject is equivalent to visiting the vertices of a linear path, LP_i, in the questionnaire's corresponding graph. We can formally define a linear path as follows.

Definition I: A linear path, LP_i, is defined as a set of vertices that includes the root and the terminal vertices of the graph, G. Further, for any vertex, v_k in LP_i, there exists one and only one vertex, v_j such that (v_k, v_j) is in A, except for the terminal vertex.

We define LP as the set of all such sets shown below.

$$LP = \bigcup_i LP_i$$

An example of a linear path can be

$LP_1 = \{v_{17}, v_{17c}, v_{17d}, v_{18}, v_{19}, v_{20}, v_{21}, v_{22}, v_{23}, v_{23a}, v_{25}, v_{25a}, v_{26}\}$ as can be constructed from Figure 2.1.

Definition II:

$PP_{s_i} = \{a_j | \, a_j \in A, a_j \in P_b^{-1}(LP_q), \exists \, a_m \in R_s^* \, s.t. \, a_m \in P_b^{-1}(LP_q) \; and \; \nexists \, LP_r \; s.t. \, R_s^* \cap P_b^{-1}(LP_q) \;\; \subset R_s^* \cap P_b^{-1}(LP_r)\}$

The *i*-th potential path for subject s, PP_{s_i}, is defined as a set of edges that connect the vertices of the *j*-th linear path, i.e., $P_b^{-1}(LP_j)$. Also, each PP_{s_i} should include all the answers in R_s (set of consistent answers). Note that there are usually more than just one potential path for each subject.

In order to explain the last condition in Definition II $\nexists \, LP_r \; s.t. \, R_s^* \cap P_b^{-1}(LP_q) \;\; \subset R_s^* \cap P_b^{-1}(LP_r)$, let's consider a subgraph consisting of only v_{22} to v_{25} in Example I with following answer set: $R_s^* = \{a_{13}, a_{15}\}$. One may imagine that there are two potential paths: $PP_{s_1} = \{a_{12}, a_{13}, a_{15}\}$ and $PP_{s_2} = \{a_{11}, a_{15}\}$. However, since $R_s^* \cap P_b^{-1}(LP_2) \subset R_s^* \cap P_b^{-1}(LP_1)$, where $LP_1 = \{v_{22}, v_{23}, v_{24}, v_{25}\}$ and $LP_2 = \{v_{22}, v_{24}, v_{25}\}$ Definition II does not allow PP_{s_2}. This is important because later in this section when the inconsistent answers are formally defined based on potential paths, it does not consider a_{13} and a_{15} inconsistent, which is intuitively the case.

Observation I: A linear path is subject independent, whereas a potential path is subject dependent.

An acyclic questionnaire graph G, can contain areas with multiple sub-linear paths *(MSLP)* and areas with a single sub-linear path *(SSLP)*. *MSLP*, $B(N_B, A_B)$, can be defined as a *connected subgraph* where

$$N_B = \left\{ v_i | v_i \in \left(P_b\left(A - \cap_j P_b^{-1}(LP_j)\right) \cup P_a\left(A - \cap_j P_b^{-1}(LP_j)\right)\right) \right\}.$$

If there is $v_i \in N_B\ s.t. v_i \in \cap_k LP_k\ and\ v_i \neq v_{BI}, v_i \neq v_{BT}$ where v_{BI} is the initial vertex of subgraph B and v_{BT} is the terminal vertex of subgraph B then $B(N_B, A_B)$ should be decomposed into sub *MSLP*'s. An example of such situation is the following:

Example II: Figure 2.1 shows a connected subgraph B'

where $N_B' = \{v_{23}, v_{23a}, v_{24}, v_{25}, v_{25a}, v_{25b}, v_{26}\}$. The initial vertex of N_B' is v_{23} ($v_{BI'} = \{v_{23}\}$) and the terminal vertex of N_B' is v_{26} ($v_{BT'} = \{v_{26}\}$). Since

$\cap_k LP_k = \{v_{23}, v_{25,} v_{26}\}$ and $v_{25} \in N_B'$ and $v_{25} \neq v_{BI'}$ and $v_{25} \neq v_{BT'}$, B' should be decomposed.

To decompose such *MSLP*'s, we define N_{B_j}: $N_{B\,j} = \left(\cup_i LP_i \Big|_{v_j}^{v_{j-1}} \right)$ where $v_j\ \in$

$\left(\cap_i LP_i \Big|_{v_{BT}}^{v_{BI}} \right)$ and v_0 is v_{BI}. Note that $\left(LP_i \Big|_{v_l}^{v_k} \right)$ refers to all linear paths starting from v_k and ending at v_l in G. Hence, the decomposition of *B*' in Example II returns two connected subgraphs B_1 and B_2 where $N_{B_1} = \left(\bigcup_i LP_i \,\Big|_{v_{25}}^{v_{23}}\right) = \{v_{23}, v_{23a}, v_{24}, v_{25}\}$ and $N_{B_2} = \left(\bigcup_i LP_i \,\Big|_{v_{26}}^{v_{25}}\right) = \{v_{25}, v_{25a}, v_{25b}, v_{26}\}$.

The edge set , A_B, of a *MSLP*, $B(N_B, A_B)$, can easily be defined as, $A_{B_i} = P_b^{-1}(N_{B_i}) \cap P_a^{-1}(N_{B_i})$. *SSLP*, $L(N_L, A_L)$, is a connected subgraph where $N_L \subseteq \; P_b\left(\cap\, P_b^{-1}(LP_i)\right) \cup P_a\left(\cap\, P_b^{-1}(LP_i)\right)$ and $A_L \subseteq \cap\, P_b^{-1}(LP_i)$.

The inconsistent data occurs when more than one alternative path of a branching question is visited. For example, subject s has answered both a_5 and a_6 (See R_s^*). However, there is no linear path (LP_i) that contains both of these answers (a_5 and a_6) as can be seen in Figure 2.1. We refer to those as mutually inconsistent responses. Similarly, according to subject s's responses, the following set of answer pairs are mutually inconsistent: (a_2, a_5), (a_4, a_5) and (a_6, a_5). However, those responses ($\{a_2, a_5, a_4, a_6\}$) are not mutually inconsistent with any other edge in the response set ($\{a_8, a_9, a_{10}, a_{15}, a_{17}, a_{19}\}$), since a linear path (LP_i) can be generated from any of those edges to the remaining edges in the response set. At this point, we formally define mutually inconsistent answers.

Definition III: a_i is inconsistent with a_j iff $a_i \neq a_j$ and $a_i, a_j \in A_{B_n}$ and a_i or $a_j \notin \left(PP_{s_q} \cap PP_{s_r}\right)$ where $a_i \in PP_{s_q}$, $a_j \in PP_{s_r}$ and $PP_{s_q} \neq PP_{s_r}$.

Two different responses ($ai \neq aj$) of a subject are mutually inconsistent with each other when they are on *exclusive paths* of the same multiple sub linear paths (MSLP).

Let's consider two scenarios: $R_s^* = \{a_1, a_3, a_{19}, a_{20}\}$.

Clearly, a_1 and a_3 are consistent with each other. However, one can construct two potential paths as follows: $PP_{s_1} = \{a_1, a_3, a_5, a_7, a_8, a_9, a_{10}, \dots, a_{17}, a_{19}\}$ and $PP_{s_2} = \{a_1, a_3, a_5, a_7, a_8, a_9, a_{10}, \dots, a_{18}, a_{20}\}$. In this example, $a_1 \neq a_3$, a_1 and $a_3 \in A_{B_n}$, $a_1 \in PP_{s_1}$, $a_3 \in PP_{s_2}$, $PP_{s_1} \neq PP_{s_2}$. Therefore, without condition a_i or $a_j \notin \left(PP_{s_i} \cap PP_{s_j}\right)$, a_1 would be inconsistent with a_3, which is not intuitively true. The second scenario: $R_s^* = \{a_{11}, a_{13}, a_{15}\}$. Obviously, a_{13} and a_{15} are consistent answers. However, one can construct two potential paths as follows: $PP_{s_1} = \{\dots, a_{10}, a_{11}, a_{15}, a_{18}, a_{20}\}$ and $PP_{s_2} = \{\dots, a_{10}, a_{12}, a_{13}, a_{15}, a_{18}, a_{20}\}$. In this example, $a_{13} \neq a_{15}$, a_{13} and $a_{15} \in A_{B_n}$, $a_{15} \in PP_{s_1}$, $a_{13} \in PP_{s_2}$, $PP_{s_1} \neq PP_{s_2}$. Therefore, without condition a_i or $a_j \notin \left(PP_{s_i} \cap PP_{s_j}\right)$, a_{13} would be inconsistent with a_{15}, which is not intuitively true. Condition a_i or $a_j \notin \left(PP_{s_i} \cap PP_{s_j}\right)$ resolves both of these counterintuitive situations explained in the above two scenarios.

Once the mutually inconsistent edges are detected, a failed edit graph (FEG) and its complement is generated to find a minimal deletion set (MDS). FEG, $F(N, A)$, is generated by creating a vertex for each inconsistent answer and then joining each pair of mutually inconsistent vertices with an edge. Figure 2.2 shows a FEG for the given example above.

Definition IV: A failed edit graph (FEG) is an undirected graph $F(N_F, A_F)$ s.t. $N_F \subseteq (A - \bigcap_l P_b^{-1}(LP_l))$ where for any $a_i, a_j \in A_{B_n}$ and $a_i, a_j \in N_F, (a_i, a_j) \in$

A_F iff $\exists PP_{s_k}\ s.t\ a_i \in PP_{s_k}\ and\ a_j \notin PP_{s_k}$ and a_i or $a_j \notin \left(PP_{s_i} \cap PP_{s_j}\right)$ where $a_i \in PP_{s_i}$, $a_j \in PP_{s_j}$ and $PP_{s_i} \neq PP_{s_j}$ for subject s.

Observation II: A different FEG is generated for each *MSLP* with inconsistent responses.

MDS is the minimum number of vertices that need to be removed from the answer set in order to generate a consistent answer set. If a_2, a_4 or a_6 is removed from the dataset, FEG will still be connected. However, if a_5 is removed, the FEG becomes disconnected, which means there are no inconsistencies. Hence, a_5 should be removed from the response set.

The new response set will be referred to by R_s where the absence of the asterisk denotes the fact that this response set is consistent:

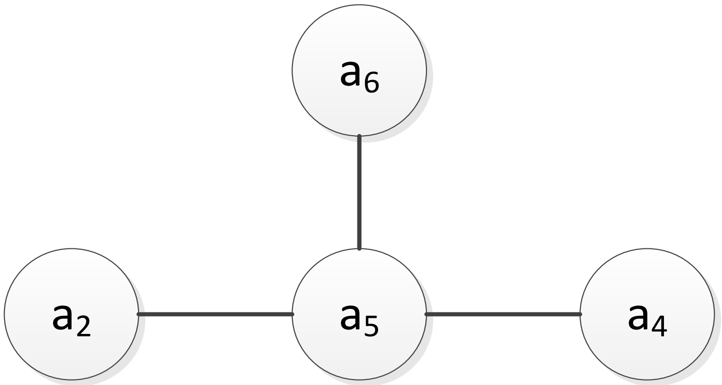

Figure 2.2- Failed Edit Graph

($R_s = \{a_2, a_4, a_6, a_8, a_9, a_{10}, a_{15}, a_{17}, a_{19}\}$). The corresponding vertices for R_s will be: $Q_s = \{v_{17}, v_{17c}, v_{17d}, v_{19}, v_{20}, v_{21}, v_{24}, v_{25}, v_{25a}\}$. Since a_5 is removed from $R^*_{s,}\, v_{17b}$ which is the corresponding vertex (question) of a_5 also needs to be removed from Q^*_s. However, the FEG is not always as simple as it is in the given example, which means finding MDS could become fairly complex. Therefore, a method is generated by Fagan and Greenberg to find MDS (Fagan et al. 1988).

The method for detecting the MDS is the following: Given a FEG, F, the vertex set in F, i.e., N and a largest maximal clique C^M in the complement of the FEG, F^c, the minimal deletion set is $N - C^M$. Note that a clique is an undirected graph such that every two vertices are connected by an edge.

The complement of F is $F^c(N_F, A^c_F)$ which consists of a set of connected graphs (cliques) each of which are denoted by $C(N_c, A_c)$ where $N_c \subseteq N_F$, $A_c \subseteq A^c_F$ and $F^C = \bigcup_i C_i$. The vertices on each clique are answers in a linear path, whereas, the vertices on two different cliques are not on the same linear path. We define $C^M(N^M_c, A^M_c)$ as the clique with a vertices set, A^M_c, of maximum cardinality (largest maximal clique). Now consider the same example given in Figure 2.2. The complement of the graph is shown in Figure 2.3, where the largest maximal clique is C^M_s. The set $N - C^M_s$returns $\{\{a_5\}\}$, which is the minimal deletion set.

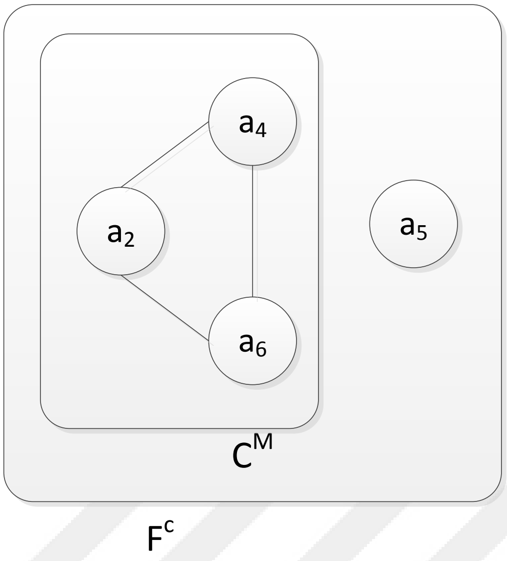

Figure 2.3-Minimal Deletion Set

Lemma I: C is complete.

Proof: Let's assume C is not complete. It means $\exists\ a_i, a_j \in N_{c_s}$, for the s-th subject, such that $(a_i, a_j) \notin A_{c_s}$, which means $(a_i, a_j) \in A_{F_s}$ for subject s (Since the cliques are generated based on the complement of the FEG). However, from the definition of C we know that $a_i, a_j \in P_b^{-1}(LP_k)$ and therefore $(a_i, a_j) \notin A_{F_s}$, which is a contradiction and our assumption that C is not complete is not true.

Lemma II: Inconsistencies only happen in a *MSLP*.

Proof: One has to show two things to prove this. 1) It is possible to have inconsistencies in *MSLP*, and 2) it is impossible to have inconsistencies in *SSLP*. We prove (1) by constructing an answer set, R_s^*, where $\exists a_i, a_j \in R_s^*$ s.t. $a_i \in P_b^{-1}(LP_k)$ and $a_j \in P_b^{-1}(LP_l)$ and $LP_k \neq LP_l$ and $a_i, a_j \in A_{B_n}$ and $R_s^* \bigcap_l P_b^{-1}(LP_l) \not\subset R_s^* \bigcap_k P_b^{-1}(LP_k)$. Based on

Definition II $a_i \in PP_{s_i}$ and $a_j \in PP_{s_j}$ where $PP_{s_i} \neq PP_{s_j}$. Definition III constitutes inconsistency between a_i and a_j. Proof of (2) can be directly derived from Definition III, where inconsistencies are only defined for responses in *MSLP* ($a_i, a_j \in A_{B_n}$), which means responses in *SSLP* ($a_i, a_j \in A_L$) cannot be inconsistent with each other. This proves Lemma II and further the latter leads to Observation II.

Note that, lack of inconsistency in only part of an MSLP does not guarantee consistency. Therefore in order to identify inconsistency an MSLP is supposed to be investigated in its entirety.

Observation III: There can be no inconsistencies in *SSLP*.

Lemma III: There can be no inconsistencies between two *MSLP'*s.

Proof: This can be directly derived from Definition III, where inconsistencies only occur within a *MSLP* ($a_i, a_j \in A_{B_k}$). Since $a_i \in A_{B_m}$ and $a_j \in A_{B_n}$ and $A_{B_m} \neq A_{B_n}$ in two different *MSLP's* a_i and a_j cannot be inconsistent.

Theorem I: The set of all inconsistent answers for subject, s, I_s can be computed by the following equation. ($N_{F_{s_n}}$ denotes the vertices of the $n-th$ failed edit graph for subject s.)

$$I_s = \bigcup_{n} (N_{F_{s_n}} - N_{c_{s_n}}^{M})$$

Proof: To show this, we need to prove two statements: 1) $\forall a_i \in I, \exists a_j \in R_s^* - I$ s.t. a_i and a_j are inconsistent, and 2) there is no inconsistency within $R_s^* - I$. We show the first statement (1) using contradiction. Assume the subject only has one failed edit graph ($n = 1$). Suppose $a_i \in N_{c_s}$ where $N_{c_s} \subseteq N_{F_s}$. Since $\exists N_{c_s}^M \subset N_{F_s}$ and $N_{c_s}^M \cap N_{c_s} = \emptyset$, $\nexists a_j \in$

$N_{c_s}^M\ s.t.\left(a_i, a_j\right) \in\ A_{c_s}^M$, which means a_i is not connected to any vertex in $N_{c_s}^M$. Therefore, a_i is inconsistent with all answers (vertices) in C_s^M, which means that $\exists a_j \in\ N_{F_s} \subseteq R_s^* - I$ s.t. a_i and a_j are inconsistent. The second statement (2) can be shown as follows: We know that $R_s^* - I =\ N_{c_s}^M\ \cup \left(R_s^*\ \cap (\bigcap_k P_b^{-1}(LP_k))\right)$. If $a_i, a_j\ \in\ N_{c_s}^M \Longrightarrow \left(a_i, a_j\right)\ \in\ A_{c_s}^M$, which means C_s^M is complete as proved in Lemma 1. Therefore, there is no inconsistency within $N_{c_s}^M$. Also, from Lemma II; we know that answers on the same linear path cannot be inconsistent. Therefore, there is no inconsistency within $\left(R_s^*\ \cap (\bigcap_k P_b^{-1}(LP_k))\right)$. Finally, we have to show that answers in $\left(R_s^*\ \cap (\bigcap_k P_b^{-1}(LP_k))\right)$ and $N_{c_s}^M$ are mutually consistent. Assume, $\exists a_i, a_j, a_i \in N_{c_s}^M, a_j \in \left(R_s^*\ \cap (\bigcap_k P_b^{-1}(LP_k))\right)$ s.t. a_i is inconsistent with a_j. This implies that $\nexists\ PP_{s_i}$ s.t. $a_i, a_j\ \in PP_{s_i}$. However, one can construct a PP_{s_i} that includes all $N_{c_s}^M$ and $\left(R_s^*\ \cap (\bigcap_k P_b^{-1}(LP_k))\right)$, which contradicts with the implication of the assumption. Therefore, the answers in $\left(R_s^*\ \cap (\bigcap_k P_b^{-1}(LP_k))\right)$ and $N_{c_s}^M$ are mutually consistent.

Theorem II: $R_s^* - I$ provides one and only one possible path iff $\exists a_i \in R_s^*$ s.t. $a_i \in A_{B_n}$ the edge set of a *MSLP*.

Proof: First, we should prove that there is a potential path using $R_s^* - I$ (1). Then, we have to show the uniqueness of this path (2).

(1) Since $\exists a_i \in R_s^*$, according to Definition II, $\exists PP_{s_i}$.

(2) a_i and a_j cannot be in A_L, because of Lemma II. Let us consider a_i, a_j in $N_{c_s}^M$. If $a_i, a_j \in N_{c_s}^M$ they cannot represent two different PP_{s_k}'s, that is in contradiction

with $PP_{s_i} \neq PP_{s_j}$. If $a_i \in N_{c_s}^M$ and $a_j \in N_{C_s} \neq N_{c_s}^M$ then $a_j \notin R_s^* - I$ which means PP_{s_j} cannot exist.

2.1.2. Incomplete Data Analysis

In this section, we introduce a method to detect incomplete data (*SPMD, GMD* and *UMD*). The first step of the analysis is to determine the potential paths for each response set. Example I shows that the potential paths for

$R_s = \{a_2, a_4, a_6, a_8, a_9, a_{10}, a_{15}, a_{17}, a_{19}\}$ are the following:

$PP_{s_1} = \{a_2, a_4, a_6, a_7, a_8, a_9, a_{10}, a_{11}, a_{15}, a_{17}, a_{19}\}$

$PP_{s_2} = \{a_2, a_4, a_6, a_7, a_8, a_9, a_{10}, a_{12}, a_{13}, a_{15}, a_{17}, a_{19}\}$

The determined potential paths can then be used in the analysis of *SPMD, GMD* and *UMD*.

2.1.2.1. SPMD Analysis. The first step of the analysis is the detection of *SPMD*. *SPMD* can only occur within *MSLP*'s. When an edge is not an element of any potential path, it will be a *SPMD* referred to by S. *SPMD* can be formally defined as;

Definition V: $a_k \in A$ is in *SPMD* if $\nexists PP_{s_i}$ s.t. $a_k \in PP_{s_i}$

Therefore, *SPMD* can be estimated by the following equation:

$$S_s = A - \bigcup_i PP_{s_i}$$

Hence; Example I shows that $S_s = \{a_1, a_3, a_5, a_{14}, a_{16}, a_{18}, a_{20}\}$.

2.1.2.2.GMD Analysis. The next step of incomplete data analysis is to determine the *GMD* referred to by M. If an edge is an element of a potential path, but not a member of the consistent answer set , R_s, it can be either a *UMD* or a *GMD*. If this condition is met ($a_m \in PP_{s_i}$ for any i, and $a_m \notin R_s$) within an *SSLP* ($a_m \in A_L$), we can refer to this edge as a *GMD*. Otherwise, if $a_m \in A_B$, we can refer to a_m as a *GMD* only if there exists at least one edge , a_k, in the same PP_i within the *MSLP* that ensures $a_k \in R_s$.

GMD (M) can be formally defined as;

Definition VI: $M = \left\{a_m \mid \forall i, a_m \in \bigcap_i PP_{s_i} \text{ and } a_m \notin R_s \right\}$

Hence, *GMD* (M) can be calculated by the following formula:

$$M_s = \bigcap_i PP_{s_i} - R_s$$

Example I shows that $M_s = \{a_7\}$ for the s^{th} subject.

Observation IV: $M_s = \bigcap_i PP_{s_i} - R_s \neq P_b^{-1}\left(\bigcap_i \left(P_b\left(PP_{s_i}\right)\right) - Q_s\right)$

This observation can be shown by a counter example shown below:

The potential paths for the response set R_s are;

$PP_{s_1} = \{a_2, a_4, a_6, a_7, a_8, a_9, a_{10}, a_{11}, a_{15}, a_{17}, a_{19}\}$

$PP_{s_2} = \{a_2, a_4, a_6, a_7, a_8, a_9, a_{10}, a_{12}, a_{13}, a_{15}, a_{17}, a_{19}\}$. The intersection of all potential paths is $\bigcap_i PP_{s_i} = \{a_2, a_4, a_6, a_7, a_8, a_9, a_{10}, a_{15}, a_{17}, a_{19}\}$. Therefore, the left hand side of the equation becomes $\bigcap_i PP_{s_i} - R_s = \{a_7\}$. The right hand side of the equation first converts the edges within the potential paths to their corresponding vertices:

$P_b(PP_{s_1}) \; = v_{17}, v_{17c}, v_{17d}, v_{18}, v_{19}, v_{20}, v_{21}, v_{22}, v_{24}, v_{25}, v_{25a}\}$ and $P_b\left(PP_{s_2}\right) = \{v_{17}, v_{17c}, v_{17d}, v_{18}, v_{19}, v_{20}, v_{21}, v_{22}, v_{24}, v_{25}, v_{25a}\}$. The intersection of those sets becomes $\bigcap_i P_b\left(PP_{s_i}\right)$=$\{v_{17}, v_{17c}, v_{17d}, v_{18}, v_{19}, v_{20}, v_{21}, v_{22}, v_{24}, v_{25}, v_{25a}\}$. Therefore, the set of genuine missing questions of the s^{th} subject is $\bigcap_i P_b\left(PP_{s_i}\right) - Q_s$= $\{v_{18}, v_{22}\}$. The corresponding edges of the genuine missing questions are $P_b^{-1}\left(\bigcap_i \left(P_b\left(PP_{s_i}\right)\right) \; - \; Q_s\right)$=$\{a_7, a_{12}, a_{11}\}$. The left hand side of the equation is not equal to the right hand side ($\{a_7\} \neq \{a_7, a_{12}, a_{11}\}$). Therefore, $\bigcap_i PP_{s_i} - R_s \neq P_b^{-1}\left(\bigcap_i \left(P_b\left(PP_{s_i}\right)\right) - Q_s\right)$.

2.1.2.3.UMD Analysis. The last step of our incomplete data analysis is to determine the *UMD* (U). *UMD* occurs when none of the questions are responded within a *MSLP* ($A_B \neq \emptyset$). In this case, due to the lack of responses in A_B, it is impossible to entitle an edge as *GMD* or *SPMD*. For such edges, the status is not determined. *UMD* (U) can be estimated by the following formula:

$$U_s = \bigcup_i PP_{s_i} - \bigcap_i PP_{s_i}$$

Clearly, *UMD* can also be estimated by; $U_s = A - R_s - M_s - S_s$

Hence $U_s = \{a_{11}, a_{12}, a_{13}\}$.

Observation V: *SPMD* and *UMD* edges can occur within a *MSLP*; however *GMD* edges can occur both within both *MSLP* and *SSLP*s.

The same method (detection of *SPMD*, *UMD*, *GMD* and inconsistent data analysis) is applied over the entire MESA data starting from the first attribute as the initial vertex and the 825th attribute as the terminal vertex.

2.2. Implementation

In this section the implementation of inconsistent and incomplete data are discussed.

2.2.1. Implementation of Inconsistent Data Analysis

In this section, we present a technique for detecting inconsistent data. According to Lemma I, all the subgraphs in the complement of the failed edit graph are complete and therefore they are already maximal cliques. Therefore, the number of vertices of each maximal clique should be counted and the one with the maximum number of vertices should be determined (i.e. largest maximal clique), which has a polynomial time complexity. Once the largest maximal clique is determined, the vertices that are not components of the largest maximal clique are removed from R_s^* $(R_s^* - N_{C_s}^M)$, in order to prevent inconsistencies.

It is expected that the number of questions being asked in a questionnaire to be very limited, since the questionnaires are designed for human. Even in the case when the input data is large and complex, the largest maximal clique detection algorithm is only applied on areas that contain multiple sub-linear paths (MSLP) with mutually inconsistent edges (not on the entire graph, G, since inconsistencies can only occur on MSLP portions). Therefore, extending this work to other datasets would not cause poor run times or inefficient memory usage.

2.2.2. Implementation of Incomplete Data Analysis

In this section, we present an implementation for detecting the potential paths for each subject. Once the potential paths are detected, the *GMD*, *SPMD* and *UMD* vertices can be calculated by the formulas defined in the previous section. The first step is to create two different matrices that represent the MESA graph (Node Matrix (M_1), and Edge Immediate Successor Matrix (M_2)). Note that, these matrices are subject-independent.

The Node Matrix shown in Table 2.2 , (M_1), corresponds to a portion of the graph in Figure 2.1 (from v_{17} to v_{20}). M_1 is generated by the following function.

$$M_1(i,j) = \begin{cases} a_m & if\ \exists\ a_m \in A\ s.t.\ a_m = (v_i, v_j) \\ 1 & if\ i = j \\ 0 & otherwise \end{cases}$$

The Edge Immediate Successor Matrix (M_2) shown in Table 2.3 is generated by the following function.

$$M_2(i,j) = \begin{cases} 1 & if\ the\ terminal\ node\ of\ a_i\ is\ the\ initial\ node\ of\, a_j \\ 1 & i = j \\ 0 & otherwise \end{cases}$$

Once the matrices are generated, Algorithm 1 is designed to find the linear and potential paths. After the linear and potential paths are determined, the formulas explained in the Method Section can be used to determine the *GMD, UMD* and *SPMD*s.

Algorithm 1-Implementation for Determining Linear and Potential Paths

/*The algorithm begins with an "initial" vertex, v_{init}. It then iteratively transitions from the current vertex to an adjacent, $P_a(P_b^{-1}(v_{init}))$, until it can no longer find an

```
unexplored vertex to transition to from its current location. The algorithm is executed $degree(v_{init})$ times.*/
Define global $k = 1,\ t = 0$                //$k$ and $t$ are arbitrary indices.
$v_{current} = v_{init}$                //$v_{init}$ is the initial vertex and $LP$ is a linear path
Define FindLinearPath($v_{current}, LP$)
{
    For all $P_b^{-1}(v_{current})$ incident to $v_{current}$
        If $P_b^{-1}(v_{current})$ is unexplored then
            $LP_k = LP_k \cup P_a\left(P_b^{-1}(v_{current})\right)$
            FindLinearPath($P_a(P_b^{-1}(v_{current})), LP_k$)
            Increment $k$
        End If
    End for
}
/*Below for each linear path, a condition is checked to see whether that linear path's corresponding edges, the consistent response set $R_s$ is a subset of $P_b^{-1}(LP_i)$. Each $P_b^{-1}(LP_i)$ satisfying this condition is assigned to a potential path. */
Define FindPotentialPath($R_s, LP$)
{
    For $i = 1: k - 1$
        If $R_s$ is a subset of $P_b^{-1}(LP_i)$ then
            $PP_{s_t} = P_b^{-1}(LP_i)$
```

```
                Increment $t$
            End If
        End For
}
```

We analyze this algorithm for its worst case time complexity starting with FindLinearPath function. The worst case happens when each vertex is connected to as many other vertices as possible. This entails $n-1$ vertices (and the same number of edges) for the first vertex and $n-2$ vertices for the second vertex and so on. The total number of recursive calls to this function would therefore be $W(n) = (n-1) + (n-2) + \cdots + 1 = n(n-1)/2 \in \Theta(n^2)$. The second function, FindPotentialPath, requires each element in R_s to be compared with elements in $P_b^{-1}(LP_i)$. This should be performed for each linear path. An upper bound for the worst case complexity for the latter function would be for a case when $Card(R_s)$, $Card(P_b^{-1}(LP_i))$ and $\#(LP_i)$ are at their maximum value, which is $Card(A)$. The notations $Card(A)$ and $\#(LP_i)$ are the cardinality of the set A (set of edges in G) and the number of linear paths, respectively. As mentioned this is an upper bound for the worst case complexity as $\#(LP_i)$ cannot be more than half of $Card(A)$ since there should always be one initial and one terminal vertices in a questionnaire's graph. Moreover, when $Card(P_b^{-1}(LP_i))$ is at its maximum value, $\#(LP_i)$ is one. Note that the largest value of $Card(A)$ in terms of n, $Card(N)$, is in the order of n^2. Therefore, the upper bound would be $O(n^6)$. We can conclude that the upper bound of the worst case time complexity of the entire algorithm is (n^6), which is a polynomial complexity.

Table 2.2-Node matrix (M_1)

	v_{17}	v_{17a}	v_{17b}	v_{17c}	v_{17d}	v_{18}	v_{19}	v_{20}	$\dots$
v_{17}	1	a_1	0	a_2	0	0	0	0	
v_{17a}	0	1	a_3	0	0	0	0	0	
v_{17b}	0	0	1	0	0	a_5	0	0	
v_{17c}	0	0	0	1	a_4	0	0	0	
v_{17d}	0	0	0	0	1	a_6	0	0	
v_{18}	0	0	0	0	0	1	a_7	0	
v_{19}	0	0	0	0	0	0	1	a_8	
v_{20}	0	0	0	0	0	0	0	1	
$\dots$									

Table 2.3-Edge Immediate successor matrix (M_2)

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	$\dots$
a_1	1	0	1	0	0	0	0	0	0	0	0	0	
a_2	0	1	0	1	0	0	0	0	0	0	0	0	
a_3	0	0	1	0	1	0	0	0	0	0	0	0	
a_4	0	0	0	1	0	1	0	0	0	0	0	0	
a_5	0	0	0	0	1	0	1	0	0	0	0	0	
a_6	0	0	0	0	0	1	1	0	0	0	0	0	
a_7	0	0	0	0	0	0	1	1	0	0	0	0	
a_8	0	0	0	0	0	0	0	1	1	0	0	0	
a_9	0	0	0	0	0	0	0	0	1	1	0	0	
a_{10}	0	0	0	0	0	0	0	0	0	1	1	1	
a_{11}	0	0	0	0	0	0	0	0	0	0	1	0	
a_{12}	0	0	0	0	0	0	0	0	0	0	0	1	
$\dots$													

2.3. Stratification

Fisher’s Exact Test is a statistical significance test used in the analysis of contingency tables. It is used for all sample sizes. The significance of the deviation from a null hypothesis can be calculated exactly. Therefore, it is not necessary to rely on an approximation that becomes exact in the limit as the sample size grows to infinity. The Fisher’s Exact Test is used to determine if there are nonrandom associations between the two variables.

As mentioned before, the dataset is divided into sub populations ($Branch_1, Branch_2$) based on each branching question. Even though, this method can be applied on each question, we limit the number of tests we are using by the branching questions for the following reason: when dividing the population into subgroups, the split is induced by a threshold which is determined by the expert knowledge for each branching question. However, for each non-branching question determining a threshold for each type of answer (categorical, binary, numeric) may lead to incorrect classifications.

Table 2.4 shows a contingency table where the p values are calculated by the following formula:

$$P \ = \frac{\left(\binom{a+b}{a}\binom{c+d}{c}\right)}{\binom{n}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

Here, $R_1, R_2, and\ R_3$ are associations rules extracted from datasets $Branch_1$ or $Branch_2$. The values a, b, c and d are the number of subjects that support/contradict the extracted rules in those two datasets. A different contingency table needs to be generated for each branching question.

Table 2.4 – Contingency Table for Stratification

		$Branch_1$	$Branch_2$	**Row Total**
$\boldsymbol{R_1}$	# of subjects that support R_1	a	b	a+b
	# of subjects that contradict R_1	c	d	c+d
	Column Total	a+c	b+d	a+b+c+d
$\boldsymbol{R_2}$	# of subjects that support R_2			
	# of subjects that contradict R_2			
…	…			
	…			
$\boldsymbol{R_n}$	# of subjects that support R_n			
	# of subjects that contradict R_n			

2.3.1. Simulation

The simulation is created by generating two independent binary datasets. The first dataset *D₁* and the second dataset *D₂* both contain 1000 subjects where $S_1 = \{S_{11}, S_{21}, S_{31}, \dots, S_{1000\ 1}\}$ is the subject set of D_1 and $S_2 = \{S_{12}, S_{22}, S_{32}, \dots, S_{1000\ 2}\}$ is the subject set of D_2 and 15 common attributes ($A = \{A_1, A_2, A_3, \dots, A_{15}\}$). Three different rules are embedded to each dataset. The rules are generated in the sense that none of the rules contradict with another rule. There is no attribute being used in more than one rule. The class labels contain both the classes from the baseline and the first follow up. '$C - I$' is an example of a response indicating that the subject was continent in the baseline and became incontinent in the first follow-up. The rules that are embedded to *D₁* are as follows:

$R_{11} = A1 = 0 \,\&\, A3 = 1 \Rightarrow I - C$

$R_{21} = A5 = 0 \,\&\, A7 = 1 \Rightarrow C - I$

$R_{31} = A10 = 1 \,\&\, A12 = 1 \Rightarrow C - I$

The rules that are embedded to D_2 are as follows:

$R_{12} = A2 = 0 \,\&\, A11 = 0 \Rightarrow C - I$

$R_{22} = A4 = 1 \,\&\, A9 = 0 \Rightarrow C - I$

$R_{32} = A6 = 0 \,\&\, A8 = 1 \Rightarrow I - C$

The two datasets are then combined. The combination $(D_1 + D_2)$, have 2000 subjects $S_{D_1+D_2} = \{S_{11}, S_{21}, S_{31}, \dots, S_{1000\ 1}, S_{12}, S_{22}, S_{32}, \dots, S_{1000\ 2}\}$ and 15 attributes $(A = \{A_1, A_2, A_3, \dots, A_{15}\})$. Three attributes are then added to the combined dataset. Those attributes each represent a branching question having binary values. First branching question BQ_1, takes value '0' for each subject existing in D_1 and value '1' for each subject existing in D_2. The second and third branching questions, BQ_2 and BQ_3 take random binary values.

The combined dataset is then separated into two subsets based on the values of each branching question. The subset for BQ_1 taking value '0' is $D_{BQ_1^0}$ *and* the subset for BQ_1 taking value '1' is $D_{BQ_1^1}$, similarly the subsets for BQ_2 are $D_{BQ_2^0}$ and $D_{BQ_2^1}$ and the subsets for BQ_3 are $D_{BQ_3^0}$ and $D_{BQ_3^1}$. Figure 2.4 shows an example of the combined dataset with branching questions that are separated.

The association rule mining algorithm, Apriori, is used to extract the rules of each subset, since it was proofed to outperform other rule extraction techniques. (Arslanturk et. al.). This is explained in detail in Section 3.3. Apriori is an association rule that iteratively reduces the minimum support until it finds the required number of rules with the given minimum confidence (W. Cohen).

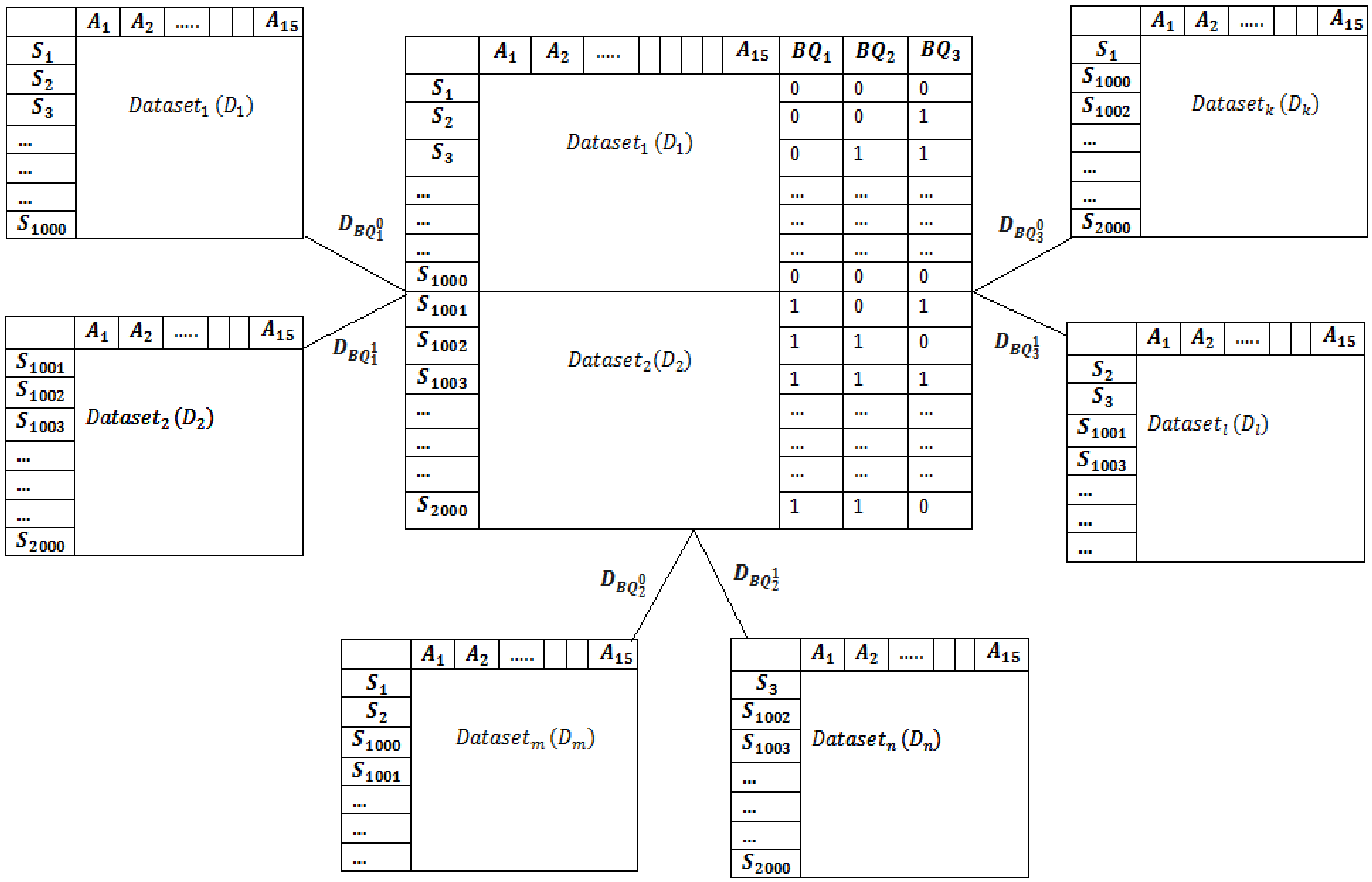
A_1 A_2 ..... A_{15}
S_1 S_2 S_3 ... ... S_{1000}
$Dataset_1\,(D_1)$
A_1 A_2 ..... A_{15}
S_{1001} S_{1002} S_{1003} ... ... S_{2000}
$Dataset_2\,(D_2)$
$D_{BQ_1^0}$
$D_{BQ_1^1}$
A_1 A_2 ..... A_{15} BQ_1 BQ_2 BQ_3
S_1 0 0 0
S_2 0 0 1
S_3 0 1 1
... ...
... ...
... ...
S_{1000} 0 0 0
S_{1001} 1 0 1
S_{1002} 1 1 0
S_{1003} 1 1 1
... ...
... ...
... ...
S_{2000} 1 1 0
$Dataset_1\,(D_1)$
$Dataset_2\,(D_2)$
$D_{BQ_3^0}$
$D_{BQ_3^1}$
A_1 A_2 ..... A_{15}
S_1 S_{1000} S_{1002} ... ... S_{2000}
$Dataset_k\,(D_k)$
A_1 A_2 ..... A_{15}
S_2 S_3 S_{1001} S_{1003} ... ...
$Dataset_l\,(D_l)$
$D_{BQ_2^0}$
$D_{BQ_2^1}$
A_1 A_2 ..... A_{15}
S_1 S_2 S_{1000} S_{1001} ... ...
$Dataset_m\,(D_m)$
A_1 A_2 ..... A_{15}
S_3 S_{1002} S_{1003} ... ... S_{2000}
$Dataset_n\,(D_n)$

Figure 2.4- Simulation of Stratification

Once the rules are extracted for each subset, a contingency table is created for each branching question. A contingency table is a matrix format that displays the frequency distribution of the variables. The rows of the contingency table denote the rules associated with that branching question. The columns are the subsets. For example, for BQ_2, the rows of the contingency table are the rules extracted from subset $\boldsymbol{D_{BQ_2^0}}$ and $\boldsymbol{D_{BQ_2^1}}$, respectively. The columns are the datasets $\boldsymbol{D_{BQ_2^0}}$ and $\boldsymbol{D_{BQ_2^1}}$**.** The contingency table is created to display the relative frequencies, i.e. the support/no support of each rule for each subset.

Once the contingency table is created, the p-values are calculated using the Fisher's Exact Test explained in the Methods Section. Based on the p-values that are calculated, we can determine the branching questions that are statistically significant.

A branching question is a good stratifying factor when it is statistically significant, since different rules (hence different risk factors and predictive factors) are extracted from its sub populations. Therefore, those two sub populations cannot be treated the same. If data is not stratified into sub populations when the branching question is determined to be significant, one may skip some important risk factors and predictive factors.

2.4. Imputation Using Rough Set Theory

In rough set theory, data is stored in an information table. Let $I = (U, A)$ be an information system, where U be a set of objects and A be a non-empty set of attributes such that $a: U \rightarrow V_a$. V_a be the set of values attribute a may take. The information table assigns a value $a(x)$ from V_a to each attribute a and object x in the universe U.

With any $P \subseteq A$ there is an equivalence relation $IND(P)$: $IND(P) = \{(x, y) \in U^2 | \forall a \in P, a(x) = a(y)\}$. The objects that are the elements of an equivalence class are indistinguishable. Let $X \subseteq U$ represent the attribute subset P; and an arbitrary set of objects comprising X, and we wish the express this subset of objects using the equivalence classes induced by attribute subset P. Since X cannot be expressed exactly (because there may cases when some objects will be included and/or excluded due to the indistinguishable relations) the lower bounds $\underline{P}X = \{x | [x]_P \subseteq X\}$ and upper bounds $\overline{P}X = \{x | [x]_P \cap X \neq \emptyset\}$ of X can be defined. In other words, either all the objects being part of the equivalence class need to be included if at least one of them is an object comprising X (upper bound), or all of the objects being part of the equivalence class need to be excluded (lower bound).

As you may recall, the *GMD*s were determined in Section 2.1.2. They are the only incomplete data type that is imputed in this study. The first step of imputation starts with dichotomizing the MESA data. 0 denotes a subject towards continence and 1 denotes a subject towards incontinence. The equivalence classes are determined from the dichotomized data. The subjects containing one or more *GMD* are assigned to the appropriate equivalence classes, if any. Note that one subject with one or more GMD can be assigned to more than one equivalence class. The algorithm implemented below (Algorithm 2) imputes the GMD by presenting a list of possible values, based on the observed data within the same equivalence class(es). The hypothesis here is that in most finite databases, a case similar to the missing data case could have been observed before.

Algorithm 2. Rough set based missing data imputation algorithm

```
Input:       Incomplete data set $\Lambda$ with $a$ attributes and $i$ instances.
             All these instances should belong to a decision $D$.
Output:      Imputed dataset.
Assumption:  $D$ and some attributes are always known.
For all $i$ do → Partition the input space according to $D$ End
For each attribute do →
The family of equivalent classes $\varepsilon(a)$ containing each object $o_i$ for all input attributes is computed.
IF $i$ has the same attribute values with $a_j$ everywhere except for the missing value, replace the missing value, $a_{missing}$, with the value $v_j$, from $a_j$, where $j$ is an index to another instance within the same equivalence class.
IF more than one $v_j$ values are suitable for the estimation, postpone the replacement for later when it will be clear which value is appropriate.
End
```

2.4.1 Imputation Validation

In MESA, 44% of the subjects (instances) contain at least one missing attribute. In order to validate the imputation technique, all the subjects containing at least one missing value are excluded from the dataset forming a subset of MESA without any missing values (referred to as $D_{complete}$). Next, the values of several attributes are randomly removed from $D_{complete}$ in order to build a new dataset that has 44% of the subjects containing at least one missing value (44% is specifically chosen in order to reflect the actual case). This dataset is referred to as $D_{simulation}$.

The rough set imputation is applied on the $D_{simulation}$ data. All the instances are partitioned into equivalence classes in the $D_{simulation}$ data. The missing attribute , $a_{missing}$, of instance i is imputed with the value v_j, from a_j, where j is an index to another instance (within the same equivalence class) that has the same attribute values with a_j everywhere except for the missing value. This process is repeated for each instance with missing attributes. If there exist more than one v_j value suitable for the missing attribute, the imputation process is postponed until only one value is appropriate. Note that, this method does not guarantee each missing attribute to be imputed. Some attributes may still remain missing at the end of imputation.

As mentioned above, the missing attributes are imputed when there exists at least one instance identical to the missing data case except for the missing attribute itself. In order to compare the performance of rough set imputation, we have extended this process to imputing the missing case only when there exists at least a predetermined number of identical instances to the instance with the missing case. The imputed values are then compared with the actual values in the $D_{complete}$ data. This process is repeated several times, by selecting a different random set of attributes to remove each time. After the removed attributes are imputed and the imputed values are compared with the actuals, the average imputation accuracy is then reported.

CHAPTER THREE

METHOD - PREDICTIVE INDEX ESTIMATION

3.1. Comparison of Attribute Selection Methods

Attribute selection techniques can be divided into three different categories: filter, wrapper and embedded methods. In filtering methods, the attribute selection method takes place before any learning algorithm. The undesirable attributes are filtered out before the classification step. All the training data is used in filtering methods (Hall 2003). In embedded methods, the learning algorithm has its own attribute selection algorithm embedded in it (Molina 2002). J48 decision tree classification algorithm is a common example of an embedded method. In wrapper mode, on the other hand, the attribute selection algorithm uses the learning algorithm as a sub-routine (John 1994).

Five different attribute selection methods are applied to the MESA dataset. Wrapper methods which are Correlation based feature selection and ReliefF, a filtering method, information gain and an embedded method J48 decision tree based attribute selection are applied to the dataset and the results are compared in terms of sensitivity and specificities.

Correlation based feature selection. Correlation based feature selection evaluates the dependencies between attributes and eliminate the ones which are correlated to each other. The irrelevant and redundant data has to be removed. After the attribute selection, the remaining data has to be highly correlated with the class and uncorrelated with each other.

As equation 1 (Ghiselli, 1964) formalizes:

$$Merit_{s_k} = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

The attribute subset S contains k different attributes where r_{cf} is the attribute to class correlation and r_{ff} is the attribute to attribute correlation. In order to have a good attribute selection algorithm the merit has to be maximized. Symmetrical uncertainty can be evaluated as follows where $H(X)$ and $H(Y)$ are marginal entropies.

$$Symmetrical\,uncertainty = 2.0x\left[\frac{gain}{H(Y)+H(X)}\right]$$

Consistency based feature selection. The consistency of the class is evaluated by first figuring out all different combinations of the attributes. For each different combination the consistency is calculated by differentiating the number of occurrences of a particular attribute from the cardinality of the majority class (Hall 1998).

Information gain. The uncertainty of the class is evaluated with and without the attribute observation (Hall 1998).

$$H(Class) = -\sum\nolimits_{c\,\in C} p(c)\log p(c)$$

$$H(Class|Attribute) = -\sum_{a\epsilon A} p(a) \sum_{c\epsilon C} p(c|a)\log p(c|a)$$

$$InfoGain(Class, Attribute) = H(Class) - H(Class|Attribute)$$

ReliefF. ReliefF algorithm assigns a relevancy score to all the attributes in descending order. The algorithm selects an instance in each iteration and finds the nearest neighbor (by Euclidean distance) from the same and opposite class. The algorithm starts with a *p*-long weight vector (W) of zeros. The closest same-class instance is called 'near-hit', and the closest different-class instance is called 'near-miss'. The weight vector is updated in each iteration as shown in the following formula:

$$W_i = W_{i-1} - (x_i - nearHit_i)^2 + (x_i - nearMiss_i)^2$$

Therefore, the weight of a given feature decreases if it differs from that feature in nearby instances of the same class more than nearby instances of the other class, and increases in the reverse case. After *m* iterations, each element of the weight vector is divided by *m* which will give the relevance vector. The algorithm can handle noise if the neighbor number, k, is increased.

J48 decision tree based attribute selection. J48 decision tree is a classification method that can also be used for attribute selection. Each node of a tree involves an attribute and the occurrence of each attribute provides information about the importance of that particular attribute (Sugumaran 2006). The importance of each attribute can be evaluated by applying the information gain formula to each node of the tree.

3.2. Selection of Potential Predictive Factors

The construction of urinary incontinence predictive index is based on subjects who were classified as continent at the baseline. The idea behind the construction of predictive index is to model what combination of factors makes a continent subject change from a continent condition to an incontinent condition. 91 subjects are classified as incontinent in the first follow up (HH2) out of 424 subjects in the baseline (HH1). As mentioned before, the factors are all dichotomized where 0 denoting a subject towards continent and 1 denoting a subject towards incontinent. A logistic regression model is used where the independent variables are the factors of baseline and the dependent variable is the class label (outcome) of HH2. The same factors from HH1, along with the outcome from HH4 are then used to determine the p-values and odd ratios. The reason of this attempt is to see if the same factors are still significant when a different outcome (response variable) is used.

Relieff attribute selection technique is also used to see whether the potential predictive factors extracted are similar with the regression method. Relieff assigns a weight to each attribute and orders the risk factors in descending order. The attributes are selected with the highest weight until a predetermined threshold value.

Interaction effects represent the combined effects of variables on the criterion or dependent measure. An *interaction* occurs when the magnitude of the effect of one independent variable (X) on a dependent variable (Y) varies as a function of a second independent variable (Z). Adding an interaction term to a model drastically changes the interpretation of all of the coefficients**.** In this study, *2-way interactions* of the regression model are determined. The new *p*-values of the potential predictive factors along with the most significant *2-way interaction* factors for HH1HH2 data (the baseline factors and the first follow-up response) are determined.

Once the potential predictive factors are extracted from the Relieff attribute selection technique and the logistic regression method, the data containing only the predictive factors along with the class labels are used for association rules generation. The class labels contain both the class from the baseline and the first follow up. The four different class labels are denoted by, ‘C-C’, ’C-I’, ’I-C’, ’I-I’ where ‘C-I’ showing the subject being continent in the baseline and incontinent in the first follow-up. The dimensionality is than further reduced by removing the subjects that have the following class labels, ‘I-I’, ‘I-C’. The reason of removing those class labels is that, we are only interested in the subjects that became incontinent over time and the ones that remain continent.

3.3. Comparison of Rule Extraction Methods

Rule extraction techniques can be categorized into several different categories. In this research we will focus on the association rule learning and classification rules. Four different rule extraction methods are studied. Apriori an association rule and PART, Prism and Jrip the classification rules are examined and the results are compared in terms of sensitivity and specificities.

Apriori is an association rule available on the data mining tool Weka (Waikato Environment for Knowledge Analysis) that iteratively reduces the minimum support until it finds the required number of rules with the given minimum confidence. PART algorithm uses a divide-and-conquer approach and builds a decision tree. The “best” leafs of the tree become part of the rules. For the Jrip, a rule learner is implemented to construct the classification rules. Prism can only deal with nominal attributes. It doesn't do any pruning. Prism algorithm cannot deal with missing values. Predictive Apriori and Tertius were also two rule extraction methods available in Weka. But those two methods were not as effective in terms of the run times compared to others. This may cause inefficient results since MESA data has a high dimensional attribute space.

3.3.1. Simulation

10 different simulation datasets are created each containing 20 attributes and 1500 subjects. In order to gain full control over the dataset the values in the dataset are formed to create rules (Agraval 1994). The last column of the dataset specifies the resultants i.e. the class labels. Resultant is defined as continence, expressed as 0, or incontinence, expressed as 1, of the subject. Resultant values are stored in the column space of the matrix.

Subjects are defined as an individual and are represented in the row space of the matrix. The number of subjects is a function of the number of rules and the partition size for each rule. For example with ten rules and two hundred subject partition size there would be two thousand subjects represented in the matrix, with subjects one to two hundred being in the first partition, subjects two hundred and one to four hundred in the second, etc.

The rules that are embedded to the simulation data are as follows:

$(A_1 = 0)\ AND(A_4 = 0) \Rightarrow Y = 0$

$(A_2 = 0)\ AND(A_8 = 1) \Rightarrow Y = 1$

$(A_{10} = 0)\ AND(A_{15} = 1) \Rightarrow Y = 0$

$(A_6 = 1)\ AND(A_{13} = 1) \Rightarrow Y = 0$

$(A_9 = 0)\ AND(A_{14} = 1) \Rightarrow Y = 1$

Having defined the base dataset, it is now possible to manipulate the dataset for further methodology testing. There are two types of modifications that can be performed to the data set for this testing. First is the inclusion of noise to the data set. The second type of modification is to incorporate missing values. The expected results of the attribute selection algorithms are the attributes that are the entities of the rules embedded to the datasets.

A simulation dataset is used to allow maximum flexibility in creating and manipulating the data sets. The simulation allows for one base dataset to be processed with multiple noise levels, multiple incomplete data levels and a mixture of both noise and incomplete data. The advantage of this methodology allows the researcher to understand the impact of varying levels of these factors on attribute selection or any other metric of interest. Further, we can determine up to what percentage of incomplete data and noise levels a given data mining tool can handle. This is extremely important when one needs to use the tool on a real application. The current version of the matrix creation algorithm was intentionally restricted to binary data only. This was chosen primarily to validate the attribute selection methods that were the driving force behind creating this simulated data. With the methodology described above, modifications to alternative data types, such as categorical and/or continuous, will be possible.

The simulation shows up to what percentage of incomplete data, noise and multicollinearity can be handled by the attribute selection and rule extraction methods. If the level of incomplete data is more than the percentage that can be handled, further preprocessing is required.

3.4. Estimating the Predictive Index

A scientifically developed and tested predictive UI index would help to identify women who are most likely to develop UI and permit widespread prevention or early treatment. Therefore a method is generated that calculates the probability of a new patient being incontinent in the future.

The HH1HH2 data is prepared for association rule mining. Apriori algorithm explained in Section 3.3 is used to determine the predictive index. However, the confidence and support of each rule were less than desired for this method. Also, not all combinations of the attribute values are listed, therefore; we cannot determine the status of a patient having a combination that was not listed in any of the rules. Since, the results of Apriori algorithm are not reliable and also limited to the listed rules; we listed the risk factors with all different combination of values they can take. This experiment will return all the rules that are available. For each rule the support is calculated, that is, the number of subjects that satisfy the rule. The probability can then simply be calculated by dividing the number of subjects that satisfy the antecedent of the rule by the number of subjects that satisfies the entire rule. This way when a new subject comes into the clinic, the probability of that subject going towards incontinence can be calculated by comparing her/his rules with the model and reporting that rule's probability measure.

3.4.1. Reliability of the Rules

Once each combination of the rules are determined, it is important whether or not these rules can be referred as reliable. For a small sample size, there may be very few or even no subjects that satisfy the rule. For example if there is a rule that is only satisfied by one subject, and contradicts with no subjects, the confidence of that rule will be 100%. An example of this rule can be the following: A10T1='(0.75-inf)' A12T1='(0.75- inf)' A3T1='(-inf-0.25]' A4T1='(0.75-inf)' *support:1* ==> class=C-I *support:1*. The antecedent of this rule is supported by 1 subject, and the rule including the consequent is supported by 1 subject. However, it will be a misleading assumption to conclude that there is a 100% probability that a new subject satisfying this rule will go towards incontinence over time. Therefore, it is important to determine a support threshold. The margin of errors is used to determine a support threshold for each rule. The rules that have less support than the threshold are considered as unreliable.

In a confidence interval, the range of values above and below the sample statistic is called the margin of error. If $npq > 5$, then the normal approximation can be used to develop a confidence interval for a binomial variable. The formula is the following:

$$|p - \hat{p}| \leq Z_{\alpha/_2} \sqrt{\frac{p_0(1 - p_0)}{n}}$$

p_0 is the probability calculated from the rule by dividing the number of the number of subjects that satisfy the entire rule by the number of subjects that satisfies the antecedent of the rule. n is the number of subjects that support the rule. $|p - \hat{p}|$ is the margin of error. The following are critical values for common levels of confidence.

A 90% level of confidence has $\alpha = 0.10$ and critical value of $z_{\alpha/2} = 1.64$.

A 95% level of confidence has $\alpha = 0.05$ and critical value of $z_{\alpha/2} = 1.96$.

A 99% level of confidence has $\alpha = 0.01$ and critical value of $z_{\alpha/2} = 2.58$.

A 99.5% level of confidence has $\alpha = 0.005$ and critical value of $z_{\alpha/2} = 2.81$.

3.4.2. Model Based Approach

This section deals with the construction of a predictive index for urinary incontinence based on predicted probabilities of all possible combinations of the selected predictive factors. Between the two regression models, with and without interaction terms, the one with higher performance will be used for the next steps. The selected predictive factors, and if needed their 2-way interaction terms, from the baseline are regressed against the class label of the follow-up. This is achieved by constructing contrasts and their corresponding 95% confidence intervals for probabilities of incontinence.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1. Incomplete and Inconsistent Data Analysis Results

The first step of MESA preprocessing is to categorize the missing values into GMD, UMD and SPMD. Table 4.1 shows the number of subjects in the MESA baseline. Table 4.2 shows the number of cells that were incomplete before the experiment, along with their percentages. Once the experiments are conducted over the entire MESA dataset, 200 responses are determined to be mutually inconsistent, 15.4% of the responses are *GMD*, 12.9% of the responses are *UMD*, and 9.8% of the responses are *SPMD*. Table 4.3 shows the results after the incomplete data analysis.

The summation of the percentages of *GMD*, *UMD* and *SPMD* for the female population shown in Table 4.3 must be equal the total percentage of incomplete data (37.1%). Since inconsistent data is not a type of incomplete data, the percentage of inconsistent data (0.021%) is not included in the summation.

Table 4.4 shows the percentage of data with the range of GMD/UMD and SPMD's. We know from our previous experiments that our methods can handle up to 12% of missing values and 15% of noise (Arslanturk et al. 2011). When the percentage of missing values and/or noise exceeds this threshold, the method shows less than desired reliability. Therefore, the dimensionality of the data is further reduced by removing the attributes and subjects that contain more than 15% of *GMD* and/or *UMD.* The attributes and subjects that contain more than 15% of *SPMD* are not removed, since those will later be used for stratification.   The size of the reduced dataset has 773 attributes and 1059 subjects.

Note that, the skip patterns cover 9.8% of the missing values. It is important to distinguish them from the GMD's. The branching questions which lead the skip patterns to occur can now be used for data stratification. This approach will help us to analyze each population with different characteristics separately.

4.2. Stratification Results

The stratification results of the simulated data and MESA data are explained in this section.

4.2.1. Simulation Results of Stratification

Table 4.5 shows the simulation results of the stratification with 0% noise. The first three rules (R_1, R_2, R_3) belong to the dataset that contains the subjects that have $BQ_1^1 = 0$, which can be denoted by $D_{BQ_1^1}$, and the last three rules (R_4, R_5, R_6) belong to the dataset that contains the subjects that have $BQ_1^2 = 1$, which can be denoted by $D_{BQ_1^2}$. Since the first branching question, BQ_1, was designed to separate D_1 and D_2, it is expected to extract the same rules listed in Section 2.3.1 for the subsets, $D_{BQ_1^1}$ and $D_{BQ_1^2}$. The rules extracted from $D_{BQ_1^1}$ are the same as the rules extracted from D_1, and the rules extracted from $D_{BQ_1^2}$ are the same as the rules extracted from D_2. Note that, the support of the first three rules (R_1, R_2, R_3) of dataset $D_{BQ_1^1}$, is much higher than the support of the first three rules (R_1, R_2, R_3) of dataset $D_{BQ_1^2}$ and the non-support of the first three

Table 4.1- The size of HH1

	# of Subjects	# of Attributes
HH1	1956	826
HH1 Female	1154	
HH1 Male	802	

Table 4.2- Missing Data in HH1, gender specified

Missing Data	# of Cells	Percentages
HH1	751132	46.4%
HH1 Female	353637	37.1%
HH1 Male	321637	48.5%

Table 4.3- *GMD*, *UMD*, *SPMD* and Inconsistent Data for Female

HH1 Female	#of Cells	Percentages
GMD	157278	15.4%
UMD	114384	12.9%
SPMD	81975	9.8%
Inconsistent	200	0.021%

Table 4.4- Average number of missing values per attribute and subject

Avg. Missing Data	By Attribute	By Subject
GMD UMD 0-15	93.5%	91.7%
GMD UMD>15	6.4%	8.2%
***SPMD* 0-15**	65.4%	0.33%
***SPMD* >15**	34.5%	99.6%

rules (R_1, R_2, R_3) of dataset $D_{BQ_1^1}$ is equal to 0. The reason of the non-support being 0 is that, the rules were generated without any conflict. Likewise, for the second branch, BQ_1^2 , we expect to see a lower support and a higher non-support compared to BQ_1^1, for the first three rules. Notice that, for the last three rules the support of BQ_1^2 is higher than BQ_1^1 and the non-support of BQ_1^2 is lower than BQ_1^1, since the last three rules were extracted from D_2.

The p-values are calculated and the results show that, for each rule, the association between the rules and two populations are considered to be extremely statistically significant. Therefore, we can define, $BQ_1,$ as a significant branching question. That means, the two populations, BQ_1^1 and BQ_1^2 has different association rules hence; they have to be analyzed separately.

Table 4.6, 4.7, and 4.8 show the same experiment with 10%, 20% and 30% noise, respectively. Note that, even if there is 30% noise in the data, the p-values are still considered to be statistically significant. However, the comparison of the p-values with different noise levels also show that, the less the noise there is in the data, the smaller the p-value becomes.

Table 4.9 shows the contingency table for the second branching question, BQ_2. As mentioned in the Methods section, the binary values of the second branching question were assigned randomly, and based on the binary values BQ_2 have, $D_1 + D_2$ dataset was separated into $D_{BQ_2^0}$ and $D_{BQ_2^1}$.

Table 4.5- Optimal Branching (OB) p-values for 0% Noise

Association Rules	**Noise: 0%**	$D_{BQ_1^0}$	$D_{BQ_1^1}$	**Row T.**	**p-value**
A1T1='(-inf -0.25]' A3T1='(0.75-inf)' 200 ==> class=I-C 200	Support	200	17	217	3.224e^{-46}
	No Support	0	61	61	
	Column Total	200	78	**556**	
A5T1='(-inf-0.25]' A7T1='(0.75-inf)' 200 ==> class=C-I	Support	200	31	231	3.730e^{-25}
	No Support	0	34	34	
	Column Total	200	65	**530**	
A10T1='(0.75-inf)' A12T1='(0.75-inf)' 200 ==> class=C-I 200	Support	200	33	233	7.54e^{-30}
	No Support	0	43	43	
	Column Total	200	76	**552**	
A2T1='(-inf-0.25]' A11T1='(-inf-0.25]' 200 ==> class=C-I 200	Support	26	200	226	4.92e^{-29}
	No Support	38	0	38	
	Column Total	64	200	**528**	
A4T1='(0.75-inf)' A9T1='(-inf-0.25]' 200 ==> class=C-I 200	Support	28	200	228	1.30e^{-29}
	No Support	40	0	40	
	Column Total	68	200	**536**	
A6T1='(-inf-0.25]' A8T1='(0.75-inf)' 200 ==> class=I-C 200	Support	10	200	210	2.74e^{-37}
	No Support	40	0	40	
	Column Total	50	200	**500**	

Table 4.6- Optimal Branching (OB) p-values for 10% Noise

Association Rules	**Noise: 10%**	$D_{BQ_1^0}$	$D_{BQ_1^1}$	**Row T.**	**p-value**
A10T1='(0.75-inf)' A12T1='(0.75-inf)' 137 ==> class=C-I 129	Support	129	16	145	$3.02e^{-17}$
	NoSupport	8	33	41	
	Column Total	137	49	**372**	
A5T1='(-inf-0.25]' A7T1='(0.75-inf)' 146 ==> class=C-I	Support	135	24	159	$3.94e^{-14}$
	NoSupport	11	34	45	
	Column Total	146	58	**408**	
A1T1='(-inf-0.25]' A3T1='(0.75-inf)' 143 ==> class=I-C 129	Support	129	7	136	$1.57e^{-33}$
	NoSupport	14	67	81	
	Column Total	143	74	**434**	
A4T1='(0.75-inf)' A9T1='(-inf-0.25]' 150 ==> class=C-I 140	Support	26	140	166	$2.70e^{-15}$
	NoSupport	36	10	46	
	Column Total	62	150	**424**	
A6T1='(-inf-0.25]' A8T1='(0.75-inf)' 147 ==> class=I-C 135	Support	12	135	147	$5.12e^{-23}$
	NoSupport	45	12	57	
	Column Total	57	147	**408**	
A2T1='(-inf-0.25]' A11T1='(-inf-0.25]' 155 ==> class=C-I 141	Support	18	141	159	
	NoSupport	35	14	49	
	Column Total	53	155	**416**	

Table 4.7- Optimal Branching (OB) p-values for 20% Noise

Association Rules	**Noise: 20%**	$D_{BQ_1^0}$	$D_{BQ_1^1}$	**Row T.**	**p-value**
A5T1='(-inf-0.25]' 169 ==> class=C-I 126	Support	126	54	180	4.47e^{-9}
	No Support	43	78	121	
	Column Total	169	132	**602**	
A5T1='(-inf-0.25]' A7T1='(0.75-inf)' 173 ==> class=C-I 128	Support	128	59	187	3.80e^{-9}
	No Support	45	84	129	
	Column Total	173	143	**632**	
A5T1='(-inf-0.25]' A7T1='(0.75-inf)' 173 ==> class=C-I 124	Support	124	54	178	4.93e^{-7}
	No Support	49	73	122	
	Column Total	173	127	**600**	
A2T1='(-inf-0.25]' A11T1='(-inf-0.25]' 149 ==> class=C-I 114	Support	46	114	160	3.06e^{-9}
	No Support	68	35	103	
	Column Total	114	149	**526**	
A4T1='(0.75-inf)' A9T1='(-inf-0.25]' 151 ==> class=C-I 115	Support	42	115	157	4.96e^{-15}
	No Support	96	36	132	
	Column Total	138	151	**578**	
A11T1='(-inf-0.25]' 161 ==> class=C-I 122	Support	43	122	165	1.36e^{-13}
	No Support	83	39	122	
	Column Total	126	161	**574**	

Table 4.8- Optimal Branching (OB) p-values for 30% Noise

Association Rules	**Noise:30%**	$D_{BQ_1^0}$	$D_{BQ_1^1}$	**Row T**	**p-value**
A12T1='(0.75-inf)' 263 ==> class=C-I 150	Support	150	105	255	0.0018
	No Support	113	139	252	
	Column Total	263	244	**1014**	
A10T1='(0.75-inf)' A12T1='(0.75-inf)' 257 ==> class=C-I 143	Support	143	90	233	7.24e^{-005}
	No Support	114	149	263	
	Column Total	257	239	**992**	
A4T1='(0.75-inf)' 249 ==> class=C-I 145	Support	96	145	241	4.69e^{-005}
	No Support	146	104	250	
	Column Total	242	249	**982**	
A9T1='(-inf-0.25]' 269 ==> class=C-I 152	Support	81	152	233	7.62e^{-005}
	No Support	131	117	248	
	Column Total	212	269	**962**	

Since the BQ_2 values are assigned randomly, once separated, $D_{BQ_2^0}$ and $D_{BQ_2^1}$ datasets each may contain subjects from both D_1 and D_2. Therefore, the rule extraction technique will extract rules belonging to both D_1 and D_2. The rules embedded to D_1 can lead conflicts, for subjects belonging to D_2, and vice versa. These conflicts will effect the support of each rule, and may also corrupt the rules.

Therefore, Table 4.9 shows that the association between the rules and two populations are considered to be not statistically significant. Table 4.10 shows the results with 10% noise where the p-value shows an increase. Therefore, BQ_2 is not being considered as a significant branching question.

Table 4.9- Bad Branching (BB) p-values for 0% Noise

Association Rules	**Noise: 0%**	$D_{BQ_2^0}$	$D_{BQ_2^1}$	**Row T.**	**p-value**
A4T1='(0.75-inf)' A9T1='(-inf-0.25]' 136 ==> class=C-I 117	Support	117	111	128	0.7326
	No Support	19	21	40	
	Column T.	136	132	**436**	
A2T1='(-inf-0.25]' A11T1='(-inf-0.25]' 132 ==> class=C-I 112	Support	112	114	226	0.8610
	No Support	20	18	38	
	Column Total	132	132	**528**	
A5T1='(-inf-0.25]' A7T1='(0.75-inf)' 124 ==> class=C-I 107	Support	124	107	231	0.7160
	No Support	17	17	34	
	Column Total	141	124	**530**	
A10T1='(0.75-inf)' A12T1='(0.75-inf)' 123 ==> class=C-I 109	Support	124	109	233	0.0962
	No Support	29	14	43	
	Column Total	153	123	**552**	

Table 4.10- Bad Branching (BB) p-values for 10% Noise

Association Rules	**Noise: 10%**	$D_{BQ_2^0}$	$D_{BQ_2^1}$	**Row T.**	**p-value**
A4T1='(0.75-inf)' A9T1='(-inf-0.25]' 152 ==> class=C-I 109	Support	109	109	218	0.3932
	No Support	43	54	97	
	Column Total	152	163	**630**	
A10T1='(0.75-inf)' A12T1='(0.75-inf)' 158 ==> class=C-I 108	Support	108	87	315	0.4590
	No Support	50	49	99	
	Column Total	158	136	**708**	
A4T1='(0.75-inf)' 156 ==> class=C-I 106	Support	106	97	203	0.4721
	No Support	50	55	105	
	Column Total	156	152	**616**	
A5T1='(-inf-0.25]' A7T1='(0.75-inf)' 159 ==> class=C-I 110	Support	96	110	206	**0.3361**
	No Support	55	49	104	
	Column Total	151	159	**620**	
A2T1='(-inf-0.25]' A11T1='(-inf-0.25]' 152 ==> class=C-I 102	Support	104	102	206	1.0000
	No Support	52	50	102	
	Column T.	156	152	**616**	

As a result, BQ_1 was defined as a factor that perfectly splits the dataset into two populations with different characteristics, BQ_2 and BQ_3 were defined as factors that randomly split the population into two groups. Our simulation results showed that BQ_1 is a significant factor, whereas BQ_2 and BQ_3 are not. Hence, this simulation shows that the Fisher's Exact Test can be used to determine the significant branching questions.

4.2.2. MESA Results of Stratification

The same experiment is applied on the MESA dataset. The branching questions that are extracted from the MESA data are listed in Table 4.11.

We have split the data into two subsets based on the following branching questions: v29, v92, v171, v180, v193, v418, v737 and v745. For most of these factors except v180 and v737, there were not enough subjects for a rule to be generated. Therefore, a contingency table could not be generated. Table 4.12 shows the contingency table for the subjects that have undergone female surgery and the ones that have not. The p-values show that the association between the rules and the data is not considered to be statistically significant. Therefore, female surgery cannot be considered as a significant branching question.

Table 4.13 shows the contingency table for the smokers and non-smokers. Here, the association between the rules and the data is considered to be statistically significant for the rules whose p-values are underlined. Therefore, smoking more than 100 cigarettes in the entire life is considered to be a significant branching question.

Table 4.11- List of Branching Questions

Label	Branching Questions
v32	Are you married, widowed, divorced, separated, or have you never married?
v41	Do you sneeze often, sometimes, rarely or never?
v53	Do you usually need to use a wheelchair, cane, crutches or walker to help you get around?
v59	Do you have any health problems which make it difficult for you to leave your home and go visiting, shopping, or to the doctor’s?
v87	Have you ever been told by a doctor that you had high blood pressure?
v90	Have you ever been told by a doctor that you had a hernia in the groin or stomach area?
v92	Have you ever had a stroke or cerebral brain hemorrhage?
v95	Have you ever been told by a doctor that you occasionally have had transient ischemic attacks or poor blood flow to the brain, where you seem to lose track of things that are happening around you for up to a few minutes?
v97	Have you ever had problems with any paralysis?
v107	Have you ever had a heart attack?
v111	Has any doctor ever told you that you have or have had arthritis or rheumatism?
v125	Have you had any other disease of the nerves or muscles?
v128	Have you ever been told by a doctor that you had cancer of any kind?
v133	Have you lost any inches in height as you have gotten older?
v135	In the last 12 months how many times have you become so dizzy that you fainted or nearly fainted?
v138	Have you broken any bones in the last 12 months?
v171	How many pregnancies have you had?
v180	Have you ever had female surgery such as on your ovaries, vagina, fallopian tubes, uterus, rectum, or urethra?
v193	Are you currently taking any female hormones?
v207	Have you ever had any other operations on your bladder, kidneys, or any other organs in your pelvic area or area normally covered by underpants or undershorts?
v219	Do you usually need help in getting into the bathroom or on or off of the toilet?
v220	Do you use any aids like a grab bar, or special toilet or anything else to help you with using the toilet in your home?
v224	Do you have your regular schedule that you usually use to get you to the toilet to urinate, for example every hour or so?
v415	Have you ever had or have you been told by a doctor that you had any kidney or bladder problems we haven’t talked about already?
v418	Do you usually drink any liquids of any kind before you go to bed at night?
v422	Did your mother of father have a urine loss condition as an adult?
v458	Do you have any health problems that require medical attention that you have not been able to get treated?

Table 4.11- List of Branching Questions- Continued

Label	Branching Questions
v496	As you know some people experience memory problems as they get older. How about your memory? Has it become worse within the last five years?
v540	Have you ever had to stay in a nursing home overnight or longer because of a health problem you had?
v543	Have you ever had to stay in a mental health facility overnight or longer, because of a mental or emotional problem that you had?
v725	Do you usually take one or more naps during the day?
v737	Have you smoked at least 100 cigarettes in your entire life?
v745	Do you drink wine, beer, or liquor?
v776	Is this the same occupation that you had for most of your life?

4.2.2.1 Analysis of the Stratums. The extracted rules of the significant branching questions can be used for prediction. For example when a new patient comes into the clinic who meets the following rule v69='(-inf-0.5]' v80='(-inf-0.5]' v124='(-inf-0.5]' v152='(-inf-0.5]' v230='(-inf-0.5]' 50 ==> class=C-C 37 conf:(0.74), the patient will remain continent with 74% of probability.

Next, the risk factors of different stratums (smokers/non-smokers for our case) are determined separately. Relieff attribute selection is used to determine the important attributes of each population, since our previous studies have shown that Relieff outperformed other attribute selection techniques on the analysis of MESA dataset (Arslanturk et al.). The extracted attributes are defined as the risk factors. Table 4.14 shows the risk factors for smokers and non-smokers.

Table 4.12- Contingency Table of Female Surgery

Female Surgery- Association Rules		D_{FS}	D_{NFS}	**Row T.**	**p-value**
1. v124='(-inf-0.5]' v152='(-inf-0.5]' v178='(-inf-0.5]' v195='(-inf-0.5]' v230='(-inf-0.5]' v231='(-inf-0.5]' 78 ==> class=C-C 49 conf:(0.63)	Support	49	38	**87**	0.4988
	No Support	29	29	**58**	
	Column T.	**78**	**67**	**145**	
2. v120='All' v124='(-inf-0.5]' v152='(-inf-0.5]' v178='(-inf-0.5]' v195='(-inf-0.5]' v230='(-inf-0.5]' v231='(-inf-0.5]' 78 ==> class=C-C 49 conf:(0.63)	Support	49	38	**87**	0.4988
	No Support	29	29	**58**	
	Column T.	**78**	**67**	**145**	
3. v121='All' v124='(-inf-0.5]' v152='(-inf-0.5]' v178='(-inf-0.5]' v195='(-inf-0.5]' v230='(-inf-0.5]' v231='(-inf-0.5]' 78 ==> class=C-C 49 conf:(0.63)	Support	49	38	**87**	0.4988
	No Support	29	29	**58**	
	Column T.	**78**	**67**	**145**	
1. v107='(-inf-0.5]' v125='(-inf-0.5]' v199='(-inf-0.5]' v432='(-inf-0.5]' 67 ==> class=C-I 43 conf:(0.64)	Support	32	43	**75**	0.1135
	No Support	33	24	**57**	
	Column T.	**65**	**67**	**132**	
2. v78='(-inf-0.5]' v107='(-inf-0.5]' v125='(-inf-0.5]' v199='(-inf-0.5]' v432='(-inf-0.5]' 67 ==> class=C-I 43 conf:(0.64)	Support	30	43	**73**	0.0782
	No Support	32	24	**56**	
	Column T.	**62**	**67**	**129**	
3. v107='(-inf-0.5]' v120='All' v125='(-inf-0.5]' v199='(-inf-0.5]' v432='(-inf-0.5]' 67 ==> class=C-I 43 conf:(0.64)	Support	32	43	**75**	0.1135
	No Support	33	24	**57**	
	Column T.	**65**	**67**	**132**	

Table 4.13- Contingency Table of Smoking

Smoke- Association Rules		D_S	D_{NS}	**Row T.**	**p-value**
1. v69='(-inf-0.5]' v80='(-inf-0.5]' v124='(-inf-0.5]' v152='(-inf-0.5]' v230='(-inf-0.5]' 50 ==> class=C-C 37 conf:(0.74)	Support	37	41	**78**	0.0064
	No Support	13	42	**55**	
	Column T.	**50**	**83**	**133**	
2. v69='(-inf-0.5]' v78='(-inf-0.5]' v80='(-inf-0.5]' v124='(-inf-0.5]' v152='(-inf-0.5]' v230='(-inf-0.5]' 50 ==> class=C-C 37 conf:(0.74)	Support	37	41	**78**	0.0064
	No Support	13	42	**55**	
	Column T.	**50**	**83**	**133**	
3. v69='(-inf-0.5]' v79='(-inf-0.5]' v80='(-inf-0.5]' v124='(-inf-0.5]' v152='(-inf-0.5]' v230='(-inf-0.5]' 50 ==> class=C-C 37 conf:(0.74)	Support	37	41	**78**	0.0064
	No Support	13	42	**55**	
	Column T.	**50**	**83**	**133**	
1. v78='(-inf-0.5]' v107='(-inf-0.5]' v138='(-inf-0.5]' v432='(-inf-0.5]' 98 ==> class=C-I 58 conf:(0.59)	Support	24	58	**82**	0.1241
	No Support	29	40	**69**	
	Column T.	**53**	**98**	**151**	
2. v28='(-inf-0.5]' v78='(-inf-0.5]' v79='(-inf-0.5]' v138='(-inf-0.5]' v432='(-inf-0.5]' 98 ==> class=C-I 58 conf:(0.59)	Support	24	58	**82**	0.0461
	No Support	34	40	**74**	
	Column T.	**58**	**98**	**156**	
3. v78='(-inf-0.5]' v107='(-inf-0.5]' v120='All' v138='(-inf-0.5]' v432='(-inf-0.5]' 98 ==> class=C-I 58 conf:(0.59)	Support	24	58	**82**	0.1241
	No Support	29	40	**69**	
	Column T.	**53**	**98**	**151**	

Table 4.14- Risk Factors of Stratums

	SMOKERS	NON-SMOKERS
RISK FACTORS	v211- Getting yourself wet	v68- Being proud of yourself
	v128- Having Cancer	v128- Having Cancer
	v229- Difficulties going to the bathroom on time	v719- Having an active hobby
	v89- Being Diabetes	v180- Undergone Female Surgery
	v69- Feeling lonely	v74- Things are going your way?

4.3. Imputation Results

The simulation results of imputation are explained in this section.

4.3.1. Simulation Results

The rough set imputation is applied on the simulated data with 44% of embedded missing values. The missing values are imputed when there exists at least 1 2,3,4 and 7, (Threshold 1, Threshold 2, Threshold 3, Threshold 4, and Threshold 7, respectively) identical instances to the instance with the missing case except for the missing attribute itself. Table 4.15 shows the accuracies of the imputation process with all different thresholds. The first column (Correct) shows the percentage of accurately imputed values while the second column (Incorrect) shows the inaccurately imputed values. At the end of the imputation process, the percentages of attributes that remain missing are shown in the third column (Not imputed). Therefore, one can see that, when the number of identical instances necessary to perform imputation increases (i.e., when the threshold increases), inaccurately imputed values decrease. At the same time, more missing values will remain to be missing at the end of the imputation process.

The 7% of incorrectly imputed values of the validation data is corrected (Threshold 1- Errors Corrected in Table 4.16). The wald score of this data (imputed data with errors corrected) is then compared both with the Threshold 1 (imputed data with 7% of error rate) data and the non-imputed data. The results in Table 4.16 show that the imputed data with the corrected errors outperformed the two other datasets.

Also, a predictive index is constructed, based on the predicted probabilities of all possible combinations of the potential predictive factors (This is explained in detail in Section 4.7). This is achieved by constructing contrasts and their 95% confidence intervals for probabilities of incontinence. For each possible combination (2^n combinations where n is the number of potential predictive factors identified) the difference between the upper and lower bounds of the confidence intervals are identified. The average difference of the n combinations is then calculated and referred as the CI_{Diff}.

Table 4.15- The average imputation results for different threshold levels for several runs

	Correct	**Incorrect**	**Not Imputed**
Threshold 1	36%	7%	56%
Threshold 2	25%	5%	70%
Threshold 3	22%	3%	74%
Threshold 4	21%	3%	75%
Threshold 7	11%	2%	88%

Table 4.16- Testing Global Null Hypothesis

	Test	Chi-Square	Pr > ChiSq
Threshold 1-Errors Corrected	Wald	29.3546	0.0003
Threshold 1-Errors Retained	Wald	28.6637	0.0004
Non-imputed Data	Wald	23.4406	0.0028

$$CI_{Diff} = \frac{\sum_{i=1}^{n}(Upper\ Bound_i - Lower\ Bound_i)}{n}$$

The results of the three different datasets, Table A.1, Table A.2 and Table A.3 (Non-Imputed Data, Threshold 1- Errors Retained and Threshold 1-Errors Corrected,) are then compared based on their CI_{Diff} values. Note that, a tighter confidence interval is preferred. The results show that, the improvement of correcting the 7% of errors with respect to retaining them is 0.009834, whereas the improvement of imputing the data with respect to no imputation is 0.02855.

These results show that, although the rough set imputation causes 7% the data to be incorrectly imputed, the improvement of imputation with respect to leaving the data as it is (i.e. no imputation) is higher. Therefore, in order to generate a more reliable predictive index, imputation needs to be preferred.

4.4. Attribute Selection Results

In this experiment we have applied five different attribute selection algorithms to the simulation dataset and the results are compared when noise, incomplete data and multicollinearity are added to the data. The first set of experiments was designed to evaluate the following attribute selection methods to see which one handles additive noise better: J48, ReliefF, information gain, Consistency based feature selection and Correlation based feature selection.

In order to refer to an attribute selection algorithm as robust, it has to both have a high sensitivity and a high specificity. Figure 4.1 shows that J48 algorithm performs well in terms of sensitivity but there is a huge decrease in the specificity curve (Figure 4.2) which makes the algorithm less desirable than the others when there is noise.

Without any noise, Information gain and ReliefF algorithms both perform well. Consistency based attribute evaluation has a low specificity and J48 decision tree classification algorithm has a low sensitivity. The results do not change in Cfs, Information Gain and ReliefF algorithms when the noise level is 2%, 5%, 10%, 15%, respectively. In spite, in J48 decision tree the specificity decreases and the sensitivity increases rapidly when the noise level increase. The sensitivity of Consistency based feature selection was 11% higher than the average sensitivity of other methods. However, its specificity was 37% lower than that of the average. It is important to note that, the method that maximizes both the sensitivity and specificity is of interest. The best method, when both sensitivity and specificity are considered, is information gain. This method outperformed the average performance of the other methods by 1% and 20.5% when we considered its sensitivity and specificity, respectively. Consistency based subset evaluation and J48 algorithms cannot handle noise.

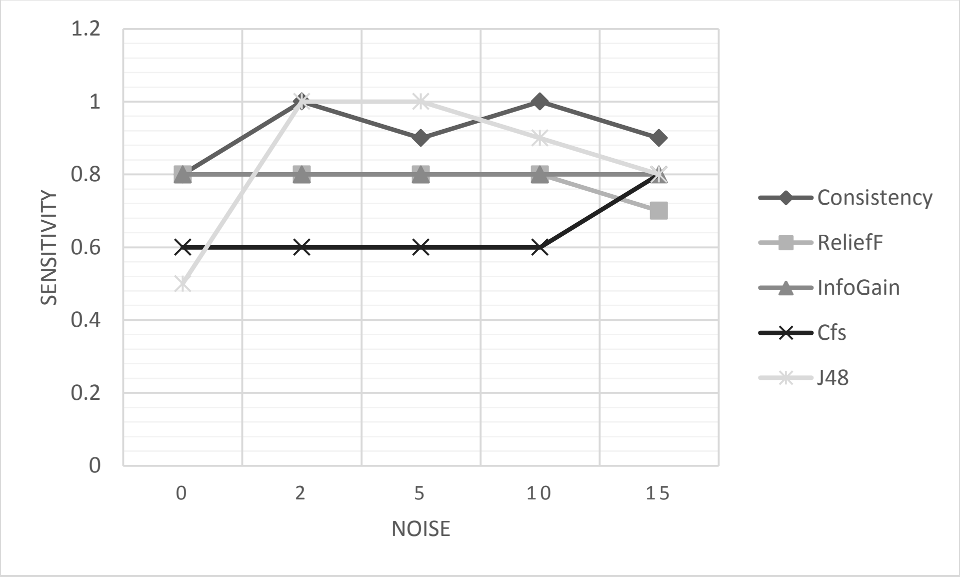

Figure 4.1- Sensitivities of Attribute Selection Methods with Noise Levels

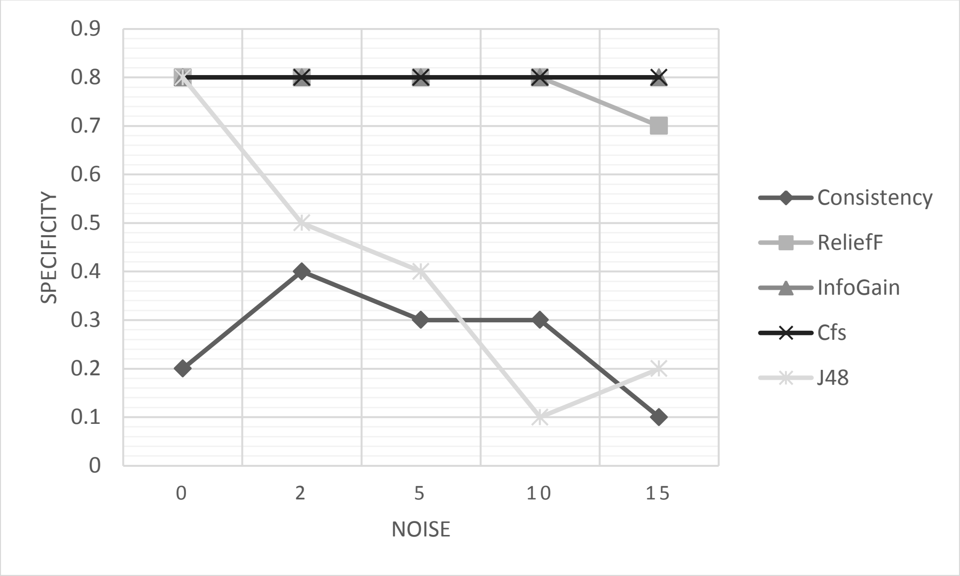

Figure 4.2- Specificities of Attribute Selection Methods with Noise Levels

When multicollinearity is embedded into the dataset without any noise and missing values, the Correlation based feature selection method outperformed other methods.

The next set of experiments is designed to evaluate the attribute selection methods to see which one handles missing values better. Figure 4.3 and Figure 4.4 show the results of different missing value levels. In this case, when considering sensitivity and specificity, ReliefF and information gain are proved to perform better compared to the other methods of our study by 7.2% and 12.4%, respectively. Despite, the Consistency based feature selection and J48 algorithms cannot handle missing values effectively.

In summary, ReliefF and information gain are the best in all three situations (noise, missing value, multicollinearity) when both sensitivity and specificity measures are considered.

It is also important to note that, the reliability of these methods decrease when there is more than 15% of noise or 12% of missing values. Therefore, 15% of noise and 12% of missing values have to be defined as thresholds and data that has more than 15% of noise or 12% of missing values need to be further pre-processed.

4.5. Potential Predictive Factors

The UI risk factors are extracted as body mass index (bmi), sneezing frequency, urine loss problems started after deliveries, trouble getting to the bathroom on time, frequency of wetting/soiling, shutting off the flow of urine in the middle of stream by using muscles and memory problems.

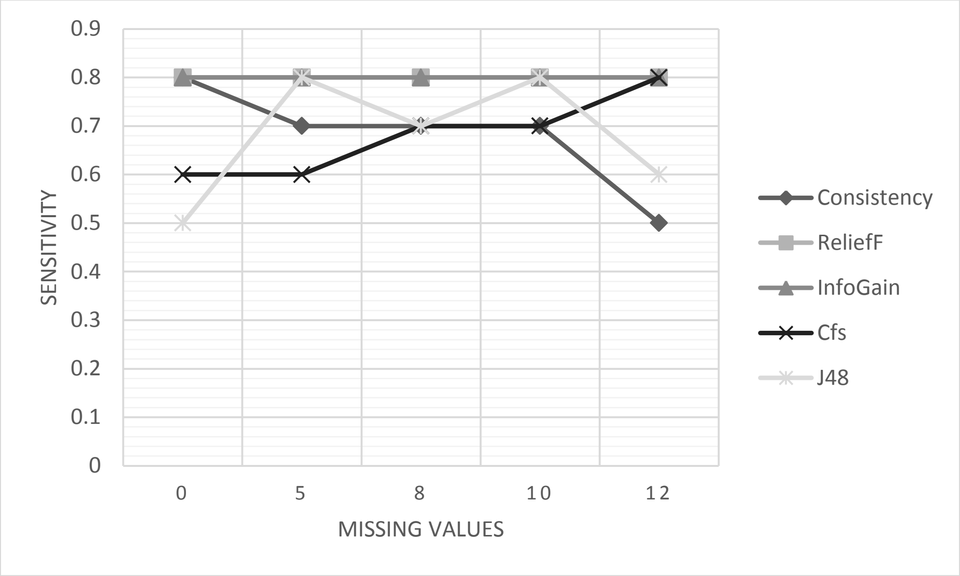

Figure 4.3- Sensitivities of Attribute Selection Methods with Missing Value Levels

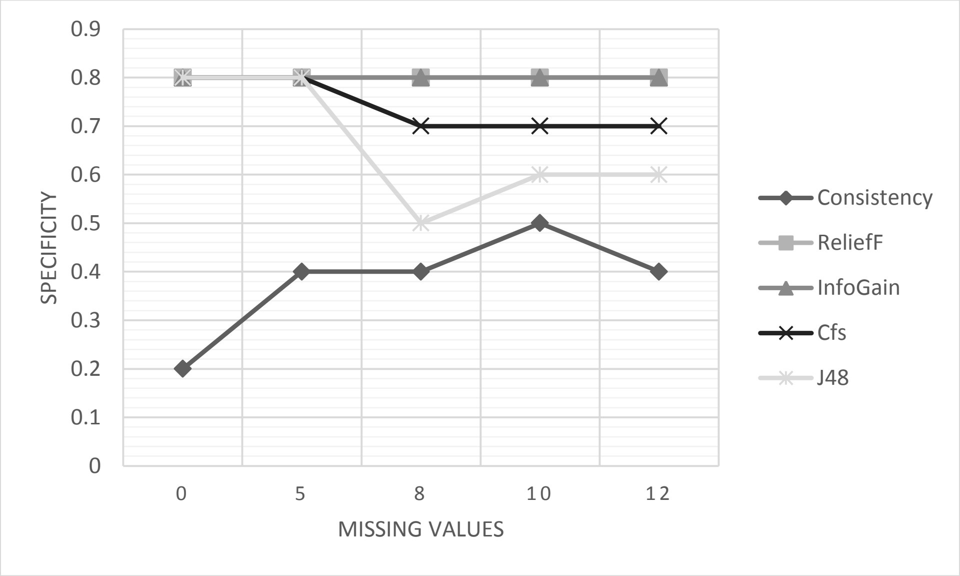

Figure 4.4- Specificities of Attribute Selection Methods with Missing Value Levels

The results of the logistic regression consisting eight most promising factors for prediction are given in Table 4.17 along with their corresponding p-values and odd ratios. The significant factors from the regression method and the Relieff algorithm were identical except a new factor (F6) which wasn't extracted as significant from the regression method. The analysis has shown that the factor (F6) is promising for both the datasets constructed having HH2 and HH4 as outcome variables (and HH1 as baseline input factors). Therefore, it is added to the list of potential predictive factors.

Table 4.18 shows that not all of the eight risk factors identified from HH1HH2 are performing well when we regressed the third follow up response on the same eight baseline factors. F3, F4, F5 and F6 remain to be significant.

However, when the outliers are removed from the dataset, all eight factors are determined as significant both for the HH1HH2 (predictors from HH1 and response from HH2) and HH1HH4 (predictors from HH1 and response from HH4) datasets. A case is described as an outlier if all or at least one half of the binary factors are 0 but the outcome is 1, or the other way round. Table 4.19 shows the results of the p-values for HH1HH2 and HH1HH4 datasets with the removed outliers.

Table 4.20 shows the p-values of HH1HH2 with 2-way interactions of some factors. Possibility of future incontinence and sneezing frequency, memory problems and bmi are the most significant interactions based on their p-values. Once the most significant interactions are determined from HH1HH2 data, HH1HH4 dataset is used to validate the results. The table shows that the same factors are still significant on HH1HH4 data.

Table 4.17- Baseline Factors and First Follow-up Outcome (HH1HH2)

Factor	*Description*	*P-value*	*OR 95% C.I*
F1	Body Mass Index (BMI) <=24; >24	0.1698	(0.827,2.944)
F2	Do you sneeze often, rarely or never?	0.0274	(1.088,4.152)
F3	Any urine loss problems that started after deliveries: Yes;No	0.0423	(1.05,15.313)
F4	Trouble getting to the bathroom on time: Yes;No	0.0769	(0.916,5.539)
F5	Frequency of wetting/soiling yourself day/night: Never; 1/week; 1 or 2/week; >3/week	0.0064	(1.573,16.021)
F6*	When you are urinating into a toilet, can you shut off the flow of urine in the middle of your stream by using your muscles if you want to?	0.0419	(1.031,5.049)
F7	As you know people experience memory problems as they get older. What about remembering names?	0.0672	(0.958,3.552)
F8	Possibility of future incontinence	0.0224	(1.112,4.014)

*:Factor Extracted From Relieff Algorithm

Table 4.18- Baseline Factors and Third Follow-up Outcome (HH1HH4)

Factor	*Description*	*P-value*	*OR 95% C.I*
F1	Body Mass Index (BMI) <=24; >24	0.6109	(0.668,1.987)
F2	Do you sneeze often, rarely or never?	0.6269	(0.527,2.894)
F3	Any urine loss problems that started after deliveries: Yes;No	0.0042	(2.097,51.848)
F4	Trouble getting to the bathroom on time: Yes;No	0.0116	(1.328,9.495)
F5	Frequency of wetting/soiling yourself day/night: Never; 1/week; 1 or 2/week; >3/week	0.0336	(1.112,13.915)
F6*	When you are urinating into a toilet, can you shut off the flow of urine in the middle of your stream by using your muscles if you want to?	0.0478	(1.007,4.639)
F7	As you know people experience memory problems as they get older. What about remembering names?	0.2564	(0.792,2.401)
F8	Possibility of future incontinence	0.1516	(0.858,2.69)

*:Factor Extracted From Relieff Algorithm

Table 4.19- Baseline Factors- First and Third Follow-up Outcome Outliers Removed

Factor	*Description*	*P-value HH1HH2*	*P-value HH1HH4*
F1	Body Mass Index (BMI) <=24; >24	0.0330	0.0028
F2	Do you sneeze often, rarely or never?	0.0058	0.0249
F3	Any urine loss problems that started after deliveries: Yes;No	0.0234	0.0076
F4	Trouble getting to the bathroom on time: Yes;No	0.0354	0.0109
F5	Frequency of wetting/soiling yourself day/night: Never; 1/week; 1 or 2/week; >3/week	0.0021	0.0086
F6*	When you are urinating into a toilet, can you shut off the flow of urine in the middle of your stream by using your muscles if you want to?	0.0149	0.0032
F7	As you know people experience memory problems as they get older. What about remembering names?	0.0204	0.0093
F8	Possibility of future incontinence	0.0062	0.0077

Table 4.20- Regression Results with 2-way Interactions

Factor	*Description*	*P-value HH1HH2*	*P-value HH1HH4*
F1	Body Mass Index (BMI) <=24; >24	0.0132	0.0008
F2	Do you sneeze often, rarely or never?	0.0065	0.0130
F3	Any urine loss problems that started after deliveries: Yes;No	0.0311	0.0086
F4	Trouble getting to the bathroom on time: Yes;No	0.0429	0.0130
F5	Frequency of wetting/soiling yourself day/night: Never; 1/week; 1 or 2/week; >3/week	0.0008	0.0071
F6*	When you are urinating into a toilet, can you shut off the flow of urine in the middle of your stream by using your muscles if you want to?	0.0069	0.0013
F7	As you know people experience memory problems as they get older. What about remembering names?	0.0073	0.0020
F8	Possibility of future incontinence	0.0015	0.0109
F8*F2	Future Incontinence * Sneezing	0.0042	0.0892
F7*F1	Memory Problems * BMI	0.0245	0.0282

Table 4.21 shows the comparison between the Wald scores of HH1 and HH4 datasets with and without the interactions. The results show that both HH1 and HH4 with the retained outliers outperformed the linear model. When the outliers are removed from the data, the linear model outperforms the model with interactions.

4.6. Rule Extraction Results

In our experiment, we have also applied four different rule extraction methods to the simulation dataset and the results are compared when noise, incomplete data and multicollinearity are added to the data.

The first set of experiments was designed to evaluate the following rule extraction methods to see which one handles additive noise better: Apriori, JRip, PART and Prism. Table 4.22 and Table 4.23 show the results of those rule extraction methods when there are 5, 8 and 10% noise. Apriori algorithm outperforms the other rule extraction methods if the noise level is 5 or 10%. But if the noise level is beyond 10% Apriori algorithm lose its effectiveness.

Table 4.21- Comparison of Wald Scores with/without Interactions

		With Interaction		***Without Interaction***	
		Chi-Square	Pr > ChiSq	Chi-Square	Pr > ChiSq
HH1 with outliers	Wald	34.89	0.0001	33.00	<.0001
HH1 without outliers	Wald	37.05	<.0001	40.50	<.0001
HH4 with outliers	Wald	29.17	0.0012	28.79	.0003
HH4 without outliers	Wald	42.15	<.0001	44.94	<.0001

Table 4.22- Sensitivity of Rule Extraction Methods at Several Noise Levels

	0%	5%	8%	10%
Apriori	100 ± 0	100 ± 0	32 ± 10	0 ± 0
PART	55 ± 10	0 ± 0	0 ± 0	0 ± 0
Prism	90 ± 20	0 ± 0	0 ± 0	0 ± 0
JRip	40 ± 0	32 ± 10	36 ± 8	40 ± 0

Table 4.23- Specificity of Rule Extraction Methods at Several Noise Levels

	0%	5%	8%	10%
Apriori	100 ± 0	100 ± 0	96 ± 2	90 ± 0
PART	75 ± 10	28 ± 3	0 ± 0	0 ± 0
Prism	98 ± 5	3 ± 10	0 ± 0	0 ± 0
JRip	100 ± 0	96 ± 0	96 ± 1	96 ± 0

Table 4.24- Sensitivity of Rule Extraction Methods at Missing Value Levels

	0%	5%	8%	10%
Apriori	100 ± 0	100 ± 0	50 ± 17	0 ± 0
PART	55 ± 10	58 ± 26	48 ± 27	26 ± 20
Prism	90 ± 20	N/A	N/A	N/A
JRip	40 ± 0	40 ± 0	40 ± 0	40 ± 0

Table 4.25- Specificity of Rule Extraction Methods at Missing Value Levels

	0%	5%	8%	10%
Apriori	100 ± 0	100 ± 0	100 ± 17	100 ± 0
PART	75 ± 10	81 ± 5	79 ± 6	80 ± 5
Prism	98 ± 5	N/A	N/A	N/A
JRip	100 ± 0	99 ± 2	97 ± 2	96 ± 2

At that point the sensitivity and specificities of JRip is better than any other rule extraction method.

Table 4.24 and Table 4.25 show the results of the rule extraction methods when there are 5, 8 and 10% missing values. Apriori algorithm outperforms other rule extraction methods when there are 5 and 8% missing values. When the missing values exceeded 8% the reliability of Apriori decreases and PART gives better results.

4.7. Prediction Results

4.7.1. Constructed Predictive Index

In this section the most promising risk factors both for the first (HH2) and for the third follow up (HH4) outcome are extracted. A table is generated (Table 4.26) that determines the support and confidence for each combination of values the factors can take. The support threshold is calculated based on the margins of error formula. Note that, we are interested in the subjects that have become incontinent over time (C-I). However, the limited number of subjects that belong to the class ‘C-I’ prevent the technique from returning a support value that is greater than the support threshold as can be seen in Table 4.26 under the assumption of 90% confidence and 10% error rate. Table 4.26 shows that only 15% of the rules (10 out of 64) are having a support threshold that is less than or equal to the support of the rule. However, eight of those rules have a confidence of 0. One has 9% and the other one has 60%.

Table 4.26- All Possible Decision Rules For Five Promising Factors

F2	F3	F6	F4	F5	HH12	Antecedent	Support	Confidence	Support Threshold	Reliable (Y/N)
0	0	0	0	0	C-C	278	167	60.0719424	64.51156151	Y
0	0	0	0	1	C-C	7	3	42.8571429	65.8677551	N
0	0	0	1	0	C-C	20	12	60	64.5504	N
0	0	0	1	1	C-C	8	3	37.5	63.0375	N
0	0	1	0	0	C-C	48	23	47.9166667	67.12326389	N
0	0	1	0	1	C-C	0	0	#DIV/0!	#DIV/0!	NA
0	0	1	1	0	C-C	6	2	33.3333333	59.76888889	N
0	0	1	1	1	C-C	1	0	0	0	Y
0	1	0	0	0	C-C	8	5	62.5	63.0375	N
0	1	0	0	1	C-C	0	0	#DIV/0!	#DIV/0!	NA
0	1	0	1	0	C-C	0	0	#DIV/0!	#DIV/0!	NA
0	1	0	1	1	C-C	0	0	#DIV/0!	#DIV/0!	NA
0	1	1	0	0	C-C	3	1	33.3333333	59.76888889	N
0	1	1	0	1	C-C	0	0	#DIV/0!	#DIV/0!	NA
0	1	1	1	0	C-C	0	0	#DIV/0!	#DIV/0!	NA
0	1	1	1	1	C-C	0	0	#DIV/0!	#DIV/0!	NA
1	0	0	0	0	C-C	26	11	42.3076923	65.64852071	N
1	0	0	0	1	C-C	4	0	0	0	Y
1	0	0	1	0	C-C	0	0	#DIV/0!	#DIV/0!	NA
1	0	0	1	1	C-C	1	1	100	0	NA
1	0	1	0	0	C-C	4	2	50	67.24	NA
1	0	1	0	1	C-C	1	0	0	0	Y
1	0	1	1	0	C-C	0	0	#DIV/0!	#DIV/0!	NA
1	0	1	1	1	C-C	2	0	0	0	Y
1	1	0	0	0	C-C	1	0	0	0	Y
1	1	0	0	1	C-C	0	0	#DIV/0!	#DIV/0!	NA
1	1	0	1	0	C-C	0	0	#DIV/0!	#DIV/0!	NA
1	1	0	1	1	C-C	0	0	#DIV/0!	#DIV/0!	NA
1	1	1	0	0	C-C	0	0	#DIV/0!	#DIV/0!	NA
1	1	1	0	1	C-C	0	0	#DIV/0!	#DIV/0!	NA
1	1	1	1	0	C-C	0	0	#DIV/0!	#DIV/0!	NA
1	1	1	1	1	C-C	0	0	#DIV/0!	#DIV/0!	NA
0	0	0	0	0	C-I	278	26	9.35251799	22.80195021	Y
0	0	0	0	1	C-I	7	4	57.1428571	65.8677551	N
0	0	0	1	0	C-I	20	7	35	61.1884	N
0	0	0	1	1	C-I	8	2	25	50.43	N

Table 4.26- All Possible Decision Rules For Five Promising Factors- Continued

F2	F3	F6	F4	F5	HH12	Antecedent	Support	Confidence	Support Threshold	Reliable (Y/N)
0	0	1	0	0	C-I	48	11	22.9166667	47.51159722	N
0	0	1	0	1	C-I	0	0	#DIV/0!	#DIV/0!	NA
0	0	1	1	0	C-I	6	2	33.3333333	59.76888889	N
0	0	1	1	1	C-I	1	0	0	0	Y
0	1	0	0	0	C-I	8	3	37.5	63.0375	N
0	1	0	0	1	C-I	0	0	#DIV/0!	#DIV/0!	NA
0	1	0	1	0	C-I	0	0	#DIV/0!	#DIV/0!	NA
0	1	0	1	1	C-I	0	0	#DIV/0!	#DIV/0!	NA
0	1	1	0	0	C-I	3	1	33.3333333	59.76888889	N
0	1	1	0	1	C-I	0	0	#DIV/0!	#DIV/0!	NA
0	1	1	1	0	C-I	0	0	#DIV/0!	#DIV/0!	NA
0	1	1	1	1	C-I	0	0	#DIV/0!	#DIV/0!	NA
1	0	0	0	0	C-I	26	4	15.3846154	35.01254438	N
1	0	0	0	1	C-I	4	1	25	50.43	N
1	0	0	1	0	C-I	0	0	#DIV/0!	#DIV/0!	NA
1	0	0	1	1	C-I	1	0	0	0	Y
1	0	1	0	0	C-I	4	2	50	67.24	N
1	0	1	0	1	C-I	1	1	100	0	N
1	0	1	1	0	C-I	0	0	#DIV/0!	#DIV/0!	NA
1	0	1	1	1	C-I	2	1	50	67.24	N
1	1	0	0	0	C-I	1	1	100	0	Y
1	1	0	0	1	C-I	0	0	#DIV/0!	#DIV/0!	NA
1	1	0	1	0	C-I	0	0	#DIV/0!	#DIV/0!	NA
1	1	0	1	1	C-I	0	0	#DIV/0!	#DIV/0!	NA
1	1	1	0	0	C-I	0	0	#DIV/0!	#DIV/0!	NA
1	1	1	0	1	C-I	0	0	#DIV/0!	#DIV/0!	NA
1	1	1	1	0	C-I	0	0	#DIV/0!	#DIV/0!	NA
1	1	1	1	1	C-I	0	0	#DIV/0!	#DIV/0!	NA

This shows us that the only reliable rule is the following:

IF v41 =0 & v178 = 0 & v215 = 0 & v229 = 0 & v230 =0 THEN HH12 = C-C

This table is generated based on only five potential predictive factors.

4.7.1.1. MESA Data Predictive Index. The MESA Data is imputed and the outliers are removed in the data preprocessing section. The predictive index is determined based on the imputed data. Table 4.27 shows all the potential risk factors along with their odd ratios and confidence limits of the imputed data.

The predicted confidence interval for incontinence for a case with any of the 256 combinations of predictive factors, based on reduced HH1HH2 data set is listed in Table A.4 (non-imputed data) and A.5 (imputed data). For instance, a case classified as having a combination of factors F7 and F8 denoted as F78 is coded as 00000011. This case has a predicted probability of 58-77% as shown in Table A.5 (where 58% is the lower limit and 77% is the upper limit) of developing incontinence in the next year.

Table 4.27- Odd Ratio Estimates of Potential Predictive Factors of the Imputed Dataset

Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
F3	4.355	1.096	17.299
F4	2.694	1.139	6.373
F5	5.256	1.500	18.411
F6	2.456	1.134	5.319
F7	2.113	1.099	4.065
F8	2.408	1.283	4.519
F1	2.202	1.163	4.168
F2	2.008	1.057	3.813

CHAPTER FIVE

SUMMARY AND CONCLUSIONS

We have successfully proposed, implemented and validated an automated method for differentiating between different types of missing values and determining inconsistent data. An automated method is particularly important when dealing with large data of complex surveys since each subject should be processed individually. Incomplete data analyses are important because different types of incomplete data (SPMD, UMD, and GMD) cannot be treated the same, e.g. SPMD should not be treated by imputation techniques. Determining and eliminating the inconsistent data, on the other hand, partially eliminates noise.

The proposed method was validated using a simulation study. It was applied on MESA data as a preprocessing step to prepare the data for further analysis with better representation and quality. In this step, the baseline of this longitudinal survey data (MESA) was analyzed focusing solely on the female population. Proposed method is a preprocessing prerequisite for any data mining of clinical survey data.

The method can be applied on any questionnaire which would be convertible to a directed acyclic graph (DAG). The method scales well to high dimensional data as the time complexity is polynomial for the proposed algorithms. Future work include: a) partially estimating noise or respondents' reliability score using inconsistent data, and b) automating the conversion of the questionnaire to its corresponding directed acyclic graph.

Once SPMD missing data is determined, the data is stratified based on the most significant branching questions. This stratification process leads us to determine the different risk factors of each stratum and shows the diverse outcomes on different populations. We achieve two goals by taking advantage of incomplete data in the stratification process: (1) we utilize the wisdom of experts embedded in the data through the questionnaire design processes; (2) We treat a number of null answers without any estimation or imputation technique, preventing any unavoidable misinformation introduced by such methods.

The GMD portion of the data is then imputed using the rough set theory. A simulated data is used to validate the results. One possible future work of this approach is to decompose the mutually exclusive paths of UMD into GMD and SPMD, and using the rough set imputation on the GMD portion of UMD as well.

We have presented a comparison between different attribute selection and rule extraction techniques. The results show that, there is no single attribute selection or rule extraction technique that gives the best results. The advantage of each technique differs in different situations. The results of the attribute selection and rule extraction techniques show that, both are only reliable up to a certain percent of noise and incomplete data level.

The most promising potential predictive factors of the imputed data are then determined using logistic regression based on the original baseline and first follow-up data sets. The results are validated based on the original baseline and third follow-up data. Then a predictive index for urinary incontinence is constructed based on the predictive probabilities of all possible combinations of the predictive factors. The predictive index can be applied for immediate screening and for predicting future urinary incontinence in older woman of comparable demographics.

APPENDIX A

PREDICTIVE INDEX

Table A.1- Predictive Index of Non-Imputed Simulation Data

Contrast Estimation and Testing Results by Row						
Contrast	**Estimate**	**Standard Error**	**Confidence**	**Limits**	**Wald Chi-Square**	**Pr > ChiSq**
0	0.5	0	.	.	.	.
F8	0.63	0.0532	0.5213	0.727	5.4446	0.0196
F7	0.5961	0.0583	0.4788	0.7034	2.5871	0.1077
F78	0.7154	0.0663	0.5703	0.8264	8.0038	0.0047
F6	0.6296	0.0598	0.507	0.7374	4.2834	0.0385
F68	0.7432	0.0663	0.5944	0.8511	9.3651	0.0022
F67	0.7149	0.0725	0.5553	0.8344	6.6753	0.0098
F678	0.8103	0.0647	0.6518	0.9069	11.9028	0.0006
F5	0.7386	0.0948	0.5191	0.8809	4.4751	0.0344
F58	0.8279	0.0795	0.6173	0.9349	7.9353	0.0048
F57	0.8065	0.0876	0.5812	0.9261	6.4717	0.011
F578	0.8765	0.0667	0.6794	0.9596	10.0986	0.0015
F56	0.8276	0.0816	0.6101	0.9364	7.5236	0.0061
F568	0.891	0.0614	0.703	0.9658	11.0381	0.0009
F567	0.8763	0.0689	0.671	0.9609	9.5004	0.0021
F5678	0.9235	0.0486	0.7584	0.9789	13.1427	0.0003
F4	0.5837	0.0766	0.4305	0.7223	1.1492	0.2837
F48	0.7048	0.0807	0.5275	0.8362	5.0341	0.0249
F47	0.6742	0.0877	0.486	0.8191	3.3131	0.0687
F478	0.7789	0.0782	0.5913	0.8956	7.6942	0.0055
F46	0.7044	0.0868	0.5129	0.8435	4.3432	0.0372
F468	0.8023	0.0757	0.6144	0.9118	8.6217	0.0033
F467	0.7786	0.0838	0.5756	0.9011	6.6932	0.0097
F4678	0.8569	0.0655	0.6775	0.9447	11.2124	0.0008
F45	0.7984	0.0853	0.5835	0.918	6.7392	0.0094
F458	0.8709	0.0665	0.679	0.9556	10.406	0.0013
F457	0.8539	0.0745	0.6445	0.9496	8.7388	0.0031

F4578	0.9087	0.0538	0.7361	0.9726	12.5359	0.0004
F456	0.8707	0.069	0.6693	0.9572	9.6745	0.0019
F4568	0.9198	0.0494	0.7554	0.977	13.289	0.0003
F4567	0.9085	0.056	0.7262	0.9738	11.6144	0.0007
F45678	0.9442	0.038	0.8043	0.9858	15.3462	<.0001
F3	0.5218	0.1304	0.2815	0.7524	0.0279	0.8674
F38	0.6501	0.1228	0.3922	0.8426	1.318	0.251
F37	0.6169	0.1362	0.3422	0.8329	0.6828	0.4086
F378	0.7328	0.1152	0.464	0.8968	2.9412	0.0863
F36	0.6497	0.1266	0.3839	0.8466	1.2322	0.267
F368	0.7595	0.105	0.5058	0.9069	4.0009	0.0455
F367	0.7324	0.1194	0.4533	0.9003	2.7306	0.0984
F3678	0.8233	0.0906	0.579	0.9404	6.1106	0.0134
F35	0.7551	0.133	0.4296	0.9266	2.4508	0.1175
F358	0.84	0.1	0.5499	0.9575	4.9704	0.0258
F357	0.8198	0.1136	0.5018	0.9536	3.8786	0.0489
F3578	0.8857	0.0799	0.6225	0.9733	6.7251	0.0095
F356	0.8397	0.102	0.5427	0.9585	4.7787	0.0288
F3568	0.8992	0.071	0.6579	0.9764	7.8149	0.0052
F3567	0.8855	0.0818	0.6139	0.9741	6.4259	0.0112
F35678	0.9294	0.0543	0.7221	0.9852	9.6938	0.0018
F34	0.6047	0.1505	0.3081	0.8402	0.4558	0.4996
F348	0.7226	0.1289	0.4248	0.9019	2.2163	0.1366
F347	0.693	0.1439	0.3749	0.8947	1.4498	0.2286
F3478	0.7936	0.1123	0.5007	0.9365	3.8583	0.0495
F346	0.7222	0.1333	0.4142	0.9053	2.0682	0.1504
F3468	0.8158	0.1021	0.539	0.9437	4.7972	0.0285
F3467	0.7932	0.1165	0.4881	0.9392	3.5823	0.0584
F34678	0.8673	0.083	0.6138	0.9641	6.772	0.0093
F345	0.8121	0.1163	0.4924	0.9506	3.6859	0.0549
F3458	0.8804	0.0827	0.6124	0.9717	6.4646	0.011
F3457	0.8645	0.095	0.5655	0.969	5.2215	0.0223

F34578	0.9157	0.064	0.6812	0.9822	8.2685	0.004
F3456	0.8802	0.0848	0.603	0.9726	6.1492	0.0131
F34568	0.926	0.0567	0.7119	0.9844	9.3237	0.0023
F34567	0.9155	0.0659	0.6713	0.9829	7.8307	0.0051
F345678	0.9486	0.0425	0.77	0.9903	11.2031	0.0008
F2	0.6226	0.0571	0.5061	0.7264	4.243	0.0394
F28	0.7375	0.0667	0.5884	0.8466	8.9853	0.0027
F27	0.7088	0.0647	0.5684	0.8181	8.0546	0.0045
F278	0.8056	0.0613	0.658	0.8993	13.1766	0.0003
F26	0.7371	0.0664	0.5887	0.8459	9.0388	0.0026
F268	0.8268	0.0608	0.6751	0.9164	13.5658	0.0002
F267	0.8053	0.0626	0.6543	0.9004	12.662	0.0004
F2678	0.8757	0.0507	0.7386	0.9461	17.5394	<.0001
F25	0.8233	0.0778	0.6204	0.93	8.2858	0.004
F258	0.8881	0.0599	0.7088	0.9628	11.8047	0.0006
F257	0.873	0.0647	0.6866	0.9557	10.9192	0.001
F2578	0.9213	0.0466	0.7685	0.9764	14.6377	0.0001
F256	0.8879	0.0601	0.7079	0.9628	11.7366	0.0006
F2568	0.931	0.0429	0.7848	0.9803	15.1994	<.0001
F2567	0.9212	0.0472	0.7659	0.9766	14.3247	0.0002
F25678	0.9522	0.0321	0.8331	0.9876	17.9647	<.0001
F24	0.6981	0.0842	0.5138	0.835	4.4007	0.0359
F248	0.7975	0.0755	0.6118	0.9078	8.6015	0.0034
F247	0.7734	0.0785	0.5865	0.8914	7.5117	0.0061
F2478	0.8532	0.0632	0.6836	0.9399	12.154	0.0005
F246	0.7972	0.077	0.6071	0.909	8.2605	0.0041
F2468	0.87	0.0607	0.7003	0.9504	12.5321	0.0004
F2467	0.8529	0.0652	0.6768	0.9414	11.4396	0.0007
F24678	0.9081	0.0478	0.763	0.9681	16.0288	<.0001
F245	0.8672	0.0659	0.6804	0.9525	10.7636	0.001
F2458	0.9175	0.0481	0.762	0.9748	14.3702	0.0002
F2457	0.906	0.0527	0.7412	0.9701	13.3876	0.0003

F24578	0.9426	0.0365	0.8139	0.984	17.184	<.0001
F2456	0.9174	0.0488	0.7586	0.9751	13.9691	0.0002
F24568	0.9498	0.0335	0.8266	0.9868	17.4799	<.0001
F24567	0.9425	0.0373	0.8096	0.9844	16.5125	<.0001
F245678	0.9654	0.0247	0.8673	0.9917	20.198	<.0001
F23	0.6428	0.1315	0.3693	0.8469	1.0521	0.305
F238	0.754	0.1104	0.4885	0.9077	3.5439	0.0598
F237	0.7265	0.1206	0.4471	0.8971	2.5922	0.1074
F2378	0.8189	0.0925	0.5711	0.9389	5.8498	0.0156
F236	0.7536	0.111	0.4866	0.908	3.4969	0.0615
F2368	0.8389	0.084	0.6065	0.9462	7.0551	0.0079
F2367	0.8186	0.0939	0.5666	0.9397	5.6823	0.0171
F23678	0.8849	0.0663	0.6821	0.965	9.813	0.0017
F235	0.8356	0.1026	0.5404	0.9565	4.7395	0.0295
F2358	0.8965	0.072	0.6544	0.9754	7.7444	0.0054
F2357	0.8824	0.0813	0.6179	0.9721	6.6233	0.0101
F23578	0.9274	0.0544	0.724	0.9842	9.9489	0.0016
F2356	0.8963	0.0725	0.6522	0.9755	7.6535	0.0057
F23568	0.9364	0.0482	0.7509	0.9863	11.0519	0.0009
F23567	0.9273	0.055	0.7205	0.9844	9.7397	0.0018
F235678	0.956	0.0354	0.8068	0.9912	13.3926	0.0003
F234	0.7162	0.1368	0.4029	0.9042	1.8915	0.169
F2348	0.8112	0.1058	0.5258	0.9434	4.4497	0.0349
F2347	0.7883	0.1173	0.4842	0.9366	3.496	0.0615
F23478	0.8638	0.0844	0.6085	0.9628	6.6284	0.01
F2346	0.8109	0.1074	0.5207	0.9442	4.3201	0.0377
F23468	0.8796	0.0763	0.6404	0.9677	7.627	0.0057
F23467	0.8635	0.0862	0.6016	0.9637	6.3676	0.0116
F234678	0.9151	0.058	0.7137	0.979	10.1329	0.0015
F2345	0.877	0.0852	0.6027	0.971	6.1882	0.0129
F23458	0.9239	0.0574	0.7103	0.9836	9.3563	0.0022
F23457	0.9132	0.0654	0.6763	0.9815	8.1401	0.0043

F234578	0.9471	0.0424	0.7728	0.9895	11.5885	0.0007
F23456	0.9237	0.0582	0.706	0.9839	9.1286	0.0025
F234568	0.9538	0.0376	0.795	0.991	12.5981	0.0004
F234567	0.947	0.0432	0.7679	0.9898	11.2215	0.0008
F2345678	0.9682	0.0272	0.8432	0.9942	14.9198	0.0001
F1	0.5635	0.0556	0.4533	0.6678	1.2784	0.2582
F18	0.6874	0.0686	0.5405	0.8043	6.0967	0.0135
F17	0.6558	0.0774	0.4931	0.7886	3.5344	0.0601
F178	0.7644	0.0729	0.5948	0.8776	8.4556	0.0036
F16	0.6869	0.0721	0.5322	0.8089	5.495	0.0191
F168	0.7889	0.0679	0.6271	0.8925	10.4671	0.0012
F167	0.764	0.0768	0.5843	0.8818	7.6127	0.0058
F1678	0.8465	0.0624	0.6827	0.9339	12.6471	0.0004
F15	0.7848	0.0963	0.5438	0.9177	5.1445	0.0233
F158	0.8613	0.075	0.6447	0.9551	8.4628	0.0036
F157	0.8433	0.0843	0.6064	0.9495	6.9539	0.0084
F1578	0.9016	0.0609	0.7047	0.9724	10.4087	0.0013
F156	0.8611	0.0763	0.6396	0.9559	8.1718	0.0043
F1568	0.9135	0.0547	0.7311	0.9762	11.5993	0.0007
F1567	0.9014	0.0623	0.6983	0.9731	9.9654	0.0016
F15678	0.9397	0.0424	0.7826	0.9854	13.4949	0.0002
F14	0.6441	0.0916	0.4527	0.7984	2.2059	0.1375
F148	0.7551	0.0846	0.557	0.8832	6.0518	0.0139
F147	0.7276	0.0946	0.5118	0.8719	4.238	0.0395
F1478	0.8198	0.0772	0.6203	0.9268	8.4036	0.0037
F146	0.7547	0.0887	0.5461	0.8872	5.5025	0.019
F1468	0.8397	0.0715	0.6491	0.9369	9.7203	0.0018
F1467	0.8195	0.081	0.6081	0.93	7.6264	0.0058
F14678	0.8855	0.0599	0.7086	0.9609	12.0102	0.0005
F145	0.8364	0.0838	0.6064	0.9444	7.1055	0.0077
F1458	0.897	0.0615	0.7028	0.9698	10.5852	0.0011
F1457	0.883	0.07	0.6669	0.966	8.9113	0.0028

F14578	0.9278	0.0484	0.7573	0.9815	12.5002	0.0004
F1456	0.8968	0.0631	0.6954	0.9707	10.0513	0.0015
F14568	0.9367	0.0434	0.779	0.9842	13.5533	0.0002
F14567	0.9277	0.0498	0.7496	0.9821	11.8106	0.0006
F145678	0.9562	0.0329	0.824	0.9903	15.406	<.0001
F13	0.5848	0.1353	0.3209	0.8077	0.3781	0.5386
F138	0.7058	0.1188	0.4387	0.8805	2.3385	0.1262
F137	0.6752	0.1348	0.3839	0.874	1.4177	0.2338
F1378	0.7798	0.1073	0.51	0.9233	4.0968	0.043
F136	0.7054	0.1216	0.4319	0.8829	2.2244	0.1358
F1368	0.803	0.0952	0.5562	0.9299	5.4511	0.0196
F1367	0.7794	0.1104	0.5009	0.9256	3.8644	0.0493
F13678	0.8575	0.0799	0.6255	0.9559	7.5338	0.0061
F135	0.7992	0.1231	0.4696	0.9471	3.2446	0.0717
F1358	0.8714	0.0884	0.5907	0.9695	5.8802	0.0153
F1357	0.8545	0.1018	0.5413	0.9669	4.6741	0.0306
F13578	0.9091	0.0692	0.6597	0.981	7.5665	0.0059
F1356	0.8712	0.0898	0.5849	0.9701	5.7052	0.0169
F13568	0.9201	0.0605	0.6963	0.983	8.8082	0.003
F13567	0.9089	0.0705	0.6528	0.9815	7.2962	0.0069
F135678	0.9444	0.0457	0.7551	0.9894	10.5802	0.0011
F134	0.6639	0.1486	0.3488	0.8793	1.0452	0.3066
F1348	0.7708	0.1197	0.4713	0.927	3.2051	0.0734
F1347	0.7445	0.1361	0.4176	0.9222	2.2351	0.1349
F13478	0.8323	0.101	0.5459	0.9535	4.9044	0.0268
F1346	0.7705	0.123	0.462	0.9292	3.0313	0.0817
F13468	0.8511	0.0898	0.5876	0.9582	6.0482	0.0139
F13467	0.832	0.1041	0.5349	0.9552	4.6131	0.0317
F134678	0.894	0.0715	0.6578	0.9737	7.9883	0.0047
F1345	0.848	0.1048	0.5313	0.9649	4.469	0.0345
F13458	0.9048	0.0718	0.6496	0.9799	7.2929	0.0069
F13457	0.8917	0.0834	0.6025	0.9781	5.9621	0.0146

F134578	0.9334	0.0547	0.7141	0.9875	8.9996	0.0027
F13456	0.9046	0.0733	0.6419	0.9805	7.0052	0.0081
F134568	0.9417	0.0478	0.7454	0.9889	10.1959	0.0014
F134567	0.9333	0.056	0.7059	0.9879	8.6061	0.0034
F1345678	0.9597	0.0354	0.7979	0.9931	11.9538	0.0005
F12	0.6805	0.073	0.5244	0.8044	5.0639	0.0244
F128	0.7839	0.0701	0.6171	0.8908	9.6926	0.0019
F127	0.7586	0.0733	0.5893	0.8731	8.1926	0.0042
F1278	0.8426	0.0614	0.6837	0.9298	13.1519	0.0003
F126	0.7835	0.0693	0.619	0.8896	9.9056	0.0016
F1268	0.8604	0.0574	0.7073	0.9402	14.4876	0.0001
F1267	0.8423	0.0618	0.682	0.9301	12.9543	0.0003
F12678	0.9009	0.0467	0.7655	0.962	17.8243	<.0001
F125	0.8575	0.0746	0.645	0.9522	8.6311	0.0033
F1258	0.9111	0.0543	0.7336	0.9744	12.0477	0.0005
F1257	0.8988	0.0601	0.7089	0.97	10.944	0.0009
F12578	0.938	0.0415	0.7891	0.9839	14.5349	0.0001
F1256	0.9109	0.0543	0.7337	0.9743	12.071	0.0005
F12568	0.9457	0.0373	0.8076	0.9863	15.4925	<.0001
F12567	0.9378	0.0417	0.7876	0.984	14.3633	0.0002
F125678	0.9625	0.0276	0.8512	0.9914	17.938	<.0001
F124	0.7491	0.0887	0.5421	0.8827	5.3718	0.0205
F1248	0.8356	0.0727	0.6431	0.9348	9.4394	0.0021
F1247	0.815	0.0786	0.6134	0.9244	8.0977	0.0044
F12478	0.8824	0.0591	0.7107	0.9582	12.5156	0.0004
F1246	0.8354	0.0735	0.6404	0.9353	9.241	0.0024
F12468	0.8963	0.0547	0.7318	0.9647	13.4504	0.0002
F12467	0.8822	0.0603	0.706	0.9589	12.0362	0.0005
F124678	0.9273	0.0423	0.7888	0.9775	16.5073	<.0001
F1245	0.894	0.0617	0.7019	0.968	10.7312	0.0011
F12458	0.9349	0.043	0.7825	0.9829	14.2282	0.0002
F12457	0.9256	0.0481	0.7599	0.98	13.0295	0.0003

F124578	0.9549	0.0322	0.8304	0.9892	16.6887	<.0001
F12456	0.9348	0.0434	0.7803	0.983	13.9864	0.0002
F124568	0.9606	0.0289	0.845	0.9909	17.446	<.0001
F124567	0.9549	0.0326	0.8274	0.9894	16.2365	<.0001
F1245678	0.973	0.0211	0.8816	0.9943	19.8443	<.0001
F123	0.6991	0.1277	0.4142	0.8842	1.9293	0.1648
F1238	0.7983	0.1009	0.5367	0.9311	4.8154	0.0282
F1237	0.7742	0.1128	0.4919	0.9239	3.6477	0.0561
F12378	0.8538	0.0825	0.6154	0.9552	7.138	0.0075
F1236	0.7979	0.1011	0.536	0.931	4.7962	0.0285
F12368	0.8705	0.0731	0.6535	0.96	8.6349	0.0033
F12367	0.8535	0.0833	0.6123	0.9555	7.0036	0.0081
F123678	0.9085	0.0568	0.7224	0.9743	11.2915	0.0008
F1235	0.8678	0.0912	0.5803	0.9689	5.6058	0.0179
F12358	0.9179	0.0619	0.691	0.9824	8.6464	0.0033
F12357	0.9064	0.0707	0.6541	0.9802	7.4212	0.0064
F123578	0.9428	0.0461	0.755	0.9888	10.7239	0.0011
F12356	0.9177	0.0621	0.6898	0.9824	8.5941	0.0034
F123568	0.95	0.0404	0.7822	0.9901	12.0039	0.0005
F123567	0.9427	0.0465	0.7525	0.9889	10.5655	0.0012
F1235678	0.9656	0.0294	0.8318	0.9937	14.1845	0.0002
F1234	0.7651	0.1274	0.4482	0.9289	2.7768	0.0956
F12348	0.8473	0.0938	0.5725	0.9583	5.5832	0.0181
F12347	0.8278	0.1059	0.5285	0.9537	4.4679	0.0345
F123478	0.8912	0.0733	0.6506	0.973	7.7445	0.0054
F12346	0.847	0.0948	0.5688	0.9587	5.4669	0.0194
F123468	0.9041	0.065	0.6844	0.9762	8.9563	0.0028
F123467	0.8909	0.0745	0.6452	0.9735	7.5078	0.0061
F1234678	0.9329	0.0488	0.7508	0.9847	11.3722	0.0007
F12345	0.902	0.0743	0.6393	0.9795	6.9757	0.0083
F123458	0.94	0.0488	0.7422	0.9884	10.1291	0.0015
F123457	0.9314	0.0561	0.7084	0.987	8.8252	0.003

F1234578	0.9586	0.0357	0.7989	0.9926	12.2099	0.0005
F123456	0.9399	0.0492	0.7391	0.9886	9.9484	0.0016
F1234568	0.9638	0.0313	0.8212	0.9936	13.3926	0.0003
F1234567	0.9585	0.0362	0.7951	0.9928	11.9082	0.0006
F12345678	0.9752	0.0225	0.8637	0.9959	15.542	<.0001

Table A.2- Predictive Index of Imputed-Errors Retained Simulation Data

Contrast Estimation and Testing Results by Row

Contrast	**Estimate**	**Standard Error**	**Confidence**	**Limits**	**Wald Chi-Square**	**Pr > ChiSq**
0	0.5	0	.	.	.	.
F8	0.6047	0.0488	0.5062	0.6954	4.3294	0.0375
F7	0.602	0.051	0.4992	0.6965	3.7846	0.0517
F78	0.6982	0.0596	0.5706	0.8011	8.7909	0.003
F6	0.6517	0.0537	0.5407	0.7483	7.0228	0.008
F68	0.7411	0.0602	0.6075	0.8411	11.2337	0.0008
F67	0.7389	0.0601	0.6059	0.839	11.1653	0.0008
F678	0.8124	0.0555	0.6796	0.8983	16.201	<.0001
F5	0.7386	0.0828	0.5494	0.8675	5.8654	0.0154
F58	0.8121	0.0744	0.6243	0.9183	9.0069	0.0027
F57	0.8104	0.0757	0.6194	0.9182	8.6934	0.0032
F578	0.8673	0.0619	0.6948	0.9494	12.1688	0.0005
F56	0.8409	0.0668	0.6652	0.9336	11.1204	0.0009
F568	0.8899	0.054	0.7329	0.9597	14.3675	0.0002
F567	0.8888	0.0545	0.7308	0.9593	14.2232	0.0002
F5678	0.9244	0.0414	0.7928	0.9751	17.8283	<.0001
F4	0.5707	0.0713	0.4291	0.7015	0.9569	0.328
F48	0.6703	0.0809	0.4981	0.8064	3.7606	0.0525
F47	0.6678	0.0793	0.4994	0.8021	3.8142	0.0508
F478	0.7546	0.0764	0.5781	0.8735	7.4143	0.0065
F46	0.7132	0.0778	0.5412	0.8398	5.7322	0.0167
F468	0.7918	0.0728	0.6156	0.9004	9.1598	0.0025
F467	0.79	0.0711	0.6188	0.8971	9.5475	0.002
F4678	0.8519	0.0601	0.6934	0.9361	13.4895	0.0002
F45	0.7897	0.075	0.6079	0.901	8.59	0.0034
F458	0.8517	0.065	0.6767	0.9403	11.52	0.0007
F457	0.8503	0.065	0.6762	0.9392	11.5785	0.0007
F4578	0.8968	0.052	0.743	0.9631	14.8331	0.0001

F456	0.8754	0.057	0.716	0.9514	13.8975	0.0002
F4568	0.9149	0.0451	0.7754	0.971	16.7928	<.0001
F4567	0.914	0.0449	0.7763	0.9702	17.1314	<.0001
F45678	0.9421	0.0337	0.8289	0.982	20.3858	<.0001
F3	0.5522	0.1285	0.3082	0.7735	0.1629	0.6865
F38	0.6536	0.1218	0.3967	0.8441	1.3931	0.2379
F37	0.651	0.1278	0.3826	0.8489	1.2294	0.2675
F378	0.7405	0.1102	0.4813	0.8977	3.3461	0.0674
F36	0.6977	0.1154	0.4413	0.8708	2.3375	0.1263
F368	0.7792	0.0972	0.5384	0.9144	4.9832	0.0256
F367	0.7773	0.1011	0.5263	0.9164	4.5791	0.0324
F3678	0.8423	0.0791	0.6244	0.9449	7.9175	0.0049
F35	0.777	0.1171	0.481	0.9291	3.4138	0.0647
F358	0.842	0.0929	0.5755	0.9545	5.7387	0.0166
F357	0.8405	0.0963	0.5631	0.9557	5.3512	0.0207
F3578	0.8897	0.0722	0.6561	0.9715	8.055	0.0045
F356	0.867	0.0811	0.6214	0.9628	7.0983	0.0077
F3568	0.9089	0.0602	0.7061	0.9764	10.0279	0.0015
F3567	0.9079	0.062	0.6972	0.9769	9.5098	0.002
F35678	0.9378	0.0443	0.7731	0.9852	12.7817	0.0004
F34	0.6211	0.1434	0.3318	0.844	0.658	0.4173
F348	0.7149	0.1286	0.4213	0.8962	2.1229	0.1451
F347	0.7126	0.132	0.4121	0.8977	1.9843	0.1589
F3478	0.7914	0.109	0.5099	0.9326	4.0812	0.0434
F346	0.7541	0.118	0.4685	0.9143	3.1034	0.0781
F3468	0.8243	0.0952	0.5639	0.9445	5.5264	0.0187
F3467	0.8227	0.0973	0.5566	0.9449	5.2937	0.0214
F34678	0.8765	0.0739	0.6507	0.9643	8.2484	0.0041
F345	0.8224	0.1025	0.5391	0.9483	4.7654	0.029
F3458	0.8763	0.0791	0.6291	0.9673	7.2044	0.0073
F3457	0.8751	0.0811	0.6206	0.9678	6.879	0.0087
F34578	0.9146	0.0596	0.7058	0.9795	9.6487	0.0019

F3456	0.8965	0.0679	0.6736	0.9733	8.6982	0.0032
F34568	0.9298	0.0495	0.7497	0.9832	11.5977	0.0007
F34567	0.9291	0.0506	0.7442	0.9833	11.2284	0.0008
F345678	0.9525	0.0357	0.8103	0.9895	14.4386	0.0001
F2	0.6023	0.0528	0.4958	0.6999	3.5493	0.0596
F28	0.6985	0.0649	0.5587	0.8091	7.4279	0.0064
F27	0.6961	0.0623	0.5625	0.8032	7.9105	0.0049
F278	0.778	0.0614	0.6358	0.8755	12.4392	0.0004
F26	0.7392	0.0609	0.604	0.8404	10.866	0.001
F268	0.8125	0.0584	0.6715	0.9019	14.6192	0.0001
F267	0.8108	0.0561	0.6766	0.8978	15.8136	<.0001
F2678	0.8677	0.0479	0.7431	0.937	20.2864	<.0001
F25	0.8106	0.0712	0.633	0.9139	9.8355	0.0017
F258	0.8675	0.0601	0.7013	0.948	12.9034	0.0003
F257	0.8662	0.0599	0.7017	0.9468	13.0782	0.0003
F2578	0.9083	0.047	0.7661	0.9677	16.5055	<.0001
F256	0.889	0.0518	0.741	0.9573	15.6966	<.0001
F2568	0.9245	0.0404	0.7975	0.9744	18.7251	<.0001
F2567	0.9237	0.0401	0.7989	0.9736	19.2403	<.0001
F25678	0.9488	0.0298	0.8478	0.984	22.6708	<.0001
F24	0.6681	0.0815	0.4949	0.8053	3.6269	0.0569
F248	0.7549	0.0804	0.5677	0.8783	6.693	0.0097
F247	0.7528	0.077	0.575	0.8727	7.2382	0.0071
F2478	0.8233	0.068	0.6508	0.9209	10.8387	0.001
F246	0.7902	0.0724	0.6155	0.8986	9.2291	0.0024
F2468	0.8521	0.0624	0.6857	0.9383	12.4967	0.0004
F2467	0.8507	0.0599	0.6932	0.9349	13.5963	0.0002
F24678	0.8971	0.048	0.7587	0.9602	17.323	<.0001
F245	0.8505	0.0619	0.6868	0.9365	12.7724	0.0004
F2458	0.8969	0.0509	0.7474	0.9624	15.4552	<.0001
F2457	0.8959	0.0499	0.7509	0.9609	16.1714	<.0001
F24578	0.9294	0.0386	0.8061	0.9766	19.2119	<.0001

F2456	0.9141	0.0431	0.7838	0.969	18.5366	<.0001
F24568	0.9421	0.0331	0.832	0.9817	21.1078	<.0001
F24567	0.9415	0.0325	0.8353	0.9808	22.2301	<.0001
F245678	0.961	0.0239	0.8757	0.9885	25.191	<.0001
F23	0.6513	0.1282	0.3819	0.8496	1.2252	0.2683
F238	0.7407	0.1124	0.4757	0.9	3.2166	0.0729
F237	0.7386	0.1155	0.4665	0.9013	3.0132	0.0826
F2378	0.8121	0.0935	0.5654	0.9349	5.7114	0.0169
F236	0.7775	0.1013	0.5257	0.9168	4.5617	0.0327
F2368	0.8424	0.0805	0.6195	0.9461	7.6348	0.0057
F2367	0.8409	0.0823	0.6129	0.9464	7.3272	0.0068
F23678	0.8899	0.0617	0.7019	0.9652	11.0238	0.0009
F235	0.8407	0.0935	0.5732	0.954	5.6751	0.0172
F2358	0.8898	0.071	0.6615	0.9709	8.3272	0.0039
F2357	0.8887	0.0727	0.6539	0.9712	7.984	0.0047
F23578	0.9243	0.0528	0.7355	0.9817	10.99	0.0009
F2356	0.908	0.0603	0.7056	0.976	10.0515	0.0015
F23568	0.9379	0.0435	0.7773	0.9849	13.1927	0.0003
F23567	0.9372	0.0444	0.7727	0.985	12.8168	0.0003
F235678	0.9581	0.0311	0.8334	0.9905	16.3028	<.0001
F234	0.7129	0.1329	0.4101	0.8986	1.9607	0.1614
F2348	0.7916	0.1111	0.5037	0.9343	3.9279	0.0475
F2347	0.7897	0.1125	0.4988	0.9341	3.8133	0.0508
F23478	0.8517	0.0878	0.5951	0.9574	6.3175	0.012
F2346	0.8229	0.0979	0.5548	0.9454	5.2343	0.0221
F23468	0.8766	0.0752	0.6454	0.9652	7.9604	0.0048
F23467	0.8754	0.0759	0.6425	0.9649	7.8566	0.0051
F234678	0.9149	0.0556	0.7263	0.9775	11.0731	0.0009
F2345	0.8752	0.0792	0.6287	0.9667	7.2145	0.0072
F23458	0.9147	0.0589	0.7096	0.9792	9.8837	0.0017
F23457	0.9139	0.0598	0.7054	0.9792	9.6696	0.0019
F234578	0.942	0.0428	0.7775	0.9869	12.6549	0.0004

F23456	0.9292	0.0494	0.7506	0.9828	11.7436	0.0006
F234568	0.9526	0.0353	0.8131	0.9893	14.7744	0.0001
F234567	0.952	0.0357	0.8109	0.9892	14.6135	0.0001
F2345678	0.9681	0.0249	0.8623	0.9933	17.9596	<.0001
F1	0.5702	0.0505	0.4698	0.6652	1.8844	0.1698
F18	0.6699	0.0629	0.5375	0.7799	6.1925	0.0128
F17	0.6674	0.0687	0.5224	0.7864	5.0614	0.0245
F178	0.7543	0.0662	0.604	0.8607	9.8738	0.0017
F16	0.7128	0.0628	0.5764	0.8191	8.7903	0.003
F168	0.7915	0.0602	0.6499	0.8859	13.3521	0.0003
F167	0.7897	0.0629	0.6413	0.8875	12.217	0.0005
F1678	0.8517	0.0529	0.7164	0.9289	17.3987	<.0001
F15	0.7894	0.0842	0.5814	0.91	6.8044	0.0091
F158	0.8515	0.0701	0.6592	0.9444	9.9255	0.0016
F157	0.8501	0.0725	0.6503	0.9453	9.3059	0.0023
F1578	0.8966	0.0562	0.7256	0.966	12.7053	0.0004
F156	0.8752	0.0616	0.6991	0.9549	11.9408	0.0005
F1568	0.9147	0.0474	0.7653	0.9725	15.2493	<.0001
F1567	0.9139	0.0486	0.76	0.9726	14.6591	0.0001
F15678	0.942	0.0356	0.8189	0.9831	18.2707	<.0001
F14	0.6381	0.0854	0.4607	0.7845	2.3525	0.1251
F148	0.7295	0.0844	0.5385	0.8618	5.3862	0.0203
F147	0.7273	0.086	0.5326	0.8619	5.1134	0.0237
F1478	0.8032	0.0754	0.6158	0.9122	8.6974	0.0032
F146	0.7674	0.0784	0.5826	0.8864	7.3949	0.0065
F1468	0.8346	0.0676	0.6592	0.9294	10.9396	0.0009
F1467	0.8331	0.0681	0.6565	0.9288	10.7702	0.001
F14678	0.8842	0.0542	0.7302	0.9556	14.7632	0.0001
F145	0.8328	0.0745	0.6358	0.9343	9.0091	0.0027
F1458	0.884	0.0602	0.7069	0.9601	11.9735	0.0005
F1457	0.8829	0.0613	0.7023	0.9601	11.6148	0.0007
F14578	0.9202	0.0466	0.7688	0.9756	14.8489	0.0001

F1456	0.9031	0.0519	0.7446	0.9675	14.1702	0.0002
F14568	0.9345	0.0392	0.8024	0.9804	17.1934	<.0001
F14567	0.9338	0.0397	0.8002	0.9803	16.978	<.0001
F145678	0.9557	0.0289	0.8501	0.988	20.3047	<.0001
F13	0.6207	0.1294	0.3578	0.8277	0.8027	0.3703
F138	0.7145	0.115	0.4531	0.8832	2.6461	0.1038
F137	0.7122	0.1224	0.4343	0.8886	2.3032	0.1291
F1378	0.7911	0.1	0.5365	0.9253	4.8467	0.0277
F136	0.7538	0.1061	0.4995	0.9038	3.8284	0.0504
F1368	0.824	0.085	0.5974	0.9366	6.9327	0.0085
F1367	0.8224	0.0897	0.5815	0.9392	6.2262	0.0126
F13678	0.8763	0.0675	0.6766	0.96	9.8986	0.0017
F135	0.8222	0.1053	0.5298	0.9499	4.5176	0.0335
F1358	0.8761	0.0803	0.624	0.9679	6.9937	0.0082
F1357	0.8749	0.0839	0.6091	0.9691	6.445	0.0111
F13578	0.9145	0.061	0.6988	0.9801	9.2348	0.0024
F1356	0.8964	0.0691	0.6681	0.9738	8.4126	0.0037
F13568	0.9297	0.0499	0.7478	0.9833	11.4529	0.0007
F13567	0.929	0.0518	0.7372	0.9839	10.7139	0.0011
F135678	0.9524	0.0362	0.8068	0.9897	14.0456	0.0002
F134	0.685	0.1387	0.3816	0.8846	1.4611	0.2268
F1348	0.7689	0.1174	0.4767	0.924	3.3083	0.0689
F1347	0.7669	0.122	0.4634	0.9261	3.045	0.081
F13478	0.8342	0.096	0.5633	0.9515	5.4133	0.02
F1346	0.8027	0.1053	0.5249	0.9374	4.4524	0.0349
F13468	0.8616	0.0815	0.62	0.9596	7.163	0.0074
F13467	0.8602	0.0842	0.6093	0.9605	6.7303	0.0095
F134678	0.904	0.0619	0.6996	0.9744	9.8949	0.0017
F1345	0.86	0.0905	0.5847	0.9641	5.8318	0.0157
F13458	0.9038	0.0674	0.6729	0.9772	8.3533	0.0038
F13457	0.9029	0.0697	0.662	0.9778	7.8735	0.005
F134578	0.9343	0.0499	0.7431	0.9859	10.6742	0.0011

F13456	0.92	0.0572	0.7148	0.9814	9.8741	0.0017
F134568	0.9462	0.0407	0.7856	0.9883	12.8331	0.0003
F134567	0.9456	0.042	0.7785	0.9885	12.2502	0.0005
F1345678	0.9638	0.0291	0.8384	0.9927	15.4787	<.0001
F12	0.6677	0.0663	0.5281	0.783	5.4608	0.0194
F128	0.7545	0.0674	0.601	0.8625	9.523	0.002
F127	0.7525	0.0686	0.5962	0.8622	9.1079	0.0025
F1278	0.823	0.0604	0.6735	0.9129	13.7312	0.0002
F126	0.7899	0.0612	0.6459	0.8857	12.876	0.0003
F1268	0.8519	0.0536	0.7145	0.9297	16.9783	<.0001
F1267	0.8505	0.0537	0.713	0.9287	16.9244	<.0001
F12678	0.8969	0.0429	0.7778	0.9558	21.7026	<.0001
F125	0.8502	0.0681	0.6657	0.9418	10.5519	0.0012
F1258	0.8967	0.0541	0.7343	0.9646	13.6936	0.0002
F1257	0.8957	0.0549	0.7306	0.9645	13.3724	0.0003
F12578	0.9293	0.0413	0.7933	0.9782	16.8218	<.0001
F1256	0.914	0.046	0.7714	0.971	16.3113	<.0001
F12568	0.942	0.0345	0.825	0.9825	19.5027	<.0001
F12567	0.9414	0.0348	0.8235	0.9823	19.3676	<.0001
F125678	0.9609	0.0251	0.8688	0.9892	22.9125	<.0001
F124	0.7276	0.0853	0.5346	0.8613	5.2067	0.0225
F1248	0.8034	0.0768	0.6118	0.9137	8.3888	0.0038
F1247	0.8016	0.0763	0.6122	0.9118	8.4804	0.0036
F12478	0.8607	0.0627	0.6891	0.9452	12.1183	0.0005
F1246	0.8333	0.0676	0.6582	0.9284	10.9366	0.0009
F12468	0.8843	0.0549	0.7274	0.9563	14.3506	0.0002
F12467	0.8832	0.0543	0.7293	0.955	14.7707	0.0001
F124678	0.9204	0.0416	0.7919	0.9723	18.6275	<.0001
F1245	0.883	0.0582	0.7145	0.9579	12.8757	0.0003
F12458	0.9203	0.0452	0.7751	0.9748	15.7367	<.0001
F12457	0.9194	0.0454	0.7746	0.9743	15.7987	<.0001
F124578	0.9458	0.0336	0.8282	0.9844	18.9635	<.0001

F12456	0.9339	0.038	0.8088	0.9792	18.5397	<.0001
F124568	0.9557	0.0281	0.8544	0.9876	21.3654	<.0001
F124567	0.9553	0.0281	0.8549	0.9873	21.6997	<.0001
F1245678	0.9703	0.0201	0.8925	0.9923	24.8694	<.0001
F123	0.7125	0.121	0.4377	0.8875	2.3599	0.1245
F1238	0.7913	0.1004	0.5351	0.9258	4.8024	0.0284
F1237	0.7894	0.1047	0.5218	0.928	4.4028	0.0359
F12378	0.8515	0.081	0.6203	0.9527	7.4314	0.0064
F1236	0.8226	0.0887	0.5849	0.9385	6.3715	0.0116
F12368	0.8764	0.0677	0.6756	0.9602	9.8179	0.0017
F12367	0.8752	0.0702	0.6657	0.9611	9.1933	0.0024
F123678	0.9147	0.051	0.7487	0.9748	13.1757	0.0003
F1235	0.875	0.081	0.6213	0.9676	6.9091	0.0086
F12358	0.9146	0.0595	0.7062	0.9795	9.6747	0.0019
F12357	0.9137	0.0615	0.6964	0.98	9.1412	0.0025
F123578	0.9419	0.0436	0.7726	0.9872	12.2087	0.0005
F12356	0.9291	0.0501	0.747	0.9831	11.4497	0.0007
F123568	0.9525	0.0354	0.8122	0.9893	14.6873	0.0001
F123567	0.952	0.0364	0.8061	0.9895	14.0474	0.0002
F1235678	0.9681	0.0252	0.8602	0.9933	17.5868	<.0001
F1234	0.7671	0.1214	0.4651	0.9258	3.0767	0.0794
F12348	0.8344	0.0967	0.5609	0.9521	5.3348	0.0209
F12347	0.8329	0.0992	0.5522	0.9527	5.0821	0.0242
F123478	0.884	0.0746	0.6469	0.9694	7.7995	0.0052
F12346	0.8604	0.0838	0.611	0.9603	6.7978	0.0091
F123468	0.9041	0.0623	0.6977	0.9747	9.7624	0.0018
F123467	0.9031	0.0636	0.6918	0.9748	9.4407	0.0021
F1234678	0.9345	0.0454	0.7692	0.9839	12.8381	0.0003
F12345	0.903	0.0677	0.672	0.9769	8.344	0.0039
F123458	0.9344	0.0489	0.7487	0.9855	11.0743	0.0009
F123457	0.9337	0.0501	0.7423	0.9857	10.6698	0.0011
F1234578	0.9556	0.0352	0.8089	0.991	13.6745	0.0002

F123456	0.9457	0.0408	0.786	0.988	12.9476	0.0003
F1234568	0.9638	0.0286	0.8423	0.9925	16.0353	<.0001
F1234567	0.9634	0.0292	0.8387	0.9926	15.6109	<.0001
F12345678	0.9758	0.0201	0.8842	0.9953	18.979	<.0001

Table A.3- Predictive Index of Imputed-Errors Corrected Simulation Data

Contrast Estimation and Testing Results by Row

Contrast	Estimate	Standard Error	Confidence	Limits	Wald Chi-Square	Pr > ChiSq
0	0.5	0	.	.	.	.
F8	0.6208	0.048	0.5233	0.7093	5.8494	0.0156
F7	0.6109	0.0506	0.5085	0.7043	4.4962	0.034
F78	0.7199	0.058	0.594	0.8186	10.7869	0.001
F6	0.648	0.0552	0.534	0.7473	6.3689	0.0116
F68	0.7509	0.0596	0.6175	0.8491	11.992	0.0005
F67	0.743	0.0601	0.6092	0.8427	11.3565	0.0008
F678	0.8255	0.0535	0.6955	0.9074	17.5026	<.0001
F5	0.7325	0.0785	0.5552	0.8573	6.3161	0.012
F58	0.8176	0.0702	0.6406	0.9185	10.1653	0.0014
F57	0.8113	0.0723	0.6301	0.9156	9.5316	0.002
F578	0.8756	0.0576	0.714	0.952	13.6121	0.0002
F56	0.8345	0.067	0.6609	0.9288	11.1279	0.0009
F568	0.8919	0.0526	0.7392	0.9601	14.9789	0.0001
F567	0.8878	0.0539	0.7326	0.9581	14.6008	0.0001
F5678	0.9284	0.0394	0.8022	0.9764	18.6868	<.0001
F4	0.5824	0.0697	0.443	0.7098	1.3481	0.2456
F48	0.6954	0.0759	0.5308	0.8217	5.3115	0.0212
F47	0.6865	0.0755	0.524	0.8133	4.9871	0.0255
F478	0.7819	0.0693	0.6178	0.8883	9.8662	0.0017
F46	0.7197	0.0772	0.5482	0.8446	6.0796	0.0137
F468	0.8078	0.0684	0.6394	0.9088	10.631	0.0011
F467	0.8013	0.0681	0.6356	0.9031	10.6323	0.0011
F4678	0.8684	0.0544	0.7221	0.9437	15.7359	<.0001
F45	0.7925	0.0714	0.6199	0.8994	9.5294	0.002
F458	0.8621	0.0599	0.6996	0.9438	13.2349	0.0003
F457	0.8571	0.0608	0.6937	0.9407	13.0027	0.0003
F4578	0.9075	0.0466	0.7679	0.9668	16.9399	<.0001
F456	0.8755	0.0565	0.718	0.951	14.1616	0.0002
F4568	0.9201	0.0426	0.7872	0.9728	17.7937	<.0001
F4567	0.9169	0.0432	0.7839	0.9711	17.8899	<.0001
F45678	0.9476	0.0308	0.8427	0.9839	21.7558	<.0001
F3	0.5327	0.1236	0.3011	0.7511	0.0696	0.7919
F38	0.6511	0.1184	0.4019	0.8382	1.4333	0.2312

F37	0.6415	0.1256	0.3803	0.8392	1.136	0.2865
F378	0.7455	0.1071	0.492	0.8986	3.6222	0.057
F36	0.6773	0.1156	0.4266	0.8555	1.9632	0.1612
F368	0.7745	0.0967	0.5373	0.9104	4.9719	0.0258
F367	0.7672	0.1021	0.5179	0.91	4.3474	0.0371
F3678	0.8436	0.078	0.6288	0.945	8.1312	0.0044
F35	0.7574	0.1181	0.4697	0.9167	3.1373	0.0765
F358	0.8363	0.0926	0.5759	0.9506	5.8198	0.0158
F357	0.8305	0.0977	0.557	0.9503	5.242	0.022
F3578	0.8892	0.0712	0.6607	0.9706	8.3078	0.0039
F356	0.8518	0.0858	0.6025	0.9561	6.6126	0.0101
F3568	0.9039	0.0619	0.6994	0.9744	9.8895	0.0017
F3567	0.9002	0.0652	0.6851	0.974	9.1859	0.0024
F35678	0.9366	0.0448	0.7712	0.9848	12.7593	0.0004
F34	0.6139	0.1395	0.334	0.8345	0.6206	0.4308
F348	0.7224	0.123	0.4389	0.8965	2.4327	0.1188
F347	0.714	0.1282	0.4218	0.8952	2.1241	0.145
F3478	0.8034	0.1023	0.5345	0.9357	4.7244	0.0297
F346	0.7454	0.1179	0.4643	0.9082	2.992	0.0837
F3468	0.8274	0.0922	0.5749	0.9444	5.8927	0.0152
F3467	0.8213	0.0961	0.5602	0.9431	5.4285	0.0198
F34678	0.8827	0.0699	0.6672	0.9658	8.9426	0.0028
F345	0.8132	0.1029	0.5359	0.9426	4.7182	0.0298
F3458	0.877	0.0769	0.6382	0.9664	7.6003	0.0058
F3457	0.8724	0.0805	0.6235	0.9658	7.0597	0.0079
F34578	0.918	0.0568	0.7187	0.98	10.2684	0.0014
F3456	0.8891	0.0707	0.663	0.9703	8.433	0.0037
F34568	0.9292	0.0494	0.7507	0.9828	11.7494	0.0006
F34567	0.9264	0.0517	0.7403	0.9823	11.1759	0.0008
F345678	0.9537	0.0348	0.8149	0.9897	14.7631	0.0001
F2	0.6073	0.053	0.5	0.705	3.8435	0.0499
F28	0.7168	0.0636	0.5779	0.8239	8.7743	0.0031
F27	0.7082	0.0619	0.5744	0.8137	8.7626	0.0031
F278	0.7989	0.0589	0.6595	0.8907	14.1582	0.0002
F26	0.74	0.0616	0.6031	0.8421	10.6703	0.0011
F268	0.8233	0.0569	0.6841	0.9093	15.4895	<.0001
F267	0.8172	0.0556	0.6831	0.9026	16.1998	<.0001
F2678	0.8797	0.0454	0.7592	0.9444	21.4607	<.0001

F25	0.8089	0.0673	0.6432	0.9086	10.9722	0.0009
F258	0.8739	0.056	0.7192	0.9494	14.5333	0.0001
F257	0.8692	0.0565	0.7149	0.9463	14.5033	0.0001
F2578	0.9158	0.043	0.7847	0.9701	18.3042	<.0001
F256	0.8863	0.0513	0.7419	0.9548	16.2822	<.0001
F2568	0.9273	0.0386	0.8058	0.9752	19.7355	<.0001
F2567	0.9245	0.039	0.8037	0.9734	20.1034	<.0001
F25678	0.9525	0.0278	0.8573	0.9852	23.806	<.0001
F24	0.6832	0.0799	0.5113	0.8164	4.338	0.0373
F248	0.7793	0.0749	0.6005	0.8924	8.3879	0.0038
F247	0.772	0.0731	0.6001	0.8842	8.6327	0.0033
F2478	0.8472	0.061	0.6878	0.9331	13.23	0.0003
F246	0.7988	0.071	0.6256	0.9042	9.7515	0.0018
F2468	0.8667	0.0578	0.7093	0.9454	14.0146	0.0002
F2467	0.8618	0.0567	0.7103	0.9406	14.7792	0.0001
F24678	0.9107	0.0428	0.7842	0.9663	19.4495	<.0001
F245	0.8552	0.0584	0.701	0.937	14.1928	0.0002
F2458	0.9063	0.0461	0.7693	0.9656	17.4511	<.0001
F2457	0.9026	0.0461	0.7682	0.9629	17.9955	<.0001
F24578	0.9382	0.034	0.8278	0.9796	21.4947	<.0001
F2456	0.9158	0.042	0.7889	0.9694	19.169	<.0001
F24568	0.9468	0.0307	0.8436	0.9833	22.3272	<.0001
F24567	0.9447	0.0307	0.8435	0.9818	23.2743	<.0001
F245678	0.9655	0.0215	0.8876	0.99	26.6637	<.0001
F23	0.638	0.1253	0.3783	0.8362	1.0907	0.2963
F238	0.7426	0.1093	0.4848	0.8984	3.4357	0.0638
F237	0.7345	0.1139	0.4683	0.8969	3.0359	0.0814
F2378	0.8192	0.09	0.5792	0.9371	6.1788	0.0129
F236	0.7644	0.1023	0.516	0.9081	4.2958	0.0382
F2368	0.8416	0.0796	0.6225	0.9448	7.8269	0.0051
F2367	0.8359	0.0828	0.6093	0.9433	7.268	0.007
F23678	0.8929	0.06	0.7089	0.9662	11.4089	0.0007
F235	0.8284	0.0944	0.5678	0.9466	5.6223	0.0177
F2358	0.8877	0.07	0.6664	0.969	8.6799	0.0032
F2357	0.8834	0.0732	0.6531	0.9683	8.1266	0.0044
F23578	0.9254	0.0513	0.743	0.9816	11.4863	0.0007
F2356	0.8989	0.0633	0.6942	0.9721	9.8434	0.0017
F23568	0.9357	0.0441	0.7756	0.9839	13.3297	0.0003

F23567	0.9331	0.046	0.7667	0.9834	12.7635	0.0004
F235678	0.958	0.0309	0.8347	0.9904	16.504	<.0001
F234	0.7109	0.1299	0.4161	0.8945	2.0275	0.1545
F2348	0.801	0.1053	0.5245	0.9362	4.4461	0.035
F2347	0.7942	0.1086	0.512	0.9342	4.1304	0.0421
F23478	0.8634	0.0813	0.6209	0.9606	7.1634	0.0074
F2346	0.819	0.0975	0.5549	0.9426	5.2673	0.0217
F23468	0.8811	0.0719	0.6587	0.966	8.5117	0.0035
F23467	0.8766	0.0742	0.6493	0.9646	8.1676	0.0043
F234678	0.9208	0.0518	0.7429	0.9791	11.9221	0.0006
F2345	0.8707	0.0789	0.6303	0.9637	7.4063	0.0065
F23458	0.9168	0.0564	0.7211	0.9792	10.5255	0.0012
F23457	0.9136	0.0586	0.7117	0.9784	10.0956	0.0015
F234578	0.9454	0.0401	0.7907	0.9875	13.4843	0.0002
F23456	0.9253	0.0508	0.7458	0.9813	11.7191	0.0006
F234568	0.953	0.0346	0.8167	0.9893	15.1434	<.0001
F234567	0.9511	0.0359	0.8108	0.9888	14.7852	0.0001
F2345678	0.9696	0.0238	0.8678	0.9936	18.4482	<.0001
F1	0.5826	0.0505	0.4816	0.6771	2.5746	0.1086
F18	0.6955	0.0615	0.5638	0.8015	8.0862	0.0045
F17	0.6866	0.0663	0.545	0.8003	6.4838	0.0109
F178	0.782	0.0619	0.6377	0.8796	12.3615	0.0004
F16	0.7198	0.0622	0.584	0.8246	9.3692	0.0022
F168	0.8079	0.0576	0.6702	0.897	14.9779	0.0001
F167	0.8013	0.06	0.6582	0.8942	13.6756	0.0002
F1678	0.8685	0.0486	0.7416	0.9382	19.7073	<.0001
F15	0.7926	0.0798	0.5962	0.9082	7.6317	0.0057
F158	0.8622	0.0646	0.6829	0.9478	11.3578	0.0008
F157	0.8571	0.0676	0.6704	0.9465	10.5432	0.0012
F1578	0.9076	0.0504	0.7516	0.9696	14.4666	0.0001
F156	0.8756	0.06	0.705	0.9539	12.5448	0.0004
F1568	0.9201	0.0444	0.7789	0.9741	16.3463	<.0001
F1567	0.917	0.0462	0.7709	0.9732	15.6851	<.0001
F15678	0.9476	0.0324	0.8343	0.9848	19.6901	<.0001
F14	0.6606	0.084	0.4829	0.8023	3.1597	0.0755
F148	0.7611	0.0785	0.5776	0.8813	7.2083	0.0073
F147	0.7535	0.0805	0.5665	0.8773	6.6402	0.01
F1478	0.8334	0.0665	0.6616	0.9275	11.2854	0.0008

F146	0.7818	0.0759	0.5997	0.8955	8.2298	0.0041
F1468	0.8544	0.0616	0.6896	0.9394	12.7582	0.0004
F1467	0.8491	0.0629	0.6825	0.9364	12.3844	0.0004
F14678	0.9021	0.0471	0.7641	0.9632	17.3482	<.0001
F145	0.842	0.0702	0.6546	0.9375	10.0579	0.0015
F1458	0.8972	0.0541	0.7345	0.9649	13.6595	0.0002
F1457	0.8933	0.0559	0.7264	0.9635	13.1514	0.0003
F14578	0.932	0.0403	0.7976	0.9794	16.9507	<.0001
F1456	0.9075	0.0498	0.7542	0.9691	14.817	0.0001
F14568	0.9414	0.0357	0.8188	0.9828	18.4177	<.0001
F14567	0.9391	0.0367	0.814	0.9819	18.1435	<.0001
F145678	0.9619	0.0253	0.8673	0.9898	21.9487	<.0001
F13	0.614	0.1267	0.3582	0.8193	0.7546	0.385
F138	0.7225	0.1117	0.4664	0.8858	2.9521	0.0858
F137	0.7141	0.1197	0.4418	0.8874	2.4372	0.1185
F1378	0.8035	0.0953	0.5561	0.9303	5.4439	0.0196
F136	0.7455	0.1061	0.4948	0.8975	3.6963	0.0545
F1368	0.8274	0.0831	0.6052	0.9375	7.257	0.0071
F1367	0.8214	0.0888	0.5842	0.9377	6.3597	0.0117
F13678	0.8827	0.0645	0.6895	0.9623	10.5021	0.0012
F135	0.8133	0.1056	0.5271	0.9445	4.4763	0.0344
F1358	0.877	0.0784	0.632	0.9673	7.3132	0.0068
F1357	0.8724	0.0832	0.6123	0.9673	6.6111	0.0101
F13578	0.918	0.0583	0.7107	0.9808	9.7429	0.0018
F1356	0.8891	0.0715	0.6592	0.9708	8.2311	0.0041
F13568	0.9292	0.0497	0.7489	0.983	11.591	0.0007
F13567	0.9264	0.0527	0.7346	0.9828	10.7397	0.001
F135678	0.9537	0.0353	0.8115	0.99	14.3374	0.0002
F134	0.6893	0.1356	0.3909	0.8847	1.5853	0.208
F1348	0.7841	0.111	0.5011	0.9293	3.8687	0.0492
F1347	0.777	0.1168	0.4816	0.9289	3.4262	0.0642
F13478	0.8508	0.0879	0.5947	0.9568	6.3168	0.012
F1346	0.8034	0.104	0.5293	0.9369	4.5725	0.0325
F13468	0.8699	0.0771	0.6376	0.9621	7.7823	0.0053
F13467	0.8651	0.0811	0.6217	0.9616	7.1572	0.0075
F134678	0.913	0.0567	0.7217	0.977	10.8576	0.001
F1345	0.8587	0.0895	0.5886	0.9627	5.9788	0.0145
F13458	0.9087	0.0639	0.6874	0.9783	8.8977	0.0029

F13457	0.9051	0.0674	0.6721	0.978	8.262	0.004
F134578	0.9398	0.046	0.7606	0.9871	11.4427	0.0007
F13456	0.918	0.058	0.7121	0.9806	9.8324	0.0017
F134568	0.9482	0.0394	0.7918	0.9888	13.1469	0.0003
F134567	0.9461	0.0414	0.7814	0.9885	12.4502	0.0004
F1345678	0.9664	0.0273	0.8472	0.9933	15.9898	<.0001
F12	0.6833	0.066	0.5427	0.7969	6.3568	0.0117
F128	0.7794	0.0647	0.6281	0.8807	11.2421	0.0008
F127	0.7721	0.0658	0.6194	0.8758	10.643	0.0011
F1278	0.8472	0.0556	0.7051	0.9279	15.9212	<.0001
F126	0.7989	0.06	0.6564	0.892	13.63	0.0002
F1268	0.8667	0.0503	0.7347	0.9385	18.4806	<.0001
F1267	0.8618	0.0508	0.73	0.935	18.4207	<.0001
F12678	0.9108	0.0388	0.8003	0.963	23.6964	<.0001
F125	0.8553	0.0637	0.6829	0.9419	11.9016	0.0006
F1258	0.9063	0.049	0.7574	0.9677	15.466	<.0001
F1257	0.9027	0.0504	0.7508	0.9662	15.0713	0.0001
F12578	0.9382	0.0363	0.8164	0.9811	18.8372	<.0001
F1256	0.9158	0.0441	0.7798	0.9709	17.3826	<.0001
F12568	0.9468	0.0317	0.8382	0.9839	20.8867	<.0001
F12567	0.9447	0.0325	0.8345	0.983	20.796	<.0001
F125678	0.9655	0.0224	0.8821	0.9905	24.5235	<.0001
F124	0.7506	0.0826	0.5589	0.8773	6.2288	0.0126
F1248	0.8313	0.0699	0.6498	0.929	10.247	0.0014
F1247	0.8253	0.0703	0.6451	0.9247	10.1521	0.0014
F12478	0.8855	0.0543	0.7304	0.9567	14.6015	0.0001
F1246	0.8471	0.0644	0.6765	0.9362	11.8624	0.0006
F12468	0.9007	0.0491	0.7555	0.9638	16.1071	<.0001
F12467	0.8969	0.0494	0.7533	0.9612	16.407	<.0001
F124678	0.9344	0.0355	0.8206	0.9779	21.0106	<.0001
F1245	0.8918	0.054	0.7337	0.961	14.2285	0.0002
F12458	0.931	0.0399	0.7999	0.9785	17.58	<.0001
F12457	0.9283	0.0406	0.7964	0.9772	17.5959	<.0001
F124578	0.9549	0.0286	0.8522	0.9873	21.1459	<.0001
F12456	0.9382	0.0358	0.8191	0.9807	19.4167	<.0001
F124568	0.9613	0.0251	0.8688	0.9894	22.6878	<.0001
F124567	0.9597	0.0255	0.8672	0.9886	23.0447	<.0001
F1245678	0.975	0.0173	0.9063	0.9937	26.524	<.0001

F123	0.711	0.1188	0.4421	0.8842	2.425	0.1194
F1238	0.8011	0.0964	0.5516	0.9295	5.2998	0.0213
F1237	0.7943	0.1017	0.5328	0.929	4.7136	0.0299
F12378	0.8634	0.076	0.6413	0.9572	8.1908	0.0042
F1236	0.8191	0.0883	0.5846	0.9358	6.417	0.0113
F12368	0.8811	0.0653	0.6859	0.9618	10.3163	0.0013
F12367	0.8767	0.0687	0.6716	0.9611	9.5186	0.002
F123678	0.9209	0.048	0.762	0.9769	13.8955	0.0002
F1235	0.8707	0.0805	0.6236	0.9648	7.1062	0.0077
F12358	0.9168	0.0572	0.7169	0.9796	10.2293	0.0014
F12357	0.9136	0.0602	0.7032	0.9793	9.5502	0.002
F123578	0.9454	0.041	0.7852	0.9879	12.9154	0.0003
F12356	0.9254	0.0512	0.7436	0.9815	11.5372	0.0007
F123568	0.9531	0.0347	0.816	0.9894	15.0525	0.0001
F123567	0.9511	0.0365	0.8071	0.9891	14.3231	0.0002
F1235678	0.9696	0.024	0.8658	0.9937	18.0502	<.0001
F1234	0.7743	0.1177	0.4781	0.9278	3.3494	0.0672
F12348	0.8489	0.0899	0.5874	0.9568	6.0718	0.0137
F12347	0.8434	0.0936	0.573	0.9558	5.6395	0.0176
F123478	0.8981	0.0669	0.6776	0.9737	8.8517	0.0029
F12346	0.8633	0.0818	0.6187	0.9609	7.0646	0.0079
F123468	0.9118	0.0579	0.7159	0.977	10.5138	0.0012
F123467	0.9084	0.0603	0.7055	0.9762	10.0188	0.0015
F1234678	0.942	0.0409	0.7894	0.986	13.8878	0.0002
F12345	0.9038	0.0661	0.6794	0.9766	8.6923	0.0032
F123458	0.939	0.0456	0.7638	0.9865	11.7959	0.0006
F123457	0.9365	0.0477	0.7538	0.9861	11.2537	0.0008
F1234578	0.9602	0.0318	0.8249	0.992	14.5815	0.0001
F123456	0.9454	0.0407	0.787	0.9878	13.0956	0.0003
F1234568	0.9659	0.0271	0.8494	0.993	16.4912	<.0001
F1234567	0.9645	0.0283	0.8433	0.9928	15.9788	<.0001
F12345678	0.978	0.0185	0.8922	0.9958	19.5793	<.0001

Table A.4- Predictive Index of Non-Imputed MESA Data

Contrast Estimation and Testing Results by Row						
Contrast	**Estimate**	**Standard Error**	**Confidence**	**Limits**	**Wald Chi-Square**	**Pr > ChiSq**
0	0.5	0	.	.	.	.
F8	0.6193	0.0419	0.5344	0.6974	7.4874	0.0062
F7	0.6065	0.0445	0.5168	0.6896	5.3799	0.0204
F78	0.7148	0.0526	0.602	0.806	12.706	0.0004
F6	0.6294	0.0508	0.5258	0.7224	5.9239	0.0149
F68	0.7342	0.0539	0.6166	0.826	13.5457	0.0002
F67	0.7236	0.0559	0.6022	0.8191	11.8594	0.0006
F678	0.8098	0.0504	0.6916	0.8899	19.61	<.0001
F5	0.7248	0.0628	0.587	0.83	9.4715	0.0021
F58	0.8108	0.057	0.674	0.8987	15.3244	<.0001
F57	0.8024	0.0613	0.6556	0.8965	13.1413	0.0003
F578	0.8685	0.0496	0.7382	0.9393	18.8936	<.0001
F56	0.8173	0.0597	0.6714	0.9074	14.035	0.0002
F568	0.8792	0.0471	0.7533	0.9455	20.0663	<.0001
F567	0.8734	0.0502	0.739	0.9438	18.0786	<.0001
F5678	0.9181	0.037	0.8103	0.9672	24.0795	<.0001
F4	0.6242	0.0566	0.5087	0.7272	4.4249	0.0354
F48	0.7299	0.062	0.5933	0.8335	9.9903	0.0016
F47	0.7191	0.0606	0.5872	0.8217	9.8232	0.0017
F478	0.8064	0.0565	0.6721	0.8943	15.5662	<.0001
F46	0.7383	0.063	0.5982	0.8425	10.1079	0.0015
F468	0.8211	0.0558	0.6855	0.9062	16.0805	<.0001
F467	0.8131	0.0557	0.6796	0.8992	16.1067	<.0001
F4678	0.8761	0.045	0.7582	0.941	22.2306	<.0001
F45	0.814	0.0534	0.6869	0.8972	17.5509	<.0001
F458	0.8768	0.0448	0.7593	0.9414	22.3516	<.0001
F457	0.8709	0.0466	0.7497	0.9382	21.2466	<.0001
F4578	0.9165	0.0359	0.814	0.9649	26.1082	<.0001
F456	0.8814	0.0451	0.7615	0.9454	21.6417	<.0001
F4568	0.9236	0.0339	0.8249	0.9688	26.8502	<.0001
F4567	0.9197	0.0353	0.8176	0.967	25.9463	<.0001
F45678	0.9491	0.0253	0.8696	0.9811	31.1451	<.0001
F3	0.6956	0.0772	0.5279	0.8236	5.1372	0.0234
F38	0.788	0.0696	0.6216	0.8937	9.9367	0.0016
F37	0.7788	0.073	0.6055	0.8899	8.8264	0.003

F378	0.8514	0.0594	0.6954	0.9349	13.8364	0.0002
F36	0.7951	0.0703	0.6248	0.9005	9.8652	0.0017
F368	0.8632	0.0559	0.7139	0.9411	15.1436	<.0001
F367	0.8568	0.0587	0.7007	0.9386	13.9719	0.0002
F3678	0.9068	0.0437	0.7795	0.964	19.3977	<.0001
F35	0.8575	0.0612	0.6928	0.9414	12.8489	0.0003
F358	0.9073	0.046	0.7703	0.9662	17.4212	<.0001
F357	0.9027	0.0491	0.756	0.9652	15.855	<.0001
F3578	0.9378	0.035	0.8229	0.98	20.4005	<.0001
F356	0.9109	0.0458	0.772	0.9686	17.0066	<.0001
F3568	0.9433	0.0322	0.8362	0.9819	21.7676	<.0001
F3567	0.9403	0.0344	0.8257	0.9813	20.2245	<.0001
F35678	0.9624	0.0235	0.8778	0.9892	24.9766	<.0001
F34	0.7915	0.0754	0.6079	0.9029	8.5227	0.0035
F348	0.8606	0.061	0.6951	0.9436	12.8256	0.0003
F347	0.854	0.0626	0.6861	0.94	12.3638	0.0004
F3478	0.9049	0.0473	0.7642	0.9654	16.8033	<.0001
F346	0.8657	0.0596	0.7023	0.9463	13.2	0.0003
F3468	0.9129	0.0443	0.7788	0.969	17.8179	<.0001
F3467	0.9086	0.0457	0.7717	0.9669	17.4206	<.0001
F34678	0.9417	0.0324	0.8353	0.981	22.149	<.0001
F345	0.9091	0.0448	0.7757	0.9666	18.0644	<.0001
F3458	0.9421	0.0323	0.836	0.9811	22.1879	<.0001
F3457	0.9391	0.034	0.8279	0.9801	21.2122	<.0001
F34578	0.9616	0.0236	0.8773	0.9888	25.3233	<.0001
F3456	0.9444	0.0316	0.8394	0.9822	22.1983	<.0001
F34568	0.9651	0.0217	0.8864	0.9899	26.4944	<.0001
F34567	0.9632	0.0229	0.8808	0.9893	25.5937	<.0001
F345678	0.977	0.0154	0.9173	0.9939	29.8834	<.0001
F2	0.6302	0.045	0.5385	0.7134	7.6086	0.0058
F28	0.7348	0.0538	0.6172	0.8265	13.6043	0.0002
F27	0.7242	0.0535	0.6084	0.8162	12.9848	0.0003
F278	0.8103	0.0512	0.6898	0.8914	19.0041	<.0001
F26	0.7432	0.055	0.622	0.8358	13.6001	0.0002
F268	0.8248	0.0499	0.7052	0.9026	20.1078	<.0001
F267	0.8169	0.0504	0.6975	0.8962	19.7105	<.0001
F2678	0.8789	0.0412	0.7726	0.9394	26.211	<.0001
F25	0.8178	0.0553	0.6845	0.9028	16.3938	<.0001
F258	0.8795	0.0455	0.759	0.9442	21.4692	<.0001

F257	0.8737	0.0478	0.7475	0.9417	19.945	<.0001
F2578	0.9184	0.0363	0.8134	0.9667	25.018	<.0001
F256	0.884	0.0455	0.7617	0.9479	20.9693	<.0001
F2568	0.9254	0.0339	0.8258	0.9701	26.3488	<.0001
F2567	0.9216	0.0356	0.8173	0.9686	25.0098	<.0001
F25678	0.9503	0.0253	0.87	0.982	30.3684	<.0001
F24	0.739	0.0611	0.6033	0.8405	10.7794	0.001
F248	0.8216	0.0565	0.6838	0.9074	15.6939	<.0001
F247	0.8135	0.0551	0.6817	0.8989	16.4546	<.0001
F2478	0.8765	0.0459	0.7554	0.9422	21.3214	<.0001
F246	0.8278	0.0546	0.694	0.9107	16.773	<.0001
F2468	0.8866	0.0441	0.768	0.9487	21.9869	<.0001
F2467	0.8811	0.0437	0.7658	0.9438	23.004	<.0001
F24678	0.9234	0.0332	0.8278	0.968	28.1859	<.0001
F245	0.8818	0.0427	0.7698	0.9433	24.0953	<.0001
F2458	0.9238	0.0332	0.8277	0.9684	27.9091	<.0001
F2457	0.92	0.0341	0.8226	0.9661	27.7908	<.0001
F24578	0.9492	0.0251	0.8709	0.9811	31.7279	<.0001
F2456	0.9268	0.0324	0.8325	0.9699	28.3178	<.0001
F24568	0.9537	0.0234	0.8795	0.9831	32.6288	<.0001
F24567	0.9513	0.0241	0.8757	0.9819	32.6342	<.0001
F245678	0.9695	0.0168	0.9124	0.9898	36.9586	<.0001
F23	0.7956	0.0667	0.6353	0.8969	10.9729	0.0009
F238	0.8636	0.0548	0.7179	0.9403	15.7464	<.0001
F237	0.8572	0.0567	0.7076	0.937	14.9542	0.0001
F2378	0.9071	0.0432	0.7814	0.9638	19.7545	<.0001
F236	0.8687	0.0535	0.725	0.9432	16.2048	<.0001
F2368	0.9149	0.0401	0.7967	0.9672	21.2595	<.0001
F2367	0.9107	0.0417	0.7888	0.9653	20.5356	<.0001
F23678	0.9431	0.0298	0.8482	0.9801	25.6202	<.0001
F235	0.9112	0.0434	0.782	0.967	18.8583	<.0001
F2358	0.9434	0.0312	0.8411	0.9813	23.104	<.0001
F2357	0.9405	0.0331	0.8323	0.9805	21.819	<.0001
F23578	0.9626	0.0229	0.8808	0.9889	26.044	<.0001
F2356	0.9457	0.0305	0.8447	0.9824	23.152	<.0001
F23568	0.9659	0.021	0.8905	0.99	27.5825	<.0001
F23567	0.9641	0.0222	0.8844	0.9895	26.3692	<.0001
F235678	0.9776	0.0149	0.92	0.994	30.7834	<.0001
F234	0.8661	0.0579	0.7086	0.9451	13.9895	0.0002

F2348	0.9132	0.0439	0.7805	0.9689	18.0884	<.0001
F2347	0.9088	0.0448	0.7755	0.9664	18.0812	<.0001
F23478	0.9419	0.0324	0.8358	0.981	22.2033	<.0001
F2346	0.9166	0.042	0.7892	0.9699	19.0308	<.0001
F23468	0.947	0.0299	0.8473	0.9829	23.3414	<.0001
F23467	0.9442	0.0307	0.8435	0.9815	23.4701	<.0001
F234678	0.965	0.0213	0.8893	0.9895	27.8018	<.0001
F2345	0.9446	0.0303	0.8457	0.9815	24.0007	<.0001
F23458	0.9652	0.0212	0.8892	0.9897	27.6211	<.0001
F23457	0.9633	0.0222	0.8847	0.989	27.1174	<.0001
F234578	0.9771	0.0151	0.9189	0.9938	30.7464	<.0001
F23456	0.9666	0.0204	0.8932	0.9901	28.2268	<.0001
F234568	0.9792	0.0139	0.9254	0.9944	32.0423	<.0001
F234567	0.9781	0.0145	0.9223	0.9941	31.6301	<.0001
F2345678	0.9864	0.00965	0.9465	0.9966	35.4426	<.0001
F1	0.5946	0.0433	0.5077	0.676	4.5454	0.033
F18	0.7047	0.0538	0.5897	0.7984	11.3081	0.0008
F17	0.6933	0.0563	0.5736	0.7916	9.4892	0.0021
F178	0.7862	0.0544	0.661	0.874	16.1823	<.0001
F16	0.7136	0.057	0.5906	0.8115	10.7187	0.0011
F168	0.8021	0.0526	0.6793	0.8858	17.8695	<.0001
F167	0.7934	0.0547	0.6662	0.8808	16.2347	<.0001
F1678	0.862	0.0451	0.7483	0.9292	23.3985	<.0001
F15	0.7944	0.0648	0.6398	0.8937	11.6185	0.0007
F158	0.8627	0.0529	0.7237	0.9378	16.9504	<.0001
F157	0.8562	0.0565	0.7078	0.9361	15.1137	0.0001
F1578	0.9064	0.0426	0.7833	0.9629	20.3956	<.0001
F156	0.8678	0.0534	0.7251	0.9423	16.3697	<.0001
F1568	0.9143	0.0396	0.7985	0.9664	21.9385	<.0001
F1567	0.91	0.0422	0.7865	0.9652	20.1635	<.0001
F15678	0.9427	0.0298	0.8479	0.9798	25.7151	<.0001
F14	0.7091	0.0641	0.5699	0.8176	8.2084	0.0042
F148	0.7985	0.0603	0.6554	0.892	13.5176	0.0002
F147	0.7897	0.0602	0.6486	0.8843	13.321	0.0003
F1478	0.8593	0.0506	0.729	0.9328	18.695	<.0001
F146	0.8054	0.0593	0.6636	0.8968	14.1132	0.0002
F1468	0.8707	0.0483	0.7439	0.9398	19.7764	<.0001
F1467	0.8645	0.0488	0.7381	0.9352	19.7685	<.0001
F14678	0.9121	0.0371	0.8071	0.9626	25.4937	<.0001

F145	0.8652	0.0506	0.7327	0.9376	18.339	<.0001
F1458	0.9126	0.039	0.8	0.9646	22.9606	<.0001
F1457	0.9082	0.0407	0.7915	0.9627	21.9899	<.0001
F14578	0.9415	0.0297	0.8485	0.9788	26.6084	<.0001
F1456	0.916	0.0383	0.8041	0.9666	22.9818	<.0001
F14568	0.9466	0.0275	0.8591	0.981	27.8633	<.0001
F14567	0.9438	0.0288	0.8529	0.9799	27.0208	<.0001
F145678	0.9647	0.0199	0.8966	0.9885	31.8924	<.0001
F13	0.7702	0.0731	0.5987	0.8828	8.5773	0.0034
F138	0.845	0.0606	0.6876	0.931	13.4378	0.0002
F137	0.8378	0.0638	0.6731	0.9284	12.2431	0.0005
F1378	0.8936	0.0488	0.7543	0.9583	17.1736	<.0001
F136	0.8506	0.0599	0.6933	0.9348	13.6185	0.0002
F1368	0.9025	0.0451	0.7722	0.962	18.845	<.0001
F1367	0.8977	0.0475	0.7609	0.9603	17.6171	<.0001
F13678	0.9345	0.034	0.8277	0.977	22.9005	<.0001
F135	0.8983	0.0512	0.7464	0.9636	15.0943	0.0001
F1358	0.9349	0.0368	0.8146	0.9791	19.4365	<.0001
F1357	0.9315	0.0393	0.8027	0.9785	17.9573	<.0001
F13578	0.9568	0.0271	0.8595	0.9877	22.2754	<.0001
F1356	0.9375	0.0361	0.8177	0.9805	19.3263	<.0001
F13568	0.9606	0.0247	0.8713	0.9887	23.86	<.0001
F13567	0.9585	0.0264	0.8629	0.9884	22.3846	<.0001
F135678	0.9741	0.0176	0.9052	0.9933	26.9052	<.0001
F134	0.8478	0.0649	0.6753	0.9371	11.6745	0.0006
F1348	0.9006	0.0495	0.7542	0.9639	15.9142	<.0001
F1347	0.8956	0.0511	0.746	0.9616	15.4443	<.0001
F13478	0.9331	0.037	0.8136	0.9781	19.7499	<.0001
F1346	0.9044	0.0478	0.7621	0.9654	16.5529	<.0001
F13468	0.939	0.0341	0.8272	0.9802	21.0569	<.0001
F13467	0.9358	0.0354	0.8211	0.9789	20.6556	<.0001
F134678	0.9595	0.0245	0.8733	0.9879	25.2162	<.0001
F1345	0.9362	0.036	0.8182	0.9795	19.8607	<.0001
F13458	0.9598	0.0251	0.8696	0.9884	23.7754	<.0001
F13457	0.9576	0.0265	0.8629	0.9878	22.8575	<.0001
F134578	0.9735	0.018	0.9037	0.9931	26.7619	<.0001
F13456	0.9614	0.0243	0.8734	0.989	24.0925	<.0001
F134568	0.9759	0.0164	0.9117	0.9937	28.1714	<.0001
F134567	0.9746	0.0173	0.9072	0.9934	27.31	<.0001

F1345678	0.9842	0.0114	0.9363	0.9962	31.3826	<.0001
F12	0.7143	0.0548	0.5963	0.8088	11.6438	0.0006
F128	0.8026	0.0535	0.6771	0.8874	17.2403	<.0001
F127	0.7939	0.0542	0.6682	0.8805	16.5985	<.0001
F1278	0.8624	0.0462	0.7449	0.9308	22.2003	<.0001
F126	0.8094	0.0525	0.6854	0.8922	18.0426	<.0001
F1268	0.8735	0.0435	0.7614	0.9373	24.0664	<.0001
F1267	0.8674	0.0444	0.7542	0.9331	23.6329	<.0001
F12678	0.9141	0.0341	0.8196	0.9614	29.6423	<.0001
F125	0.8682	0.051	0.7334	0.9403	17.9223	<.0001
F1258	0.9146	0.0389	0.8015	0.966	22.6965	<.0001
F1257	0.9103	0.0409	0.7916	0.9644	21.3709	<.0001
F12578	0.9429	0.0295	0.8492	0.9797	26.1266	<.0001
F1256	0.9179	0.0381	0.8059	0.9679	22.8149	<.0001
F12568	0.9479	0.0272	0.861	0.9816	27.8377	<.0001
F12567	0.9452	0.0286	0.8539	0.9807	26.6176	<.0001
F125678	0.9656	0.0197	0.8977	0.989	31.6236	<.0001
F124	0.8059	0.0588	0.6653	0.8967	14.342	0.0002
F1248	0.871	0.0493	0.7408	0.941	18.9614	<.0001
F1247	0.8649	0.049	0.7378	0.9357	19.5994	<.0001
F12478	0.9124	0.0381	0.8036	0.9636	24.1785	<.0001
F1246	0.8758	0.0469	0.7518	0.9426	20.5188	<.0001
F12468	0.9198	0.0357	0.8162	0.9673	25.405	<.0001
F12467	0.9158	0.0359	0.8137	0.9644	26.3097	<.0001
F124678	0.9465	0.0261	0.8657	0.9798	31.1566	<.0001
F1245	0.9162	0.0372	0.809	0.9658	24.4232	<.0001
F12458	0.9468	0.0273	0.8603	0.9809	28.2911	<.0001
F12457	0.944	0.0282	0.8557	0.9796	28.0611	<.0001
F124578	0.9648	0.0199	0.8969	0.9886	31.9622	<.0001
F12456	0.9489	0.0262	0.8655	0.9817	29.1899	<.0001
F124568	0.968	0.0183	0.9048	0.9897	33.35	<.0001
F124567	0.9663	0.019	0.9015	0.989	33.2464	<.0001
F1245678	0.979	0.0129	0.9315	0.9938	37.406	<.0001
F123	0.851	0.0579	0.7001	0.9332	14.5641	0.0001
F1238	0.9028	0.0447	0.7739	0.9618	19.1516	<.0001
F1237	0.898	0.0465	0.7649	0.9597	18.3479	<.0001
F12378	0.9347	0.0339	0.8281	0.977	22.9344	<.0001
F1236	0.9066	0.0431	0.7818	0.9633	19.9823	<.0001
F12368	0.9404	0.031	0.842	0.979	24.824	<.0001

F12367	0.9373	0.0324	0.8355	0.9778	24.091	<.0001
F123678	0.9605	0.0225	0.8838	0.9873	28.9357	<.0001
F1235	0.9377	0.0347	0.8243	0.9797	20.7998	<.0001
F12358	0.9607	0.0242	0.8743	0.9885	24.8281	<.0001
F12357	0.9587	0.0256	0.8672	0.988	23.6256	<.0001
F123578	0.9742	0.0174	0.9069	0.9932	27.6366	<.0001
F12356	0.9623	0.0234	0.8782	0.9891	25.1988	<.0001
F123568	0.9765	0.0158	0.9153	0.9938	29.4027	<.0001
F123567	0.9752	0.0167	0.9104	0.9935	28.2495	<.0001
F1235678	0.9846	0.0111	0.9387	0.9963	32.439	<.0001
F1234	0.9047	0.0469	0.7658	0.965	17.1412	<.0001
F12348	0.9391	0.0341	0.8275	0.9803	21.0726	<.0001
F12347	0.936	0.035	0.823	0.9787	21.0611	<.0001
F123478	0.9597	0.0245	0.8729	0.988	24.9957	<.0001
F12346	0.9416	0.0324	0.8357	0.9808	22.3255	<.0001
F123468	0.9633	0.0225	0.883	0.9891	26.4473	<.0001
F123467	0.9613	0.0232	0.8798	0.9883	26.5694	<.0001
F1234678	0.9758	0.0157	0.9161	0.9934	30.6933	<.0001
F12345	0.9615	0.0236	0.8774	0.9887	25.4458	<.0001
F123458	0.976	0.0161	0.9133	0.9937	28.9229	<.0001
F123457	0.9747	0.0169	0.9096	0.9933	28.4232	<.0001
F1234578	0.9843	0.0113	0.9371	0.9962	31.9038	<.0001
F123456	0.977	0.0154	0.917	0.9939	29.8042	<.0001
F1234568	0.9857	0.0103	0.9427	0.9966	33.4472	<.0001
F1234567	0.9849	0.0108	0.9402	0.9963	33.028	<.0001
F12345678	0.9907	0.00711	0.9592	0.9979	36.668	<.0001

Table A.5- Predictive Index of Imputed MESA Data

Contrast Estimation and Testing Results by Row

Contrast	Estimate	Standard Error	Confidence	Limits	Wald Chi-Square	Pr > ChiSq
0	0.5	0	.	.	.	.
F8	0.6081	0.0383	0.5311	0.6801	7.479	0.0062
F7	0.5925	0.0403	0.5117	0.6685	5.0246	0.025
F78	0.6928	0.0489	0.5898	0.7797	12.5356	0.0004
F6	0.6105	0.0469	0.5157	0.6975	5.1955	0.0226
F68	0.7086	0.0514	0.5988	0.7985	12.7337	0.0004
F67	0.695	0.0523	0.5842	0.787	11.1503	0.0008
F678	0.7795	0.0495	0.6677	0.8615	19.1909	<.0001
F5	0.6963	0.0676	0.5506	0.811	6.731	0.0095
F58	0.7806	0.0628	0.6342	0.8795	11.9717	0.0005
F57	0.7692	0.0654	0.6181	0.8728	10.6669	0.0011
F578	0.838	0.0555	0.6988	0.9202	16.1454	<.0001
F56	0.7823	0.0656	0.628	0.8843	11.0152	0.0009
F568	0.8479	0.0544	0.7091	0.9273	16.578	<.0001
F567	0.8393	0.0566	0.6964	0.9224	15.5106	<.0001
F5678	0.8902	0.0443	0.7695	0.9516	21.3672	<.0001
F4	0.6214	0.0517	0.5162	0.7163	5.0858	0.0241
F48	0.718	0.0569	0.5949	0.8154	11.079	0.0009
F47	0.7047	0.0565	0.5836	0.8024	10.2667	0.0014
F478	0.7873	0.0538	0.6635	0.8742	16.5724	<.0001
F46	0.7201	0.059	0.5917	0.8203	10.4101	0.0013
F468	0.7996	0.054	0.6733	0.8854	16.8491	<.0001
F467	0.789	0.054	0.6644	0.876	16.5311	<.0001
F4678	0.853	0.0456	0.7399	0.9221	23.3658	<.0001
F45	0.79	0.0565	0.6588	0.88	15.1545	<.0001
F458	0.8538	0.0488	0.7307	0.9263	20.3598	<.0001
F457	0.8454	0.0501	0.7207	0.9206	19.6444	<.0001
F4578	0.8946	0.0404	0.7858	0.9515	24.9714	<.0001
F456	0.855	0.0497	0.7288	0.9283	19.59	<.0001
F4568	0.9015	0.0392	0.7938	0.956	25.1165	<.0001
F4567	0.8955	0.0403	0.7864	0.9523	24.8272	<.0001
F45678	0.9301	0.0305	0.8415	0.9709	30.5174	<.0001
F3	0.676	0.0771	0.5115	0.8062	4.3697	0.0366
F38	0.764	0.071	0.5994	0.8751	8.8976	0.0029

F37	0.7521	0.0748	0.5801	0.8695	7.6477	0.0057
F378	0.8248	0.0632	0.6663	0.9173	12.5326	0.0004
F36	0.7658	0.0729	0.5959	0.8788	8.5023	0.0035
F368	0.8354	0.0605	0.6817	0.9232	13.6127	0.0002
F367	0.8262	0.0636	0.6661	0.9189	12.3742	0.0004
F3678	0.8806	0.0498	0.7447	0.9491	17.8109	<.0001
F35	0.8271	0.0697	0.6478	0.9256	10.3004	0.0013
F358	0.8813	0.0549	0.7262	0.9541	14.5797	0.0001
F357	0.8743	0.0582	0.7111	0.9516	13.3954	0.0003
F3578	0.9152	0.0437	0.7814	0.9702	17.8123	<.0001
F356	0.8823	0.0556	0.7242	0.9554	14.1638	0.0002
F3568	0.9208	0.0413	0.7929	0.9725	18.7212	<.0001
F3567	0.916	0.0438	0.7814	0.9708	17.6346	<.0001
F35678	0.9442	0.0315	0.8397	0.982	22.3531	<.0001
F34	0.774	0.0755	0.5951	0.8887	8.1316	0.0044
F348	0.8416	0.0629	0.6783	0.9305	12.541	0.0004
F347	0.8327	0.0655	0.6645	0.926	11.6538	0.0006
F3478	0.8854	0.0514	0.7412	0.9542	16.3071	<.0001
F346	0.843	0.063	0.6787	0.9317	12.4779	0.0004
F3468	0.8928	0.0488	0.7541	0.9577	17.2974	<.0001
F3467	0.8864	0.0509	0.7436	0.9545	16.5455	<.0001
F34678	0.9237	0.0378	0.8088	0.9719	21.6169	<.0001
F345	0.887	0.0517	0.7408	0.9557	15.9804	<.0001
F3458	0.9241	0.0389	0.8041	0.9731	20.274	<.0001
F3457	0.9194	0.041	0.7942	0.9712	19.362	<.0001
F34578	0.9466	0.0298	0.8479	0.9825	23.7474	<.0001
F3456	0.9248	0.0389	0.8043	0.9736	20.1263	<.0001
F34568	0.9502	0.0281	0.8563	0.9839	24.6465	<.0001
F34567	0.9471	0.0296	0.8492	0.9827	23.9318	<.0001
F345678	0.9652	0.0209	0.8914	0.9895	28.5702	<.0001
F2	0.5862	0.0397	0.5069	0.6613	4.5336	0.0332
F28	0.6874	0.0511	0.5798	0.7779	10.9838	0.0009
F27	0.6732	0.0512	0.5661	0.7648	9.624	0.0019
F278	0.7617	0.0523	0.645	0.849	16.2472	<.0001
F26	0.6895	0.0539	0.5756	0.7843	10.0522	0.0015
F268	0.7751	0.0524	0.6565	0.8613	16.9324	<.0001
F267	0.7635	0.0526	0.6459	0.8511	16.1799	<.0001
F2678	0.8336	0.0462	0.7227	0.9059	23.3728	<.0001

F25	0.7646	0.0648	0.6159	0.8681	10.7019	0.0011
F258	0.8344	0.0563	0.6941	0.918	15.7738	<.0001
F257	0.8252	0.0582	0.6816	0.9124	14.7917	0.0001
F2578	0.8799	0.047	0.7538	0.946	20.0146	<.0001
F256	0.8358	0.0572	0.6923	0.9201	15.256	<.0001
F2568	0.8876	0.0454	0.764	0.9507	20.6237	<.0001
F2567	0.881	0.047	0.7546	0.9469	19.9413	<.0001
F25678	0.9199	0.0356	0.8166	0.9673	25.5121	<.0001
F24	0.6993	0.0572	0.5771	0.7985	9.6245	0.0019
F248	0.783	0.056	0.6542	0.8731	15.1751	<.0001
F247	0.7717	0.0554	0.646	0.8623	14.9867	0.0001
F2478	0.8399	0.049	0.7198	0.9146	20.7102	<.0001
F246	0.7847	0.0558	0.656	0.8744	15.3228	<.0001
F2468	0.8497	0.0479	0.7304	0.9219	21.2868	<.0001
F2467	0.8412	0.0478	0.7242	0.9145	21.6784	<.0001
F24678	0.8915	0.0386	0.7899	0.9473	27.8722	<.0001
F245	0.8421	0.0502	0.718	0.9178	19.69	<.0001
F2458	0.8922	0.0412	0.7814	0.9504	24.3685	<.0001
F2457	0.8857	0.0421	0.7743	0.946	24.2409	<.0001
F24578	0.9232	0.0327	0.8295	0.9674	29.032	<.0001
F2456	0.8931	0.041	0.7825	0.951	24.3898	<.0001
F24568	0.9284	0.0314	0.8371	0.9703	29.4283	<.0001
F24567	0.9239	0.0322	0.832	0.9675	29.7474	<.0001
F245678	0.9496	0.0238	0.8768	0.9804	34.9032	<.0001
F23	0.7473	0.072	0.5835	0.8619	8.0877	0.0045
F238	0.821	0.0624	0.6661	0.9134	12.8548	0.0003
F237	0.8113	0.0653	0.6506	0.9084	11.6758	0.0006
F2378	0.8696	0.0528	0.7282	0.9432	16.6336	<.0001
F236	0.8225	0.0627	0.6664	0.9149	12.759	0.0004
F2368	0.8779	0.0499	0.7427	0.9471	17.9451	<.0001
F2367	0.8707	0.0522	0.7307	0.9436	16.909	<.0001
F23678	0.9127	0.0396	0.7979	0.9651	22.3083	<.0001
F235	0.8714	0.0571	0.7139	0.9485	14.087	0.0002
F2358	0.9132	0.0435	0.7819	0.9686	18.3554	<.0001
F2357	0.9079	0.046	0.7703	0.9666	17.3023	<.0001
F23578	0.9386	0.0338	0.829	0.9797	21.6643	<.0001
F2356	0.914	0.0435	0.7822	0.9692	18.2282	<.0001
F23568	0.9428	0.0317	0.839	0.9812	22.7397	<.0001

F23567	0.9392	0.0334	0.8305	0.9799	21.8447	<.0001
F235678	0.9599	0.0237	0.8772	0.9877	26.4756	<.0001
F234	0.8291	0.0644	0.6656	0.9221	12.0646	0.0005
F2348	0.8828	0.0514	0.7398	0.9522	16.5131	<.0001
F2347	0.8758	0.0534	0.7293	0.9486	15.8285	<.0001
F23478	0.9163	0.0406	0.795	0.9686	20.431	<.0001
F2346	0.8838	0.0508	0.7427	0.9525	16.8428	<.0001
F23468	0.9219	0.0382	0.8066	0.9709	21.6204	<.0001
F23467	0.9171	0.0397	0.7989	0.9685	21.1649	<.0001
F234678	0.9449	0.0289	0.8522	0.9808	26.1175	<.0001
F2345	0.9175	0.0405	0.7957	0.9695	20.2372	<.0001
F23458	0.9452	0.0299	0.8477	0.9817	24.3467	<.0001
F23457	0.9418	0.0314	0.8406	0.9802	23.6998	<.0001
F234578	0.9617	0.0225	0.8836	0.9881	27.8834	<.0001
F23456	0.9458	0.0295	0.8494	0.9818	24.6278	<.0001
F234568	0.9644	0.0211	0.8906	0.989	28.9558	<.0001
F234567	0.962	0.0221	0.8856	0.9881	28.556	<.0001
F2345678	0.9752	0.0155	0.9182	0.9928	32.9761	<.0001
F1	0.5974	0.0392	0.5189	0.6712	5.8734	0.0154
F18	0.6972	0.0494	0.5928	0.7845	12.7203	0.0004
F17	0.6833	0.0515	0.5751	0.7747	10.4465	0.0012
F178	0.77	0.0513	0.6549	0.8552	17.4048	<.0001
F16	0.6993	0.0508	0.5915	0.7888	12.1898	0.0005
F168	0.783	0.0492	0.6716	0.8643	19.6185	<.0001
F167	0.7717	0.0506	0.6581	0.8559	17.9661	<.0001
F1678	0.8399	0.044	0.7341	0.9088	25.6208	<.0001
F15	0.7728	0.0672	0.6164	0.8781	10.2341	0.0014
F158	0.8407	0.057	0.6963	0.924	15.2856	<.0001
F157	0.8318	0.0597	0.6817	0.9195	14.0191	0.0002
F1578	0.8847	0.0474	0.7553	0.9502	19.2269	<.0001
F156	0.8421	0.0573	0.6961	0.9254	15.0787	0.0001
F1568	0.8922	0.0449	0.7682	0.9538	20.5068	<.0001
F1567	0.8857	0.047	0.7572	0.9506	19.4454	<.0001
F15678	0.9232	0.0352	0.8196	0.9695	25.0716	<.0001
F14	0.7089	0.0588	0.5822	0.8097	9.7686	0.0018
F148	0.7907	0.0559	0.6609	0.8799	15.4949	<.0001
F147	0.7798	0.0566	0.6498	0.8711	14.7048	0.0001
F1478	0.846	0.0489	0.7248	0.9197	20.616	<.0001

F146	0.7924	0.055	0.6647	0.8802	16.0702	<.0001
F1468	0.8555	0.0466	0.7389	0.9253	22.2897	<.0001
F1467	0.8473	0.0472	0.7306	0.919	22.024	<.0001
F14678	0.8959	0.0376	0.7961	0.95	28.4784	<.0001
F145	0.8481	0.0525	0.7151	0.9255	17.7753	<.0001
F1458	0.8965	0.0421	0.7807	0.9547	22.6533	<.0001
F1457	0.8903	0.0436	0.7717	0.9512	21.9545	<.0001
F14578	0.9264	0.0333	0.8287	0.9704	26.938	<.0001
F1456	0.8974	0.0416	0.783	0.955	23.0293	<.0001
F14568	0.9314	0.0314	0.8384	0.9726	28.2258	<.0001
F14567	0.9271	0.0325	0.832	0.9703	27.9134	<.0001
F145678	0.9518	0.0237	0.8775	0.9819	33.244	<.0001
F13	0.7559	0.0726	0.5889	0.87	8.262	0.004
F138	0.8277	0.0619	0.6724	0.9183	13.0841	0.0003
F137	0.8182	0.0655	0.655	0.9143	11.6648	0.0006
F1378	0.8748	0.0522	0.7331	0.9467	16.6574	<.0001
F136	0.8291	0.0616	0.6743	0.9192	13.2028	0.0003
F1368	0.8828	0.0485	0.7502	0.9497	18.5272	<.0001
F1367	0.8758	0.0513	0.7368	0.9468	17.157	<.0001
F13678	0.9163	0.0386	0.8034	0.967	22.6694	<.0001
F135	0.8765	0.0579	0.7132	0.953	13.4156	0.0002
F1358	0.9168	0.0436	0.7824	0.9712	17.634	<.0001
F1357	0.9117	0.0464	0.7695	0.9696	16.436	<.0001
F13578	0.9412	0.0337	0.8292	0.9814	20.7497	<.0001
F1356	0.9175	0.0433	0.7836	0.9716	17.6926	<.0001
F13568	0.9452	0.0313	0.8407	0.9826	22.1988	<.0001
F13567	0.9418	0.0332	0.8314	0.9815	21.0925	<.0001
F135678	0.9617	0.0234	0.8784	0.9887	25.7154	<.0001
F134	0.8356	0.0649	0.668	0.9277	11.8263	0.0006
F1348	0.8875	0.0511	0.7431	0.9556	16.2797	<.0001
F1347	0.8808	0.0536	0.731	0.9526	15.3583	<.0001
F13478	0.9198	0.0403	0.7973	0.9709	19.9655	<.0001
F1346	0.8884	0.0502	0.747	0.9555	16.7893	<.0001
F13468	0.9251	0.0374	0.8107	0.9727	21.6303	<.0001
F13467	0.9205	0.0392	0.8019	0.9707	20.8655	<.0001
F134678	0.9473	0.0284	0.8551	0.982	25.8775	<.0001
F1345	0.921	0.0412	0.7934	0.9725	18.8094	<.0001
F13458	0.9476	0.03	0.8468	0.9834	22.93	<.0001

F13457	0.9442	0.0317	0.8387	0.9822	22.0445	<.0001
F134578	0.9633	0.0225	0.8827	0.9892	26.2405	<.0001
F13456	0.9481	0.0296	0.8491	0.9834	23.3794	<.0001
F134568	0.9659	0.0209	0.8909	0.9899	27.7385	<.0001
F134567	0.9637	0.0221	0.8852	0.9892	27.0366	<.0001
F1345678	0.9763	0.0153	0.9183	0.9934	31.4904	<.0001
F12	0.6777	0.0518	0.5691	0.7699	9.8161	0.0017
F128	0.7654	0.0534	0.6456	0.8538	15.8293	<.0001
F127	0.7535	0.0544	0.6324	0.8445	14.5268	0.0001
F1278	0.8259	0.0492	0.708	0.9027	20.6874	<.0001
F126	0.7672	0.0523	0.6498	0.854	16.5636	<.0001
F1268	0.8364	0.0465	0.7244	0.9086	23.0996	<.0001
F1267	0.8273	0.0473	0.7146	0.9016	22.3893	<.0001
F12678	0.8814	0.0389	0.782	0.939	29.1236	<.0001
F125	0.8282	0.0598	0.6789	0.9166	13.9871	0.0002
F1258	0.8821	0.0483	0.7507	0.9489	18.7989	<.0001
F1257	0.8751	0.0504	0.7396	0.9453	17.8584	<.0001
F12578	0.9158	0.0386	0.8032	0.9666	22.7871	<.0001
F1256	0.8831	0.0477	0.7532	0.9492	19.1229	<.0001
F12568	0.9214	0.0362	0.8149	0.969	24.2652	<.0001
F12567	0.9165	0.0377	0.8068	0.9665	23.5984	<.0001
F125678	0.9446	0.0276	0.8583	0.9796	28.8909	<.0001
F124	0.7753	0.0577	0.6432	0.8685	13.9726	0.0002
F1248	0.8426	0.0508	0.7165	0.919	19.1964	<.0001
F1247	0.8338	0.0513	0.7083	0.912	18.9854	<.0001
F12478	0.8861	0.042	0.775	0.9462	24.3427	<.0001
F1246	0.8439	0.0489	0.7231	0.918	20.645	<.0001
F12468	0.8935	0.0395	0.7882	0.9498	26.2751	<.0001
F12467	0.8872	0.04	0.7824	0.945	26.6924	<.0001
F124678	0.9242	0.0307	0.8377	0.9665	32.4797	<.0001
F1245	0.8878	0.0441	0.7687	0.9496	21.8555	<.0001
F12458	0.9247	0.0341	0.8248	0.9697	26.3058	<.0001
F12457	0.92	0.0352	0.8183	0.9671	26.0885	<.0001
F124578	0.9469	0.0262	0.8654	0.9802	30.6336	<.0001
F12456	0.9254	0.0332	0.8285	0.9695	27.4008	<.0001
F124568	0.9506	0.0245	0.8738	0.9816	32.1476	<.0001
F124567	0.9474	0.0253	0.8694	0.9799	32.3682	<.0001
F1245678	0.9655	0.0182	0.9056	0.9879	37.2192	<.0001

F123	0.8144	0.064	0.6569	0.9095	12.2124	0.0005
F1238	0.8719	0.052	0.7322	0.9443	16.9826	<.0001
F1237	0.8645	0.0547	0.7186	0.9409	15.7438	<.0001
F12378	0.9082	0.0421	0.7863	0.9638	20.6237	<.0001
F1236	0.873	0.0509	0.7364	0.9442	17.6109	<.0001
F12368	0.9143	0.0389	0.8014	0.9658	22.7811	<.0001
F12367	0.9091	0.0409	0.7914	0.9634	21.7047	<.0001
F123678	0.9394	0.03	0.8465	0.9776	27.012	<.0001
F1235	0.9096	0.0459	0.7711	0.9678	17.1116	<.0001
F12358	0.9398	0.0337	0.8291	0.9805	21.242	<.0001
F12357	0.936	0.0358	0.8195	0.9792	20.205	<.0001
F123578	0.9578	0.0255	0.868	0.9874	24.4084	<.0001
F12356	0.9403	0.0332	0.8316	0.9805	21.6776	<.0001
F123568	0.9607	0.0236	0.8776	0.9882	26.0614	<.0001
F123567	0.9582	0.025	0.8708	0.9873	25.172	<.0001
F1235678	0.9726	0.0174	0.9077	0.9923	29.6485	<.0001
F1234	0.8781	0.0532	0.7311	0.9502	15.7881	<.0001
F12348	0.9178	0.0406	0.7957	0.9698	20.1395	<.0001
F12347	0.9128	0.0424	0.7867	0.9674	19.459	<.0001
F123478	0.942	0.0311	0.8416	0.9803	23.9185	<.0001
F12346	0.9186	0.0394	0.8008	0.9694	21.1647	<.0001
F123468	0.946	0.0288	0.8531	0.9814	25.835	<.0001
F123467	0.9425	0.0301	0.8468	0.9799	25.3998	<.0001
F1234678	0.9622	0.0214	0.8892	0.9878	30.1964	<.0001
F12345	0.9429	0.0316	0.8393	0.9812	22.8169	<.0001
F123458	0.9624	0.0227	0.8822	0.9887	26.7389	<.0001
F123457	0.96	0.0239	0.8764	0.9878	26.1093	<.0001
F1234578	0.9738	0.0168	0.9109	0.9927	30.0972	<.0001
F123456	0.9628	0.0222	0.885	0.9886	27.6623	<.0001
F1234568	0.9757	0.0155	0.9175	0.9931	31.8006	<.0001
F1234567	0.9741	0.0163	0.9136	0.9926	31.4057	<.0001
F12345678	0.9832	0.0113	0.9388	0.9955	35.623	<.0001

REFERENCES

1. A Manca, S Palmer, “Handling missing data in patient-level cost-effectiveness analysis alongside randomised clinical trials*.” Appl Health Econ Health Policy 2005; 4(2):65-75.*
2. A. Ahmed, M. W. Rich, P. W. Sanders, G. J. Perry, G. L. Bakris, M. R. Zile, T. E. Love, I. B. Aban, and M. G. Shlipak. Chronic kidney disease associated mortality in diastolic versus systolic heart failure: a propensity matched study. The American journal of cardiology, 99(3):393–398, 2007.
3. A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences, 2(1):183–202, 2009.
4. A.C. Diokno, B.M. Brock, M.B. Brown, et al., “Prevalence of urinary incontinence and other urological symptoms in the noninstutionalized elderly,” J Urol 1986; 136:1022.
5. A.C. Diokno, C.M. Sampselle, A.R. Herzog, et al., “Prevention of urinary incontinence by behavioral modification program: a randomized, controlled trial among older women in the community,” J Urol 2004; 171: 1165.
6. A.C. Diokno, M.B. Brown, B.M. Brock, et al., “Clinical and cystometric characteristics of continent and incontinent noninstitutionalized elderly,” J Urol 1988; 140: 567.

7. A.T.Sadiq, M.G. Duaimi, S.A.Shaker, “Data Missing Solution Using Rough Set Theory and Swarm Intelligence”, International Journal of Advanced Computer Science and Information Technology (IJACSIT), Vol. 2, No. 3, 2013, Page 1-6
8. Acock A.C. (2005). Working with missing values. J Marriage Fam, 67, 1012-1028
9. Ambler G, Omar RZ, Royston P, “A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome”, Statistical Methods in Medical Research 2007, 16 pp.277-98.
10. B. Qian, X. Wang, F. Wang, H. Li, J. Ye, and I. Davidson. Active learning from relative queries. In Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, pages 1614–1620. AAAI Press, 2013
11. B.Azhagusundari, A.S. Thanamani, “Feature Selection based on Information Gain” International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-2, Issue-2, January 2013.
12. Bazan, J. G. & Szczuka, M. (2005), The rough set exploration system, in ‘Transactions on Rough Sets III’, Springer, pp. 37–56.
13. Berry, Michael J.A., and Gordon Linoff, Data Mining Techniques For Marketing, Sales and Customer Support, Wiley, 1997
14. C. Zhang, Y.Zu, J. Zhang, S. Zhang, “Clustering-based Missing Value Imputation for Data Preprocessing”, *International Conference on Industrial Informatics, 2006*
15. Cendrowska J.(1987). PRISM, “An algorithm for inducing modular rules”. International Journal of Man-Machine Studies.

16. Clark T.G., Altman D.G. (2003). Developing a prognostic model in the presence of missing data. An ovarian cancer case study. J Clin Epidemiol, 56, 28-37.

17. D.A. Dillman, L.C.Baxter, A. Jackson, “Skip-Pattern Compliance in Three Test Forms: A Theoretical and Empirical Evaluation”, Technical Report #99-01 of the Social & Economic Sciences Research Center.

18. Dilworth S. E., Riley E. D., Perry, H.I, “SAS ® Data Management: How Raw Variables with Complex Skip Patterns and Missing Values are Interpreted with Derived Variables”, SAS Global Forum 2011.

19. Donders, A. R. T., van der Heijden, G. J., Stijnen, T. & Moons, K. G. (2006), ‘Review: a gentle introduction to imputation of missing values’, Journal of clinical epidemiology 59(10), 1087–1091.

20. Dong, L., Xiao, D., Liang, Y., & Liu, Y. 2008. Rough set and fuzzy wavelet neural network integrated with least square weighted fusion algorithm based fault diagnosis research for power transformers. Electric Power Systems Research, 78,129–136.

21. Düntsch I, Gediga G, “Rough set data analysis: A road to non-invasive knowledge discovery”, Metho_os Publishers (UK), 2000.

22. E. M. L. Beale, R. J. A. Little, “Missing Values in Multivariate Analysis”, *Journal of the Royal Statistical Society. Series B, 1975.*

23. Erden C, Tuysuz F, “An Application of Rough Sets Theory on Traffic Accidents”, An International Conference on Engineering and Applied Sciences Optimization, Kos Island, Greece, 4-6 June 2014.

24. F. V. Nelwamondo and T. Marwala, “Rough Sets Computations to Impute Missing Data”, arXiv:0704.3635v1, 26 Apr 2007.

25. Frank E.,Witten I. H., “Generating Accurate Rule Sets Without Global Optimization”. In: Fifteenth International Conference on Machine Learning, 144-151, 1998.

26. G. Heijden, A.Donders, T. Stijnen, K. Moons. “Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example”, *Jounal of Clinical Epidemiology 2006*

27. Ghiselli E. E., Theory of Psychological Measurement 1964.

28. Gongzhu Hu, Feng Gao, “Rearrangement of Attributes in Information Table and its Application for Missing Data Imputation”, Proceedings of the International Conference of Machine Vision and Machine Learning Prague, Czech Republic, August 14-15, 2014.

29. Graham J.W. (2009). Missing data analysis: making it work in the real world. Annu Rev Psychol, 60,549-576.

30. Grzymala-Busse, J. W. & Grzymala-Busse, W. J. (2007), An experimental comparison of three rough set approaches to missing attribute values, in ‘Transactions on rough sets VI’, Springer, pp. 31–50.

31. H Zhong, “The impact of missing data in the estimation of concentration index: a potential source of bias.” *Eur J Health Econ. 2010 Jun;11(3):255-66. doi: 10.1007/s10198-009-0170-5. Epub 2009 Jul 15.*

32. H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, M. Kolehmainen, “Methods for imputation of missing values in air quality data sets*”,* Volume 38, Issue 18, June

2004, Pages 2895–2907.

33. H. Zou and T. Hastie. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2):301–320, 2005.

34. Hosmer, D.W. and Lemeshow, S. (2000) Applied Logistic Regression, John-Wiley & Sons Inc., New York

35. I. Kononenko, “Estimating attributes: Analysis and extensions of relief,” in Proceedings of the Seventh European Conference on Machine Learning. 1994, pp. 171–182, Springer-Verlag

36. J. Fagan, B. V. Greenberg, “Using Graph Theory to Analyze Skip Patterns in Questionnaires,” Bureau of the Census, Statistical Research Division Report Series, SRD Research Report Number: Census/SRD/RR-88/06, 1988.

37. J. Liu, J. Chen, and J. Ye. Large-scale sparse logistic regression. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 547–556. ACM, 2009

38. J. R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA., 1993.

39. J. W. Grzymala-Busse, “A Rough Set Approach to Data with Missing Attribute Values”, Springer-Verlag Berlin Heidelberg 2006.

40. J.M. Jerez, I. Molina, et al. , “Missing Data Imputation Using Statistical and Machine Learning Methods in a Real Breast Cancer Problem” Artificial Intelligence in Medicine, 50, pp.105-115 (2010).

41. Jaccard, J. (2001) Interaction Effects in Logistic Regression, Series: Quantitative Applications in the Social Sciences, Sage Publications Inc., CA

42. K. I. Penny, T. Chesney, “Imputation Methods to Deal with Missing Values when Data Mining Trauma Injury Data”, *28th Int. Conf. Information Technology Interfaces ITI 2006, June 19-22, 2006, Cavtat, Croatia*

43. K. Kira and L. Rendell, “A practical approach to feature selection,” in Proceedings of the Ninth International Conference on Machine Learning. 1992, pp. 249–256, Morgan Kaufmann.

44. Kennedy, Ruby L, et,al, Solving Data Mining Problems Through Pattern Recognition, Prentice-Hall, 1997

45. Kirsopp, C., and Shepperd, M. Case and Feature Subset Selection in Case-Based Software Project Effort Prediction, Proc. 22nd SGAI Int’l Conf. Knowledge-Based Systems and Applied Artificial Intelligence, December 2002.

46. L. Molina, L. Belanche, A. Nebot, "Feature Selection Algorithms : A survey and Experimental Evaluation".

47. Li, J., and Ruhe, G. Attribute Selection and Weighting Using Rough Sets for Effort Estimation by Analogy—Initial Results, Technical Report SEDS-TR-05102, Software Engineering Decision Support Laboratory at the University of Calgary, Canada, October 2005.

48. Lindsey R. Haas, MPH; Paul Y. Takahashi, MD; Nilay D. Shah, PhD; Robert J. Stroebel, MD; Matthew E. Bernard, MD; Dawn M. Finnie, MPA; and James M.

Naessens, ScD, “Risk-Stratification Methods for Identifying Patients for Care Coordination”, Am J Manag Care 2013;19(9):725-732

49. Little, R.J.A., & Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: John Wiley & Sons.
50. Lung, S.-Y. 2007. Efficient text independent speaker recognition with wavelet feature selection based multilayered neural network using supervised learning algorithm. 40, 3616–3620.
51. M. A. Hall. “Correlation-based Attribute Subset Selection for Machine Learning,” Hamilton, New Zealand, 1998
52. M. Dash, H. Liu, “Feature Selection For Classification”, Intelligent Data Analysis, 1997.
53. M. Hall, G. Holmes, “Benchmarking Attribute Selection Techniques For Discrete Class Data Mining”, Transaction on Knowledge and Data Engineering, Vol. 15, No. 3, 1998.
54. M. Hall, L. A. Smith, “Feature Selection For Machine Learning: Comparing a Correlation Based Filter Approach to the Wrapper”, American Association of Artificial Intelligence, 2003.
55. Myrtveit, I., Stensrud, E., and Olsson, U. H. Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods, IEEE Transactions on Software Engineering, 27, 11(2001) 999-1013.
56. Patetta, M. (2002) Categorical Data Analysis Using Logistic Regression Course Notes, Copyright © 2002 by SAS Institute Inc., Cary, NC 27513, USA.

57. Pires M. and Branco J.A., “Comparison of Multinomial Classification Rules”, 1997.

58. R. Agrawal, R. Srikant “Fast Algorithms for Mining Association Rules”, Proceedings of the 20th VLDB Conference Santiago, Chile, 1994.

59. S. Arslanturk, M. R. Siadat, T. Ogunyemi, I. Sethi, A.Diokno. “Comparison of Attribute Selection Techniques Using Fully Controlled Simulation Based Datasets”, 2nd *International Conference on Information Management and Evaluation 2011,* Toronto, Canada

60. S. Arslanturk, M-R. Siadat, T. Ogunyemi, K. Demirovic, A. Diokno, “Skip Pattern Analysis for Detection of Undetermined and Inconsistent Data”, 2012 5th International Conference on BioMedical Engineering and Informatics (BMEI 2012).

61. S. Dumais, J. Platt, D. Heckerman, and M. Sahami, “Inductive learning algorithms and representations for text categorization,” in Proceedings of the International Conference on Information and Knowledge Management, 1998, pp. 148–155.

62. Schaffer J.L., Graham J.W. (2002). Missing data: our view of the state of the art. Psychol Methods, 7, 147-177.

63. Song, Q., Shepperd, M., and Mair, C. Using Grey Relational Analysis to Predict Software Effort with Small Data Sets, METRICS'05: Proceedings of the 11th IEEE International Software Metrics Symposium. Como, Italy, 2005, 35-45.

64. V. Ouzienkio, Z. Obradovic, “Imputation of missing links and attributes in longitudinal social surveys”, Journal of Machine Learning, Volume 95 Issue 3, June 2014, Pages 329-356.

65. V. Sugumaran, V. Muralidharan, K. I. Ramachandran, “Feature Selection Using Decision Tree and Classification through Proximal Support Vector Machine for fault Diagnostics of Roller Bearing”, Mechanical Systems and Signal Processing, 2006.

66. W. W. Cohen: “Fast Effective Rule Induction. In: Twelfth International Conference on Machine Learning,” 115-123, 1995.

67. Wang F., Zhang P., Qian B, Wang X., Davidson I., “Clinical risk prediction with multilinear sparse logistic regression”, Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. Pages 145-154, 2014.

68. Weiss, Sholom M., and Nitin Indurkhya, Predictive Data Mining: A Practical Guide, Morgan Kaufmann, 1998

69. WR. Lenderking, JF. Nackley, RB. Anderson, MA. Testa, “A review of the quality of life aspects of urinary urge incontinence”, Pharmacoeconomics 1996:1:11-23.

70. X. Su, C-L. Tsai, H. Wang, D. M. Nickerson, B. Li, “Subgroup Analysis via Recursive Partitioning”, Journal of Machine Learning Research 10 (2009)141-158.

71. Ye, J., Chow, J.-H., Chen, J., Zheng, Z. 2009. Stochastic gradient boosted distributed decision trees. In Proceeding of the 18th ACM conference on Information and knowledge management. 2061–2064.

72. Yiming Yang and Jan O. Pedersen, “A comparative study on feature selection in text categorization,” in International Conference on Machine Learning, 1997, pp. 412–420.

73. Yu, S.N., & Chen, Y.-H. 2007. Electrocardiogram beat classification based on wavelet transformation and probabilistic neural network. Pattern Recognition Letters, 28, 1142–1150.

74. Yuanyuan Li, L.E. Parker, “Full Length Article: Nearest neighbor imputation using spatial-temporal correlations in wireless sensor networks”, Journal Information Fusion, Volume 15, January, 2014, Pages 64-79.

75. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M., Shenker, S., Stoica, I. 2012. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation.