

2017

Data Masking, Encryption, and their Effect on Classification Performance: Trade-offs Between Data Security and Utility

Juan C. Asenjo

Nova Southeastern University, asenjo.juanc@gmail.com

This document is a product of extensive research conducted at the Nova Southeastern University [College of Engineering and Computing](#). For more information on research and degree programs at the NSU College of Engineering and Computing, please click [here](#).

Follow this and additional works at: https://nsuworks.nova.edu/gscis_etd



Part of the [Computer Engineering Commons](#), and the [Computer Sciences Commons](#)

Share Feedback About This Item

NSUWorks Citation

Juan C. Asenjo. 2017. *Data Masking, Encryption, and their Effect on Classification Performance: Trade-offs Between Data Security and Utility*. Doctoral dissertation. Nova Southeastern University. Retrieved from NSUWorks, College of Engineering and Computing. (1010)
https://nsuworks.nova.edu/gscis_etd/1010.

This Dissertation is brought to you by the College of Engineering and Computing at NSUWorks. It has been accepted for inclusion in CEC Theses and Dissertations by an authorized administrator of NSUWorks. For more information, please contact nsuworks@nova.edu.

Data Masking, Encryption, and their Effect on Classification Performance:
Trade-offs Between Data Security and Utility

by

Juan C. Asenjo

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in
Information Systems

College of Engineering and Computing
Nova Southeastern University

2017

We hereby certify that this dissertation, submitted by Juan Asenjo, conforms to acceptable standards and is fully adequate in scope and quality to fulfill the dissertation requirements for the degree of Doctor of Philosophy.



Junping Sun, Ph.D.
Chairperson of Dissertation Committee

6/23/2017
Date



Ling Wang, Ph.D.
Dissertation Committee Member


6/23/2017
Date



Osiris Villacampa, Ph.D.
Dissertation Committee Member

6/23/2017
Date

Approved:



Yong X. Tao, Ph.D., P.E., FASME
Dean, College of Engineering and Computing

6/23/2017
Date

College of Engineering and Computing
Nova Southeastern University

2017

An Abstract of a Dissertation Report Submitted to Nova Southeastern University in
Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Data Masking, Encryption, and their Effect on Classification Performance: Trade-offs Between Data Security and Utility

by
Juan C. Asenjo

June 2017

As data mining increasingly shapes organizational decision-making, the quality of its results must be questioned to ensure trust in the technology. Inaccuracies can mislead decision-makers and cause costly mistakes. With more data collected for analytical purposes, privacy is also a major concern. Data security policies and regulations are increasingly put in place to manage risks, but these policies and regulations often employ technologies that substitute and/or suppress sensitive details contained in the data sets being mined. Data masking and substitution and/or data encryption and suppression of sensitive attributes from data sets can limit access to important details. It is believed that the use of data masking and encryption can impact the quality of data mining results. This dissertation investigated and compared the causal effects of data masking and encryption on classification performance as a measure of the quality of knowledge discovery. A review of the literature found a gap in the body of knowledge, indicating that this problem had not been studied before in an experimental setting. The objective of this dissertation was to gain an understanding of the trade-offs between data security and utility in the field of analytics and data mining. The research used a nationally recognized cancer incidence database, to show how masking and encryption of potentially sensitive demographic attributes such as patients' marital status, race/ethnicity, origin, and year of birth, could have a statistically significant impact on the patients' predicted survival. Performance parameters measured by four different classifiers delivered sizeable variations in the range of 9% to 10% between a control group, where the select attributes were untouched, and two experimental groups where the attributes were substituted or suppressed to simulate the effects of the data protection techniques. In practice, this represented a corroboration of the potential risk involved when basing medical treatment decisions using data mining applications where attributes in the data sets are masked or encrypted for patient privacy and security concerns.

Dedication

This mid-life journey into my Ph.D. is dedicated to my mother Angeles Mayoral Asenjo. Thank you for your example and for teaching me to set goals and to never stop learning. One of my fondest memories of childhood are the many afternoons you spent helping my brothers and I with homework at the kitchen table. I am forever grateful.

Acknowledgements

This dissertation would not have been possible without the support of many people. I would like to recognize and extend my sincere gratitude to my advisor, Dr. Junping Sun, who during initial course work sparked my interest in the science of data mining, and who later provided the invaluable insight and direction needed for this dissertation. Thank you for your kind help and patience in guiding me through this complex endeavor. I would also like to thank my dissertation committee members, Dr. Ling Wang and Dr. Osiris Villacampa, who carefully reviewed and helped bring focus to my initial research proposal, and later provided thoughtful and practical input that helped enrich this report.

To my classmate and colleague, Dr. Richard Maiti, I am very grateful for your time, advice, and reassurance. To my employer, Thales e-Security and to Ms. Cynthia Provin, thank you for the sponsorship and the opportunity to pursue my academic goals.

Finally, and most importantly, I would like to thank my family. To my loving wife Anjanette Asenjo, who walked through this journey with me hand in hand, endured the long nights and my many absences, and always offered endless support, advice and encouragement. Thank you for your patience and for listening and always offering your valuable insight. To my children, Marielle and Katja, I cannot tell you how proud I am of you. To you I owe the drive that led me to take on this great adventure.

Table of Contents

Abstract	iii
List of Tables	viii
List of Figures	ix

Chapters

1. Introduction	1
Background	1
Problem Statement	5
Dissertation Goal	5
Research Questions and Hypotheses	6
Relevance and Significance	7
Barriers and Issues	9
Assumptions, Limitations, and Delimitations	10
Definition of Terms	11
List of Acronyms	13
Summary	13
2. Review of Literature	15
Overview of Topic	15
Justification	17
Previous Research	18
Data Quality	20
Security Policies, Regulations, and Data Protection	24
Current State of the Art	25
Gap in the Literature	26
Analysis of Research Methods	28
Synthesis	28
Summary	30
3. Methodology	32
Overview	32
Data	33
Knowledge Discovery Process	35
Data Selection	37
Data Preprocessing	37
Data Transformation	41
Data Mining	47
Interpretation and Evaluation	50
Sample	50
Variables	51
Research Method	53
Internal and External Validity	54

Algorithms	59
Experimental Design	72
Resources Used	77
Hardware	77
Software	77
Summary	78
4. Results	79
Introduction	79
Findings	79
Control Group	81
Experimental Group – Data Masking	85
Experimental Group – Data Encryption	89
Data Analysis	99
Statistical Significance	100
Summary	105
5. Conclusions, Implications, Recommendations, and Summary	110
Introduction	110
Conclusion	111
Implications	112
Recommendations	112
Summary	112
Appendixes	
A. SEER Data Record Description Summary	114
B. Attribute Evaluation and Ranking Results	118
C. Experimental Setup	120
D. Configuration of Classification Algorithms	125
E. Classification Results	126
F. Data Analysis	162
G. Critical Value Table	166
Reference List	168

List of Tables

Tables

1. Control Group Weighted Results 82
2. Experimental Group Weighted Results Using Data Masking 86
3. Experimental Group Weighted Results Using Data Encryption 90
4. Relative Classifier Performance Measured Between Groups 94
5. Average Performance Measured Across Control and Experimental Groups 95
6. Variance of Average Performance Metrics Measured Between Groups 95
7. Variability Among Values Collected from Experiment 102
8. Statistical Significance of Classification Parameters 104
9. Variability of Measured Results by Group 106
10. Multiple Comparison Statistical Significance 108

List of Figures

Figures

1. Schematic Representation of the Knowledge Discovery Process 2
2. KDD Model Followed in Research Study 36
3. Raw Benchmark Data Set from which Balanced Sample was Derived 40
4. Ranking of Most Influential Attributes on the Outcome Variable (Survival) 42
5. Data Transformation Process Model Used in Research Study 44
6. Schematic Representation of Data Preprocessing and Transformation Process 46
7. Code Defining Steps Taken to Determine Information Gain 49
8. Theoretical Framework of Research Scenario 52
9. Confusion Matrix and Precision Calculation 55
10. Representative ROC Graph Depicting Potential Costs as AUC 58
11. Example of Classification Split and Plotted Values Delivered by ZeroR 60
12. Framework of Decision Tree Induction Algorithms 62
13. Performance Results and Plotted Values Delivered by J48 63
14. Code Defining Steps Taken by Naïve Bayes 65
15. Performance Results and Plotted Values Delivered by Naïve Bayes 66
16. Code Defining Steps Taken by AdaBoost Ensemble 68
17. Performance Results and Plotted Values Delivered by AdaBoost Ensemble 69
18. Framework of Ensemble Algorithms 70
19. Performance Results and Plotted Values Delivered by Random Forest Ensemble 71
20. Algorithm Training and Validation Process 76

21. Graphical Representation of Control Group Performance Metrics Delivered by J48, Naïve Bayes, and AdaBoost and Random Forest Ensemble Algorithms Against Baseline ZeroR Classifier 83
22. Overlaid ROC Graph Produced by Classifiers for the Control Group Data Set 84
23. Graphical Representation of Experimental Group Performance Metrics Delivered by J48, Naïve Bayes, and AdaBoost and Random Forest Ensemble Algorithms Against Baseline ZeroR Classifier when Data Masking was Used 87
24. Overlaid ROC Graph Produced by Classifiers for the Experimental Data Masking Group Data Set 88
25. Graphical Representation of Experimental Group Performance Metrics Delivered by J48, Naïve Bayes, and AdaBoost and Random Forest Ensemble Algorithms against Baseline ZeroR Classifier when Data Encryption was Used 91
26. Overlaid ROC Graph Produced by Classifiers for the Experimental Data Encryption Group Data Set 92
27. Graphical Representation Comparing Combined Performance Variance Between Control and Experimental Groups 97
28. Average Performance Variance Between Control and Experimental Groups 98

Chapter 1

Introduction

Background

Data mining has become an incredibly useful technology in business and science. However, if used casually, data mining results can mislead decision-makers and cause costly mistakes. Therefore, the quality of knowledge discovered through data mining is critically important to ensure trust in the technology.

Data mining is increasingly used in decision-making to help explain past and present events, and to predict future states. Among the techniques used to develop predictive models, classification is one of the most widely employed (Tan, Steinbeck, & Kumar, 2006). In the medical field, its use has been shown to be useful in classifying diseases and in helping physicians decide on the most appropriate treatment protocols (Salama, Abdelhalim, & Zeid, 2012). The practice derives knowledge from vast volumes of raw data (also referred to as big data) collected from distributed networked databases to find associations and trends, and to discover new knowledge that would have otherwise remained buried in storage (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Knowledge discovery analyzes data using algorithmic methods until inherent relationships become visible (Fan, 2008). The process of data mining and knowledge discovery fulfills the quest to seek new insight from available data resources (Ahmadi & Abadi, 2013). A schematic representation of the main steps involved in the knowledge discovery process is shown in Figure 1.

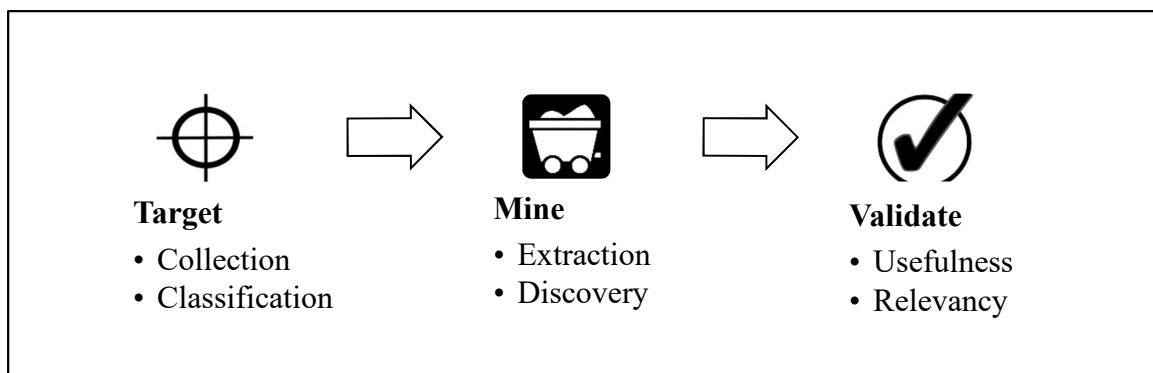


Figure 1. Schematic Representation of the Knowledge Discovery Process

In parallel with the development of knowledge discovery techniques, data security policies and regulations have also gained greater attention. Persistent attacks on government and enterprise computing systems, and high profile data breaches have made data security policies and regulations increasingly common across banking, financial services, and healthcare, among others business areas. Incomplete data sets, missing values, and errors in data entry have been recognized to impact the quality of mining results. However, a review of the literature has shown that the impact of data security policies and regulations and the use of techniques such as data masking and encryption, which substitute sensitive data and suppress important attributes for confidentiality and privacy protection, have not been the focus of in-depth study. While organizations that collect their own data can protect its sensitive aspects while maintaining visibility of these fields or attributes for analytical processing, when researchers collect data from distributed repositories, clear text access to protected fields is not always possible. Access to complete data sets is not always available due to protection given to certain fields as a result of data privacy regulations.

At a time when data mining is increasingly used for decision-making, more data may be masked or encrypted due to its sensitivity, impacting the quality of data mining results and the trustworthiness of the technology. In the healthcare field for example, protected health information is often stripped of important personal characteristics when used for research purposes. This type of data substitution and associated attribute suppression is believed to have a causal impact on the quality of data mining results.

This report is organized in five chapters; introduction, review of literature, methodology, results, and conclusions. Chapter 1 presents the problem investigated, and

describes its persistence and its effects in academia, government, and business environments. The goal and scope of the research are presented, and three specific questions are posed to help guide the literature review and help formulate the hypothesis for the quantitative analysis. The relevance and significance of the research is then explained in light of the affected population. The chapter closes with an identification of barriers and issues encountered during the course of the research, including the steps taken for their mitigation. Specialized terms used throughout the report are also defined.

Chapter 2 presents the literature review with a chronological account of related works. Associated questions, hypotheses, and findings by the various researchers are examined, and the methods developed to study data quality and knowledge discovery are compared and contrasted. The chapter synthesizes available works on the subject and identifies the gap in the body of knowledge that the dissertation fulfills.

Chapter 3 outlines the methodology followed to conduct the study and identifies the research model and tools used to collect the data representative of the problem in question. The chapter also describes how the methodology, model, and tools were tested to validate their feasibility. The chapter describes the variables involved, how the data sample was collected, and how it was analyzed to test the postulated hypothesis. Internal and external validity implications are also discussed. Resources used to conduct the research initiative, including hardware, software, tools, and access to representative data are also described.

Chapter 4 defines the experimental design model and presents the empirical findings. Classification performance parameters calculated by each of the algorithms employed for the control and experimental groups representative of the measured impact of data

masking and encryption on classification performance, are tabulated and graphed for comparison. Results of a statistical analysis are also presented. Finally, an analysis of the results obtained is presented in Chapter 5. The chapter draws conclusions, states associated implications, and presents recommendations for further study.

Problem Statement

Data protection techniques can create inconsistencies and gaps in historical records that can affect the completeness of data (Grimmer & Hinrichs, 2001). Data masking and encryption respectively substitute and suppress important attributes in data sets that can affect knowledge discovery. Increased use of these data protection techniques puts in question the quality of data mining results.

It is believed that the effects of masking and encryption on classification performance, and thereby on the quality of knowledge discovery, has become more acute in recent years as the use of these data mining tools has become more prevalent. With growing use of this technology and increased security awareness, the impact that data protection techniques can have on knowledge discovery is likely to become an even more important subject of study (Ahmadi & Abadi, 2013).

Dissertation Goal

The goal of this research was to develop an experimental model to investigate and compare the causal effect between the use of data masking and encryption on the quality of knowledge discovered through data mining. The research observed the impact that data masking (through sensitive attribute substitution) and encryption (through sensitive attribute suppression) had on knowledge discovery by measuring classification performance parameters including accuracy, precision, and recall among others. As a

dimension of data quality, classification accuracy is an objective metric that determines the capability of algorithms to correctly classify instances in data sets (Pipino, Lee, & Wang, 2002). Precision refers to the degree of separation between the predicted values (Bhuvaneswari, Prabakaran, & Subramaniaswamy, 2015). The lower the number of false positives that a classifier calculates, the higher the precision of the classifier (Tan, Steinbach, & Kumar, 2006). Classifiers are the algorithms that systematically build the data groupings in a data mining application (Tan, Steinbach, & Kumar, 2006). Recall measures completeness of results and aligns with the proportion of positive cases that are correctly predicted to be positive.

By developing a knowledge discovery quality metric, the study demonstrated causal impact and provided a testable scenario that enables repeatable validation of the trade-offs between data security and utility.

Research Questions and Hypotheses

Three specific research questions were postulated to frame the literature review and to focus the objective of the research:

1. Is masking and encryption of attributes in data sets impacting classification performance and the quality of knowledge discovery?
2. Can the impact in performance and quality of knowledge discovery be objectively measured between masking and encryption?
3. Is the measured impact of masking and encryption radically different and statistically significant?

Given the stated problem and the questions that the research study sought to answer, the null and alternate hypotheses included:

H_0 = Data masking and encryption of attributes in data sets have no effect on classification performance and quality of data mining results

H_1 = Data masking and encryption of attributes in data sets have an effect on classification performance and quality of data mining results

Given that classification performance metrics are parameters of data quality, their measurement determines the quality of knowledge discovery. The empirical manner by which this was conducted is described in the methodology outlined in Chapter 3.

Relevance and Significance

The trustworthiness of knowledge discovered through data mining is increasingly critical to organizational decision-making. Factors affecting the data mining process and the knowledge derived from this activity can put in question its credibility (Fan, 2008). While the literature confirms that researchers have examined many aspects of knowledge discovery including quality, the impact of applied data protection techniques such as data masking and encryption, has not been specifically studied. An assessment of the impact of these practices on the quality of knowledge discovery will help prepare decision-makers when using derived business intelligence.

Dependence on data mining and concerns over the reliability and trustworthiness of the derived knowledge is of interest to all those who increasingly use this technology, including academia, government, and enterprise. This research study was meant to be especially useful to regulated industries such as banking, financial services, insurance, and healthcare that handle massive volumes of private and sensitive data subjected to security policies and regulations, and that increasingly use aggregate data from distributed resources with protected fields, and data mining techniques for business

intelligence purposes. Examples in the literature that illustrate the effects of data quality include the work of Grimmer and Hinrichs (2001) and Hipp, Gützler, and Grimmer (2001) showing how poor data quality affected decision support systems in manufacturing. The work of Buja and Lee (2001) revealed how clinical trials depended on data mining techniques and regression and classification process, to gain insight into patient data.

With growing security awareness and increasing use of comprehensive data protection technologies, the impact that data masking and encryption may have on knowledge discovery is likely to become more significant. Haug and Arlbjørn (2011) found that poor data quality impacted the enterprise bottom line and highlighted how decision-makers often could not trust available business intelligence. The increasing use of data mining as a business tool requires an assessment of the quality of extracted information (Ahmadi & Abadi, 2013). Ensuring that data mining results are trustworthy and dependable will become critically important as the technology increasingly shapes organizational decision-making (Alkharboush, 2013). Assessing the quality of discovered knowledge is therefore a task that requires further study.

Given that the literature review presented herein showed a gap in this area of the body of knowledge, this research study was built upon existing studies that have used methodologies already developed and proven reliable by researchers in the field (Al-Badrashiny & Bellaachia, 2016; Al-Bahrani, Agrawal, & Choudhary, 2013; Bellaachia & Guven, 2006; Bostwick & Burke, 2001; Bradley, 1997; Delen, Walker, & Kadam, 2004; Endo, Shibata, & Tanaka, 2008; Rosenberg, Chia, & Plevritis, 2005). The study filled the identified gap where the impact of masked or encrypted data on data mining results have

not been addressed, and provided an original contribution to the literature, establishing a platform for analyzing a real-world application problem.

Barriers and Issues

Assessing the quality of knowledge discovered through data mining in a real-world non-contrived setting is difficult and impractical due to the invasive nature of the process. Access to real-world sensitive data can also infringe on privacy. Organizations that process sensitive data are often prohibited from sharing the data. Sensitive data includes personally identifiable information that has to be protected by law (Cios & Moore, 2002).

The fact that the subject of the research involved organizations that enforce data security policies and regulations and the use of data protection techniques, creates certain barriers. Organizations that use data protection techniques tightly control access to their systems to mitigate risks of data breaches and system compromise (Lu & Miklau, 2008). The use of real-world business data for this study allowed the researcher to observe the phenomena first-hand within the natural environment, but ethical, legal, and operational requirements made this impractical. Prokosch and Ganslandt (2009) studied the reuse of electronic medical records for clinical research and found that there were impediments to the reuse of this data. “Consideration of regulatory requirements, data privacy issues, data standards as well as people/ organizational issues are prerequisites in order to vanquish existing obstacles” (Prokosch & Ganslandt, 2009, p. 38).

Wang (2009) found that environmental factors also had an impact on data quality. Ramakrishnan, Jones, and Sidorova (2011) examined external environmental factors on knowledge discovery and found that in business settings, competitive pressure can be an issue. Specific characteristics of the natural setting being studied that were outside of the

researcher's control, would have resulted in observations that would not have been able to be generalized and applied to the broader population affected by the research problem.

Not having access to the natural environment and the data resources representative of the research questions were nonetheless manageable barriers. Using an alternative research enabled the study to be carried out in a manner that was able to produce repeatable results. An alternative method employing available benchmark data sets was used for experimental purposes. Publicly available benchmark data sets enables real-world records, representative of the natural environment, to be readily used for experimentation (Scalzo, Burlison, Fernandez, Ault, & Kline, 2007). While a full institutional review board appraisal was not needed, given the nature of the benchmark data used, a filing was made along with a formal request for authorization to use the data for research purposes. No other problems were encountered during the performance of the experiment.

Assumptions, Limitations and Delimitations

The research study was based on one main assumption: that the benchmark data set utilized was representative of the real world environment, and that the suppressed sensitive attributes in the data set were indeed relevant to the classification process. These assumptions took for granted that demographics might not be significant factors in all cases. Therefore, it was incumbent on the researcher to ensure that the conditions analyzed, and the results that the data mining exercise predicted, were dependent on at least a subset of the critical demographic values that were masked or encrypted.

Limitations of the study include the fact that recorded demographic factors of the test population were dependent on the accuracy of the original patient records. Any alteration

of these historical records is beyond the control of the researcher and would impact the extent to which one could draw cause and effect relationships.

To limit the degree to which such conditions could impact the internal validity of the research, the scope of the study was controlled, while at the same time ensuring that observations and conclusions of the initiative could be applied to a general population. To this end, delimitations imposed included the selection of a focused data set where the demographics of the sample populations had already been proven to be significant factors in the incidence of the disease being studied (i.e., breast cancer). Additionally, to ensure external validity of observed results, a large sample size was chosen.

Definition of Terms

Terminology used throughout this report is defined as follows:

AdaBoost – Adaptive boosting type of ensemble learning algorithms that uses iterative weighted results of multiple data set instances and classifiers.

Classification accuracy – Rate of correctly identified attributes within a data set.

Benchmark – A test methodology based on real-world use of computer systems.

Data warehousing – Practice where data from distributed database resources are maintained in a centralized location for ease of access and retrieval.

Data mining – Process of extracting knowledge from vast amounts of distributed data to help explain past and present events, and to predict future states.

Ensemble – Library of classification algorithms.

Entropy – Ratio of binary alternatives of an attribute's occurrence within a data set(s).

Imputation – Substitution of missing values with existing similar values found in the data set. The term was previously used in the literature to refer to prepositioning.

Information gain – Amount of useful information contained in set of data.

J48 – WEKA's implementation of the Quinlan C4.5 decision tree-based classifier.

Learning – Formation of classification rules based on training data.

Materialized views – Pre-computed results of frequent database queries.

Naïve Bayes – Probabilistic classifier algorithm used in experiment.

Overfitting – Condition where training data fits too tightly and leads to a useless classification process where nodes have only single branches and no decision point(s).

Precision – Closeness of the various measures recorded.

Predictive accuracy – Capacity of classification algorithm to categorize data tuples for which classification label is not known.

Predictors – Independent variables or attributes that are known and used to train the algorithm being employed for data classification purposes.

Prepositioning – Process of replacing missing values with commonly occurring ones in the training data.

Quality – Accuracy and usefulness of knowledge obtained from a data set after mining for hidden insight.

Quasi-experiment – Type of experimental test where multiple measures are taken before and after intervention or treatment of the independent or predictor variable(s).

Random Forest – Ensemble algorithm method that uses multiple training inputs.

Recall – Completeness of results obtained from an analysis.

Referential integrity – Critical property of relational databases.

Utility – Usefulness of knowledge obtained from a data set(s) for the particular purpose for which it was mined.

ZeroR – Simple classifier used to determine majority category of outcome variable and baseline performance.

List of Acronyms

Acronyms used throughout this report are defined below:

ANOVA – Analysis of variance

AUC – Area under curve

CRISP-DM – Cross-industry standard process for data mining

CSV – Comma separated value

HSD – Honest significance difference

KDD – Knowledge discovery in databases

NAACCR – North American Association of Central Cancer Registries

NHIA – NAACCR Hispanic Identification Algorithm

NCI – National Cancer Institute

NIH – National Institutes of Health

ROC – Receiver operating characteristic

SEER – Surveillance, Epidemiology, and End Results

SEMMA – Sample-explore-modify-model-assess

WEKA – Waikato Environment for Knowledge Analysis

Summary

Data mining has evolved from an experimental technology to an applied scientific and business tool. During this evolution, the use of data protection techniques such as masking and encryption has also proliferated. This has impacted many industries, and has put into question the trustworthiness of knowledge discovery.

The results of this research and quantitative analysis provided experimental evidence showing variances in classifier performance measures between control and experimental groups. By developing a knowledge discovery quality metric, the study demonstrated causal impact and provided a testable scenario for repeatable validation.

Chapter 2

Literature Review

Overview of Topic

A literature review draws a chronological account of related works in the subject area to demonstrate how research has evolved (Salkind, 2012). The historical account facilitates the identification of gaps and the substantiation of how the objective of a research initiative fits within the broader body of knowledge (Hart, 1998). An assessment of over 50 peer-reviewed works on related subject areas was conducted to understand how research in this field had matured, what areas had been examined, and what gaps remained to be studied. Subject areas included data quality, data mining, knowledge discovery, security policies and regulations, and data protection techniques.

In the context of data mining, data quality had been the subject of focused research for over 20 years. Missing values had been recognized as a problem and studied in the context of incomplete data sets and errors in data entry (Farhangfar, Kurgan, & Dy, 2008). The impact of missing values and the effect on classification accuracy had also been studied (Acuña & Rodríguez, 2004). However, an investigation and comparison of the causal effect between the use of data masking and encryption on the quality of knowledge discovery through data mining had not been the subject of focused experimental study, nor had a methodology and tool been developed for repeated validation.

Data anonymization tools using data masking techniques are commonly built into commercial database systems (Vinogradov & Pastyak, 2012). Data masking substitutes

sensitive attribute values in data sets with fictitious ones to hide their sensitivity, protect their confidentiality, and prevent them from being used to re-identify personal identities (Dhir & Garg, 2017). The process is irreversible, and does not allow reconstitution of the original data element once executed. Given that data masking produces a fundamentally comparable data element and data set, its use generally has minimum or no impact to business processes (Ogigau-Neamtii, 2016). Data masking can be static or dynamic. Static masking replaces sensitive attribute values with constant values already present in the data set. Dynamic masking replaces attribute values with random ones within the range represented in the data set (IBM Knowledge Center, 2016). G. K., Rabi, and TN (2012) found that dynamic data masking with random replacement of sensitive values yielded high security with the added convenience of not having to alter processes to accommodate changes in data structure.

Some authors consider encryption to be a form of data masking (Dhir & Garg, 2017). However, encryption performs a significantly different process that altogether suppresses the sensitive data, making it illegible to the naked eye and to data mining algorithms. Unlike masking, encryption is reversible, and relies on the use of cryptographic keys (symmetric or asymmetric) to transform data back to its original legible state (Ogigau-Neamtii, 2016). Encrypted data blocks typically are also structurally different than the original ones. As most encryption algorithms expands the data block, this often requires the alteration of database processes to enable encrypted data to fit within existing application table formats. This, and the complexity associated with managing large groups of cryptographic keys, can make encryption a less favorable data protection alternative for certain applications. Nonetheless, its reversibility continues to make it an

indispensable data protection technique. In terms of the effect on data mining, the process of replacing sensitive attributes with fictitious masked values is believed to have a greater impact on knowledge discovery than the suppression of values through encryption.

Given the fundamental differences in technique, this research differs from the problem of missing data because it measured the effects that the fictitious data may have on knowledge discovery in contrast to the effects of data suppression. With a knowledge discovery quality metric obtained from both masking and encryption, the research compared this trade-off by measuring the classification performance of data mining results obtained using complete, masked, and encrypted attributes in a common data set.

As data masking is increasingly used to de-identify and protect the confidentiality of data, an understanding of how the technology impacts knowledge discovery is important. The contrast between the effects of data masking and encryption is valuable for organizations trying to pick the best solutions to protect their sensitive data, while still being able to maximize the utility of the distributed data resources.

The literature review was separated into two sections. The sections include the research that focused on data quality and how it can be measured, and the research that contributed to furthering the understanding of evolving data security policies, regulations, and data protection techniques.

Justification

Data quality is defined based on five critical characteristics: accuracy, completeness, consistency, actuality, and relevance (Luebbbers, Grimmer, & Jarke, 2003). With growing data security awareness, cryptographic techniques are commonly used to protect the confidentiality of sensitive data. As more databases are subject to stricter security policies

and regulations, it is important to consider the effects that these may have on the five critical characteristics defining data quality, and on the resulting knowledge discovery. An assessment of the literature on this specialized field enabled identification of the gap in the body of knowledge that this study addressed.

Previous Research

Early research in knowledge discovery dates back to the development of mechanisms to uncover hidden rules in relational databases through attribute-oriented induction (Han, Cai, & Cercone, 1993). The concept of the data warehouse as a centralized repository that brought together distributed databases was first developed to improve data availability and performance. The storage of pre-computed results of frequent queries into materialized views optimized efficiency and overall process quality (Gupta, Mumick, & Subrahmanian, 1993). Widom (1995) recognized the advantage of data warehousing over traditional database querying, and identified areas needing dedicated research to take the technology forward. As data mining technology began to evolve, data warehousing became an important enabler. Fayyad, Piatetsky-Shapiro, and Smyth (1996) described how data warehousing aggregated data from multiple distributed sources and compiled the data into common frameworks from where mining algorithms could then analyze and extract meaningful insight and knowledge. Using sophisticated algorithms, data mining applications were then able to analyze more complex interactions between data sets in data warehouses and across heterogeneous networks, and discover knowledge inherently hidden in the data.

As the volume of available data grew, computational resources needed to analyze these for pattern recognition and identification of associations, began to hit performance

limits. While materialized views enabled these challenges to be managed in a more effective way, maintaining them became a critical factor to ensure data completeness (Wu & Buchmann, 1997). Driven by the commercialization potential of data warehousing and data mining technology, Wu and Buchmann (1997) identified areas that needed further study; including data warehouse architecture, data loading, cleansing and purging, data indexing, and query optimization among others. The focus on data cleansing and purging provided one of the first instances that can be linked to the direct effects of data security policies and regulations. García-Molina, Labio, and Yang (1998) recognized that while data volume management was important to ensure optimum performance, indiscriminate purging of expired data could violate referential integrity across databases and adversely affect the stability of data warehouses. Inconsistencies in databases and data warehouses were also found to have a consequential effect on the degree of accuracy and consistency of queries. Accuracy and consistency of data, along with its timeliness, were found to be key characteristics defining data quality (Jeusfeld, Quix, & Jarke, 1998).

Buja and Lee (2001) and Alkharboush (2003) studied how organizations used data mining to discover interesting associations between unrelated data sets, and to discover hidden patterns that could provide insight and greater understanding of available data resources. The tremendous drop in the cost associated with maintaining very large data repositories and the capability to link these distributed sets across large geographies in an economical way, propelled the development and adoption of data mining (Kurgan & Musilek, 2006). Yang and Wu (2006) surveyed the data mining research at the time and identified the 10 most challenging problems in the field to include security, privacy, and

data integrity as problems needing critical attention due to their ability to distort results. However, their study was limited to identifying the factors and not trying to measure their effect. Similarly, Kurgan and Musilek (2006) conducted a survey of knowledge discovery process models with the objective of consolidating research in the specialized field and promoting development of standardized methodologies to ensure greater acceptance in industry. While these studies examined various dimensions of knowledge discovery and its broad applications, the impact of data protection techniques as their use became pervasive, had not been addressed. Roski, Bo-Linn, and Andrews (2014) studied the opportunities offered by mining healthcare data stored across interconnected infrastructures, and found that the value gained could be limited by data security practices designed to protect patient privacy. Their study proposed a series of steps for implementing big data solutions in healthcare organizations focused on identifying patients only by derived insight, to reduce cost and improve quality.

Data Quality

Early work by Redman (1998) recognized the threat that the growing problem of poor data quality represented to enterprise operations and analyzed the tactical and strategic impacts to create greater awareness. In an effort to study undetected inconsistencies in databases and quality of the data mining, Grimmer and Hinrichs (2001) developed the concept of data quality management to determine the degree of confidence in association rule mining. Using a deviation detection technique, they developed a process to find inconsistent associations between items in large databases. A survey conducted by Lee, Strong, Kahn, and Wang (2002), summarized academic research on information quality at the time and identified the different dimensions of importance to users of information.

The survey also classified the quality dimensions in terms of the intrinsic, contextual, representational, and accessibility values of importance to consumers of data. In terms of the intrinsic value of data quality, 13 dimensions were identified in the survey including accuracy, believability, reputation, objectivity, factuality, credibility, consistency, completeness, precision, reliability, freedom from bias, correctness, and unambiguity (Lee et al., 2002). Among the 11 researchers named in the aforementioned survey, accuracy was the dimension most often identified as being critically important to users. Pipino, Lee, and Wang (2002) assessed data quality levels using subjective and objective measures with the goal of developing quality metrics for organizations to be able to put in practice. Luebbers, et al. (2003) defined an objective set of data quality dimensions that included accuracy along with completeness, consistency, actuality, and relevancy.

As a dimension of data quality, accuracy had been used as one of the more objective metrics, particularly in terms of the capability of algorithms to correctly classify instances in data sets. Acuña and Rodríguez (2004) found that the effects of missing values on classifier accuracy was dependent on the number of instances present in the specific data sets. McGarry (2005) conducted a survey of what constituted an “interestingness” measure for knowledge discovery, and found that these fell into two categories: objective, which were based on statistical correlation between data sets, and subjective, which were based on what users anticipated. Yang and Wu (2006) found that algorithms used to mine data could modify or hide certain parameters for privacy and security reasons, distorting the knowledge that could be derived from them. It is at this point in the chronology of the literature that one of the specific parameters of data security (i.e., privacy) is specifically linked to data quality. Examples in healthcare included research in the incidence of

certain diseases, and on medical informatics. Bellaachia and Guven (2006) studied breast cancer survivability and compared the accuracy of various data mining techniques in predicting these incidences. Conversely, Malazizi, Neagu, and Chaudhry (2006) surveyed data quality assessment methods in predictive toxicology and identified significant deficiencies affecting scientific research. Kumari and Godara (2011) studied the performance of various classification techniques in predicting cardiovascular disease using sensitivity, specificity, accuracy, and error rate as comparative metrics. Dimitoglou, Adams, and Jim (2012) also studied cancer patient's rate of survival using various classification algorithms. The general procedures used by Dimitoglou, Adams, and Jim (2012) to measure algorithm performance served as a model for the dissertation work. However, while their research addressed the capabilities that different classifiers had in predicting an outcome given an initial set of input conditions, it did not address the potential effects of data protection techniques used to safeguard the privacy of patients.

While the problem of data quality in data mining has been fully recognized, the measurement of the impact that data substitution and suppression may have on the quality of mined results, has not been experimentally quantified. One of the reasons why this has not been done is because data quality cannot be measured as a stand-alone value. Data quality must be assessed within the context of usage. Shankaranarayanan and Cai (2006) found that no matter the dimensions used to determine data quality, the context for which the data was used also needed to be assessed. What could be considered high quality results for one task, had little or no value for another (Shankaranarayanan & Cai, 2006).

Wang (2009) focused on referential integrity as the key factor affecting data accuracy, consistency, and dependability. His findings showed that data quality also included the

dimensions of intrinsic, contextual, representational, and accessibility values. Intrinsic values were accuracy and objectivity. Accessibility was the ability to be only available to authorized entities. Contextual characteristics represented its relevancy, timeliness and completeness. And finally, its representational characteristics involved interpretability, ease of understanding, and consistency. Wang (2009) indicated that “without a solid foundation of high-quality data, ‘dirty data’ can chip away at an organization’s ability to function effectively” (p.3). Agarwal and Yiliyasi (2010) studied the characteristics of data quality in social media environments and the unique challenges that its informal and unstructured nature had on data mining and machine learning. They defined data quality in social media in the context of four dimensions; intrinsic, contextual, representational, and access value, and singled-out security as the degree to which information was protected against unauthorized access without considering its integrity.

Out of the 13 data quality dimensions identified by Wang and Strong (1996), Blake and Mangiameli (2011) singled out accuracy, completeness, consistency, and timeliness as the more objective ones having a direct impact on data mining classification methods. While the complexity of the classification problem was the principal factor in determining classification outcomes, higher rates of correctly classified positives and correctly classified negatives were indicative of higher accuracy. As a recognized dimension of data quality when using classification algorithms, accuracy was defined as the rate of correctly identified attributes (Blake & Mangiameli, 2011). Sidi et al. (2012) studied dependencies between data quality dimensions and found that accuracy, currency, consistency, and completeness could improve knowledge discovery. Palepu and Rao (2012) studied quality control mechanisms in data warehousing and found that lack of

quality data due to degradation over time, and improper handling, rendered it useless for mining purposes.

Security Policies, Regulations, and Data Protection

Early research on the impact of security policies and regulations on data quality was conducted by García-Molina, Labio, and Yang (1998). Their work focused on the preservation of database referential integrity in light of regulations requiring databases to be purged of sensitive and expired data. Greater awareness of the need for data security polices drove many organizations to take pre-emptive steps to protect the confidentiality, integrity, and privacy of their sensitive data, and their corporate entity from data breaches and associated liabilities. Organization's own policies, as well as industry and government regulations led to the growing use of data protection options like masking and encryption in business applications. Grimmer and Hinrichs (2001) found that these practices, where organizations screened-off sensitive data, created gaps in available records. A more focused study of data security policies and regulation on operational data handling was conducted by Cios and Moore (2002). Their study highlighted ethical, legal, and social issues involved in medical research, and identified specific requirements such as patient data de-identification as one having significant implications on how data was collected and analyzed. Wilson and Rosen (2003) studied the effects of perturbation, or the addition of noise to databases to protect the confidentiality of attributes. Findings of the exploratory study revealed initial evidence of the introduction of a type of data mining bias on results, but did not examine the characteristics of the data. Not covered in Wilson and Rosen (2003) research, but left open for future study, was the need to assess the impact that confidential / protected attributes might have on the ability to discover

knowledge – the area of focus of this dissertation. Islam, Barnaghi, and Brankoviz (2003) also studied the effects of data perturbation on predictive accuracy of decision trees, and found it to be inconsistent in determining resulting data quality. Fan (2008) studied data security policies and regulations and found that they created inconsistencies in databases, and proposed the use of data cleansing algorithms to remove conflicting data in an effort to manage quality. Similarly, Lu and Miklau (2008) studied the impact that data security policies and regulations had on auditing, and also examined technologies to cleanse data warehouses from sensitive data, while maintaining records for auditing purposes.

Farhangfar, et al. (2008) found that classification accuracy when using data sets that were subjected to imputation, was higher than when letting data go missing. Imputation is a technique where missing values in a data set are “filled-in” with similar values based on estimates between existing attributes (Luengo, García, & Herrera, 2011). The research conducted by Farhangfar, et al. (2008) presented a similar analysis to the one conducted in this dissertation, but did not measure the impact that masking and encryption had on data mining results as a function of classification performance. The effect of imputation draws a parallel to data masking, and the effect of missing values to encryption.

Motiwalla and Li (2013) postulated that encryption of healthcare data had the same effect as its removal, since it could not be used for any practical purpose. Motiwalla and Li (2013) researched the use of data masking as an alternative for protecting patient data, and developed a system to protect privacy without removing sensitive attributes.

Current State of the Art

Benitez and Malin (2009) studied the risks of data re-identification within specific demographics and determined that these were not generally recognized by policy makers

when deciding to share records containing sensitive data. To address this problem, current state of the art in data mining included innovative techniques such as privacy preservation data mining. The concept recognized the threat that data mining could pose to privacy, and enables individual database contributors to provide anonymized data to a trusted third party service broker so that the privacy of the source can be maintained (Fletcher & Islam, 2015; Keshavamurthy, Khan, & Toshniwal, 2013).

More dedicated research in the field of knowledge discovery began to focus on the specific characteristics of data attributes, and the accuracy of their results. This research showed that mining on the attributes that were most significant to the question under study yielded more accurate results (Ahmadi & Abadi, 2013; Blake & Mangiameli, 2011; Farhangfar, et al., 2008). However, attributes that were most significant could often be masked or suppressed. Current developments in the field include specialized algorithms that protect sensitive data while enabling it to be used for knowledge discovery (Kalariya, Shah, & Vala, 2015). While these studies recognized the effects of data security policies and regulations on such aspects as operations and auditing, they did not drill down into the impact that the use of data masking and encryption could have on knowledge discovery. Given that data masking and encryption are increasingly used to protect sensitive data, the extent to which they may impact the quality of data mining results was deemed deserving of the investigation undertaken by this study.

Gap in the Literature

While the problem of data quality has been extensively studied and the dimensions of data quality have been generally agreed upon by the academic community, only the individual effects of these dimensions have been analyzed (Blake & Mangiameli, 2011).

Farhangfar et al. (2008) studied the effects of missing values on classification error, but the impact of missing values brought about by data masking and encryption on mining results was not assessed.

As one can infer from the previous sections, while a significant amount of literature has focused on general database quality metrics, the research on the specific causal effect of the use of data masking and encryption on data quality has received little attention. When one dives down to the level of their effect on knowledge discovery, there appears to be no significant research on this specialized topic. This gap in the body of knowledge might be attributable to a number of factors. First, data quality in the context of databases and data warehouses is a mature subject area. Data mining on the other hand is a newer technology that is just starting to become widely used in the scientific community and in industry. Early adopters of the technology, some of which are significant players, had the luxury of using vast amounts of publicly available data that at least until now, has not been subject to protection policies and regulations requiring masking and encryption. Second, industries who process sensitive data are just now starting to exploit data mining. Prokosch and Ganslandt (2009) found that, while the use of electronic medical records was widespread in the United States and Europe, mining these is still in its infancy due to technological, structural, and procedural aspects. Third, there appears to be a general lack of understanding of what constitutes quality as a measure of knowledge discovery. While many data quality dimensions have been postulated, putting these in the context where they could be consistently measured, still needs to be developed.

Analysis of Research Methods

Even though data mining is a relatively new subject area, past research provides rich examples of validated methodologies employed to study this discipline. Many of the methods utilized included techniques such as classification and association rule mining, vector machines, decision trees, instance-based learning, Naïve Bayes classifiers, clustering, numeric prediction, outlier detection, ranking weight decision tables, and data labeling techniques. Wilson and Rosen (2003) used classification algorithms to measure accuracy across various sets of perturbed and non-perturbed data. Other methods used by researchers focused on the process of data mining and not on the specific quality aspect of knowledge discovery. An example of this is the study of association rule mining conducted by Mutter, Hall, and Frank (2004). Their work found that, as a technique used to uncover relationships between data sets across large data repositories, association rule mining lacked quality parameters. They proposed the use of confidence-based measures to instill higher assurance and quality in discovered associations between data sets. Another example includes Sheng, Provost, & Ipeirotis (2008) research of repeated labeling techniques, which found them to improve data quality. Yet another case where data labeling was used as a research method was Iyer (2013) study of online streaming data using data quality tags.

Synthesis

The objective of the literature review was to assemble available academic works on the subject under study into a comprehensive framework that addressed the body of knowledge from multiple angles (Levy & Ellis, 2006). The literature review process provided a structured mechanism to select, compare, and contrast related concepts and

ideas expressed by previous researchers. The chronological review of the literature on data quality and data security practices began to tell a story of how these distinct areas of research developed over the years and at times crossed paths.

When assessing the research on these topics over the last 20 years in aggregate, it becomes apparent that certain areas have been the focus of more attention than others. Researchers have shown a particular interest for defining the concept of data quality. The notion of multiple dimensions has been postulated by many, and this conceptualization has led to the definition of both objective and subjective parameters. However, data quality has been defined in the context of the unit of record itself, and not from the larger standpoint of knowledge acquisition. The relative novelty of data mining as an applied scientific and business tool can also be said to be a reason why focus research in this area is lacking maturity.

At the other end of the spectrum, data security has received less attention as a subject of academic research. The story that one can put together from the literature associated to data security is not as clear. Its goals are not singular, and different works appear to focus on divergent areas. A clear story, as is the case with the research on data quality, cannot be drawn from the available body of knowledge. Possible reasons for this phenomena could be the perception that the subject is more a matter of business rather than academic inquiry.

Notwithstanding these observations, there are select areas where the two subject have crossed paths and the two problems have received at least a passing mention. One of these areas is privacy, and how increased concerns over data confidentiality led to increasing use of data protection technologies resulting from stricter data security policies

and regulations in the business environment. Taken as a whole, a synthesis of the literature led one to conclude that while data quality has been extensively studied, and the impact of security policies and regulations is recognized, the connection between specific data protection techniques such as data masking and encryption, particularly in the area of knowledge discovery, is a field ripe for new research.

Summary

An initial assessment of academic research on the closely defined topic of data quality and knowledge discovery uncovered over 50 peer-reviewed works. Only works dealing with data quality and works addressing security policies and regulations in the context of knowledge discovery and data mining were selected. Works that did not fit within the scope of the research question were excluded from the review.

The literature review analyzed and compared previous research, and divided them into works addressing data quality and those addressing security policies, regulations, and data protection. Within this framework, the works were examined for their strong and weak aspects, and the gaps in the literature became apparent. The effects of data protection on the quality of knowledge discovery has not been an area of research that has received a great deal of attention. While many of the other works quoted in this literature review studied related topics to the main focus of this dissertation, they took different angles and none researched the specific impact of now popular data security techniques such as masking and encryption.

Bringing this gap in the body of knowledge to the surface, the literature review also provided an overview of valuable and reliable research methods previously utilized and

validated to study data quality. Examination of these proven methods provided insightful ideas on how to execute the dissertation work.

Chapter 3

Methodology

Overview

This research study examined the causal effect of data masking and encryption on the performance of classification algorithms. The research was conducted as an experimental test, using benchmark medical data. A benchmark is “a well-defined testing methodology based on real-world use of a computer system” (Scalzo, Burleson, Fernandez, Ault, & Kline, 2007, p.19). Benchmarking provides a practical resource, allowing the researcher to use representative data, otherwise not accessible in the natural environment. The experiment measured the performance impact that data masking and encryption of demographic attributes had on the classification algorithms’ ability to predict patient survival. Multiple parameters were used to measure algorithm performance.

The predictor variables included common attributes that could be used to re-identify the individual patients represented in the sample population. Patient health records can be used for research purposes only after de-identification and removal of personally identifiable information (U.S. Department of Health and Human Services, Health Insurance Portability and Accountability Act, 2015). While data de-identification involves the purging of distinct attributes that can be directly attributable to the individual such as name, personal identification number, and address, other non-distinct attributes can also be used to single out and re-identify records from a select population. Sweeney (2002) estimated that 87% of the population of the United States could be re-identified using attributes not normally classified as sensitive. Narayanan and Shmatikov (2007)

demonstrate how identities could be determined using non-sensitive, non-distinct attributes. In this research, the non-distinct attributes used as predictor variables included marital status, race/ethnicity, Hispanic origin, and year of birth. The dependent/outcome variable was patient survival beyond five years from diagnosis of the chronic disease.

The parameters used to measure the classification algorithm's performance included accuracy, precision, and recall. The harmonic mean between precision and recall, also known as the f-measure, was also used. This metric combined precision and recall into a single representative ratio for ease of analysis (Bhuvanewari, Prabakaran, & Subramaniaswamy, 2015). Additional derivative metrics including the weighted accuracy, receiver operating characteristic (ROC), and area under curve (AUC) were used to calculate, visualize, and compare algorithms' performance, as presented by related works found in the literature (Bradley, 1997; Fawcett, 2006). Derivative metrics were calculated from the initial true positive (TP) and false positive (FP) rates measured by each classification algorithm. In the study, the parameters were used to graphically illustrate relative performance competencies of the algorithms. A detailed description of each of these metrics is included in the Data Analysis section in this chapter.

Data

The data benchmark used was the National Institutes of Health (NIH) National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) Program database. The complete database contains over 9 million cancer patient records collected across nine geographical areas in the United States, and extends over a 30-year period (SEER, 2015). The areas represented include a cross-section of the U.S. population's race and ethnic backgrounds. As a national resource, the SEER is the most comprehensive, up to

date, and accurate repository of cancer incidence and survival statistics available to researchers (Cox, 1984; Duggan, Anderson, Altekruze, Penberthy, & Sherman, 2016; Hankey, Ries, & Edwards, 1999). No direct human subjects were contacted in the study. However, real patient records contained in the benchmark data set were used. The records documented clinical factors related to patient demographic, treatment, and survival statistics.

Nine text files, formatted in the American Standard Code for Information Interchange (ASCII), were contained in the database. Each data set corresponded to a different type of cancer. The SEER breast cancer data set provided the right parameters for masking and encrypting select non-distinctive attributes. Although the SEER breast cancer data set was already de-identified, the personal attributes it contained held the potential for it to be used to re-identify subject patients and compromise their privacy. Because of this characteristic, the data set proved to be particularly useful for the research study.

A total of 769,261 instances were present in the raw breast cancer data set. The number of instances was subsequently reduced to a balanced sample of 100,000 to ensure the internal validity and confidence in the experimental setting. A total of 121 attributes were initially contained in the SEER breast cancer raw data set. These included categorical, numeric, and string attributes including patient personal details such as marital status at diagnosis, race/ethnicity, Hispanic origin, and year of birth. These attributes were used as the predictor variables. The data set was used to train multiple supervised classification algorithms to predict patient survival. Classification, along with regression, clustering, and association rule mining, are methods used in predictive analytics (Fletcher & Islam, 2015). By measuring the performance of the classification

algorithms' ability to accurately predict the outcome of the dependent variable when the predictor variables were subject to treatment, enable the determination of the effect of masking and encryption on the quality of knowledge discovery.

Knowledge Discovery Process

The research followed the Knowledge Discovery in Databases (KDD) conceptual model for applying data mining to practical scenarios. The process delineates sequential steps that enable the extraction of hidden knowledge from vast amounts of data (Fayyad et al., 1996). The KDD is one of three data mining models used by practitioners and researchers, along with Cross-Industry Standard Process for Data Mining (CRISP-DM), and the Sample-Explore-Modify-Model-Assess (SEMMA) process. Azevedo and Santos (2008) studied these standards in data mining and found that CRISP-DM and SEMMA could be considered implementations of the KDD process. However, Azevedo & Santos (2008) concluded that CRISP-DM and SEMMA may not fully embody all the steps delineated in the KDD. According to Shafique and Qaiser (2014), the KDD offered a more complete and accurate model for carrying data mining exercises given its iterative and interactive nature. For this reason, KDD was used to answer the stated research questions and to test the postulated hypothesis. Fayyad, et al. (1996) defined five stages in the KDD model:

1. Data selection
2. Data preprocessing
3. Data transformation
4. Data mining
5. Data interpretation and evaluation

The implementation of this general model in the execution of this study is outlined in Figure 2.

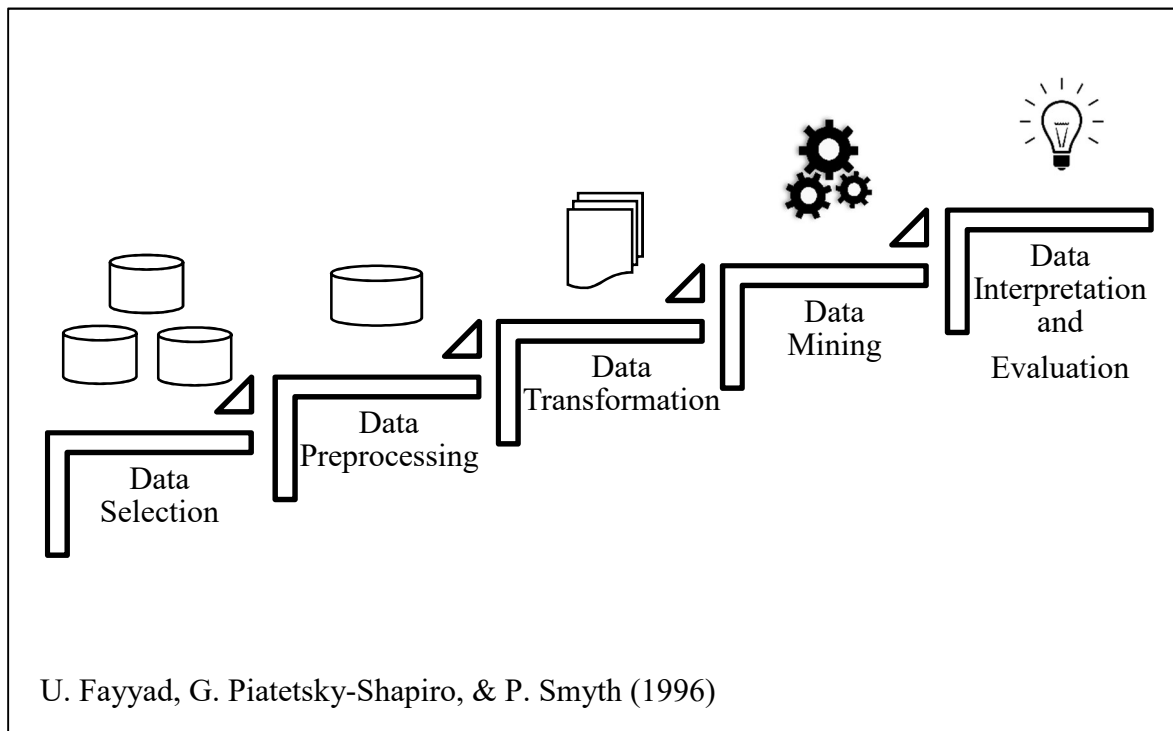


Figure 2. KDD Model Followed in Research Study

Data Selection

Following the KDD process, the first step undertaken in the study involved the selection of data for analysis and observation. Typically, it is important to assess the validity and reliability of instruments used to collect data for experimental purposes. However, since an existing benchmark was used for the study and no data collection was conducted, this assessment was not necessary. Having selected the SEER breast cancer benchmark data set, an end user agreement was filed with the NIH/NCI to gain access to the resource.

Data Preprocessing

The second step in the study involved preprocessing of the raw SEER breast cancer data to enable proper reading by a data mining tool. The process of preparing data for mining is critically important and can be very time consuming (Cios & Moore, 2002). In this study, the preprocessing phase examined the data for consistency and completeness using a method similar to one employed by Prandini, Campi, Marzolla, and Melis (2014). This method ensured that only instances with common characteristics were used, and that numeric and string attributes were discretized prior to mining. Of the total 121 attributes listed in the original SEER breast cancer data set, 79 were removed as they were only recorded during specific years and were not common across the entire population. In addition, 22 indexing attributes such as month and year of diagnosis were removed. Other attributes determined by the SEER Research Data Record Description not to have complete coverage of the cases diagnosed across the sample period were also removed. An excerpt from the SEER Research Data Record Description is included in Appendix A. Using this process, the number of attributes in the data set was reduced to 18 plus the

classification attribute (survival). An additional preprocessing step included the removal of instances that contained incomplete data. An example of this was the use of codification "999" for the classification attribute to indicate unknown survival.

Ideally, training and testing of classification algorithms and models should be performed using data set samples selected from an infinite population. For this research, a sizeable benchmark containing over 700,000 instance was used. From this benchmark, the research study used a sample size of 100,000 instances. The reduction was done for two reasons. The first was to balance the sample of the outcome classification label (i.e., survival) to remove biases. The second, was to avoid overfitting that could have resulted from a sample skewed in favor of one of the outcome alternatives.

Once the sample was balanced, it was important to observe the classification or predicted value to validate the population split. This was done by using the simple classifier ZeroR to establish the baseline for comparison. ZeroR focuses on classification, and does not examine other attributes. The outcome classification variable (survival), which recorded the number of months that patients lived beyond diagnosis, was then set to a binary nominal value. Consistent with previous research on cancer patient survival using period analysis, patients who survived the disease were selected as those that were still alive after 60 months from initial diagnosis (Brenner, Gefeller, & Hakulinen, 2002; Cox & Oakes, 1984; Delen, Walker, & Kadam, 2005; Rosenberg, Chia, & Plevritis, 2005). Values below 60 month were set to a categorical NO, and values over 60 months were set to a categorical YES. Records were then sorted using the binary value of the classification variable, and a balanced sample of 100,000 records selected using the first 50,000 instances with value YES and last 50,000 with value NO.

To also ensure that the data mining tool would be able to read the data set correctly and not confuse indexing values with input variables, additional modifications were made before running the algorithms. The selection of the sample data set from the original benchmark is illustrated in Figure 3 and described in detail in Appendix C.

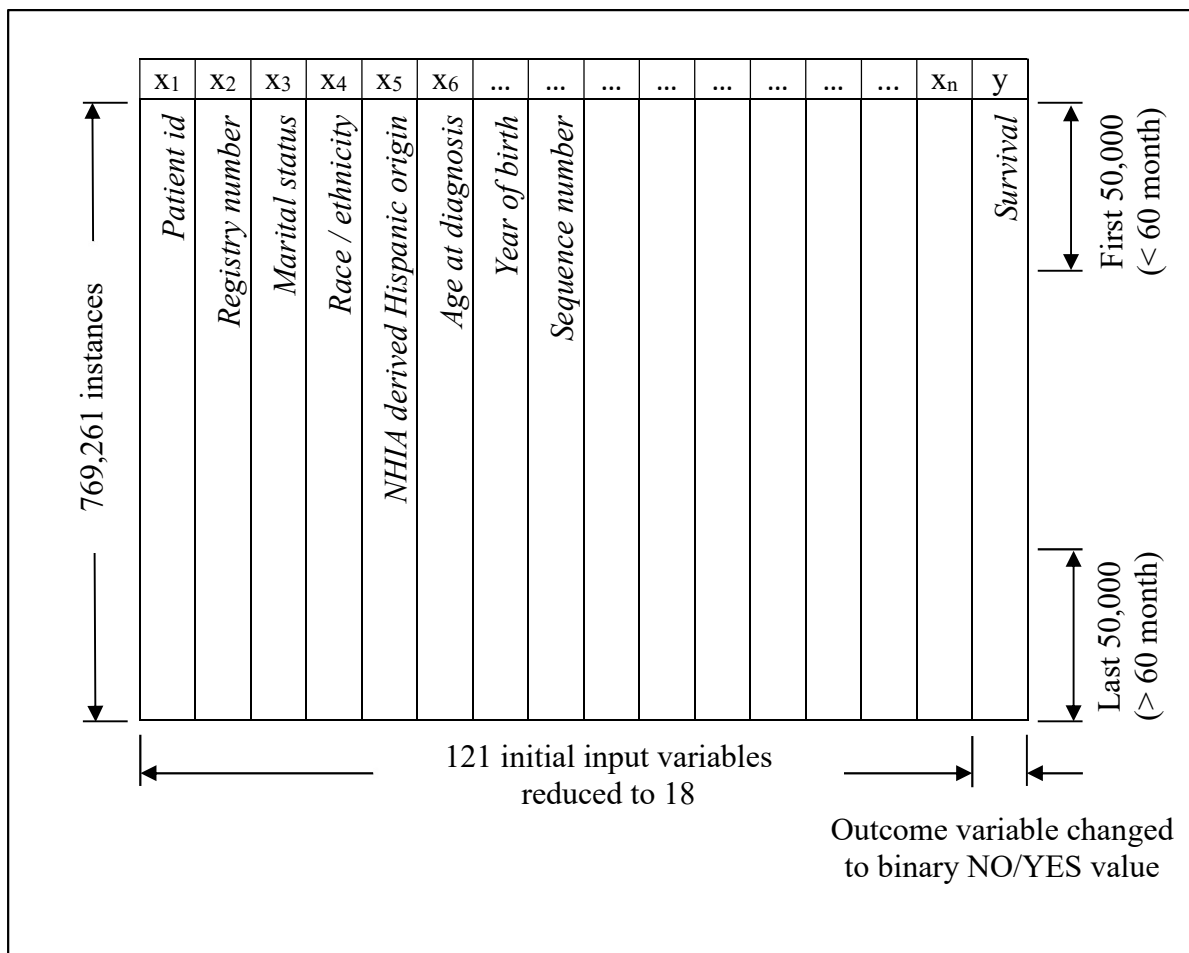


Figure 3. Raw Benchmark Data Set from which Balanced Sample was Derived

Data Transformation

Step three of the KDD process used in this study transformed the balanced data to prepare it for the mining and classification performance analysis. During data transformation, missing values were replaced with the mode or most frequently appearing value for the respective predictor variable attributes in the data set. The values were obtained from the SEER Data Record Description Summary. The number of attributes was also further reduced to the top 12 most influential ones on classification outcome. Measuring the information gain of each attribute relative to classification, the attributes were listed in order of merit as shown in Figure 4. Details of the process undertaken to determine information gain using the data mining tool are included in Appendix B. The fact that all four of the critical attributes under study (i.e., marital status at diagnosis, race/ethnicity, Hispanic origin, and year of birth) were found by the filter to be influential in predicting the outcome variable confirmed their importance and relevancy.

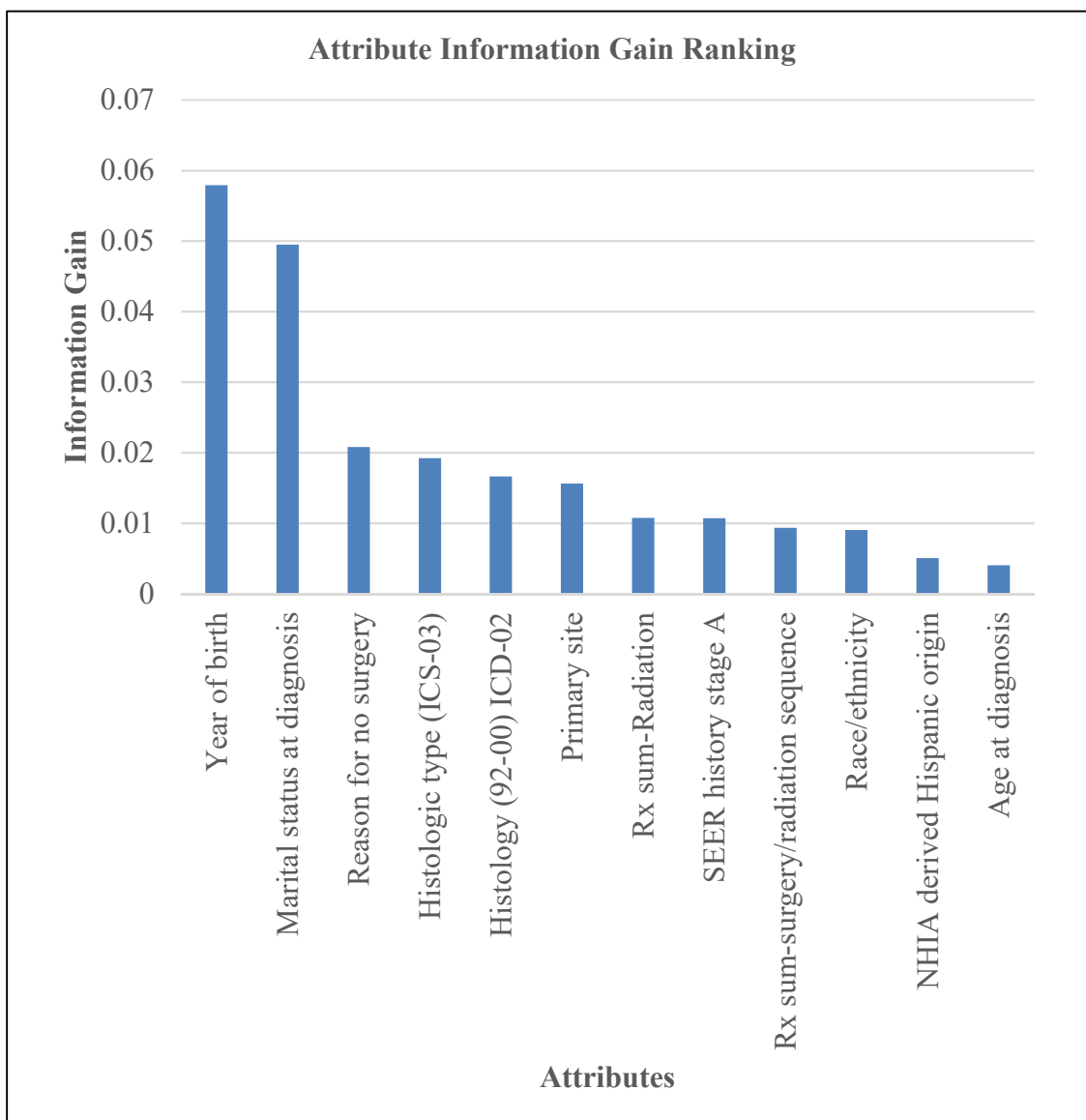


Figure 4. Ranking of Most Influential Attributes on the Outcome Variable (Survival)

The resulting data set of 100,000 instance and 13 attributes (i.e., 12 plus classification variable) produced the control group used in the experiment. Using additional filters in the mining tool, the experimental data sets were then derived.

The control group was representative of the data set not subjected to treatment of select attributes. In this group, predictor variables were not subject to replacement or suppression.

The experimental groups were representative of the data sets subjected to treatment. Two of these groups were created, a group subjected to the effects of data masking, and a group subjected to the effects of data encryption. Both experimental groups were based on the same control group data, and contained the same original number of instances. Data masking engines can be static or dynamic, and typically substitute select sensitive attributes with either constant or randomly changing values (IBM Knowledge Center, 2016). The research study used static data masking, replacing selected attributes with constant values based on commonality of attributes in the data set. The group subjected to encryption had the select attributes removed. Removal of the attributes emulated the use of an encryption engine (Motiwalla & Li, 2013). The emulation process was based on one postulated by Bhuvanewari, Prabaharen, and Subramaniaswamy (2015). The derivative transformation of the control group data set into the respective experimental masking and encryption data sets is illustrated in Figure 5.

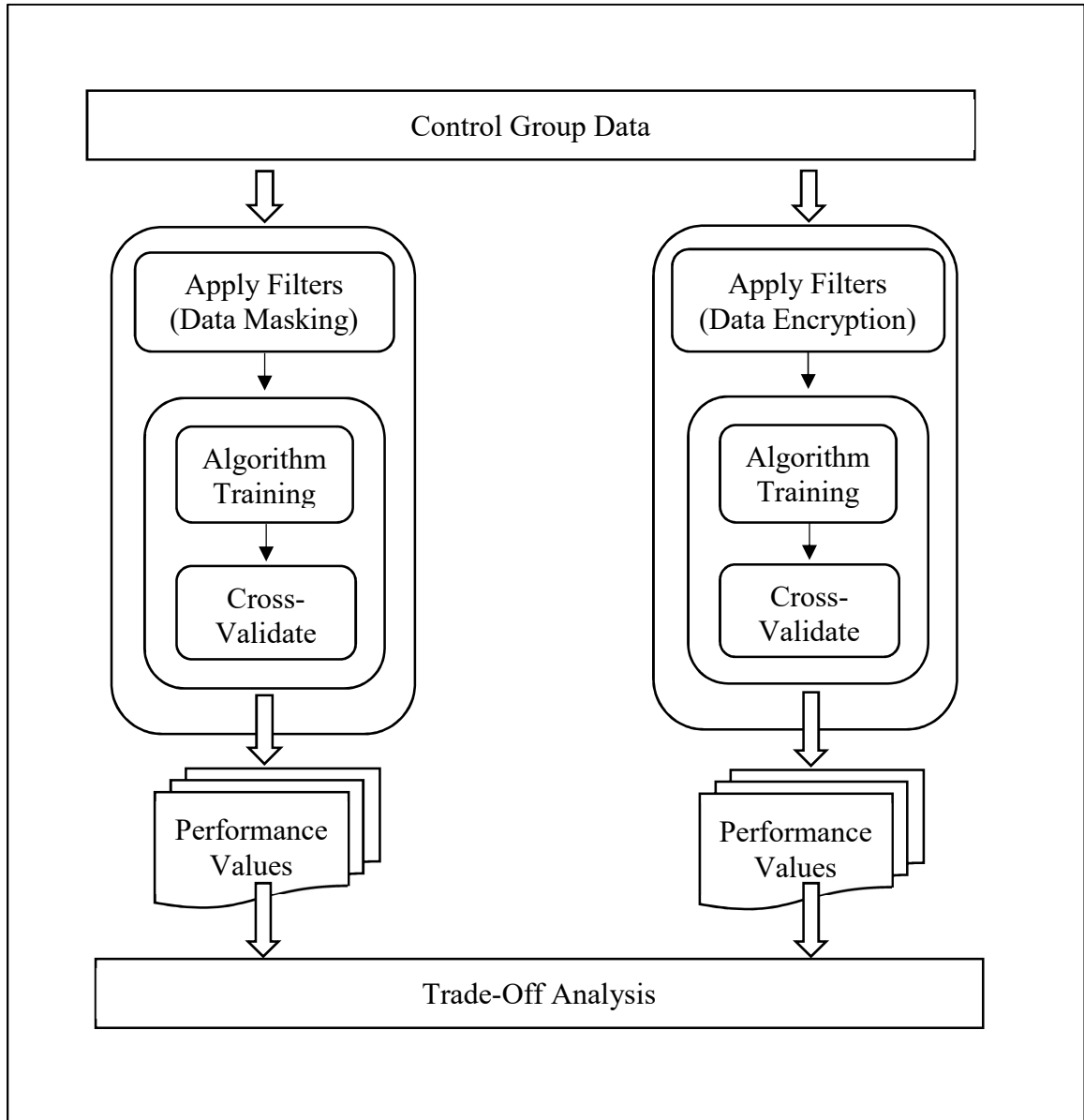


Figure 5. Data Transformation Process Model Used in Research Study

According to Xue, Zhang, and Browne (2012), filters perform feature selection to find the attributes that are most useful to the classification process. Data transformation used the mining tool's filtering and wrapping capabilities to remove and consolidate the data into a form that was easier for the classifiers to process. Filtering techniques include the purging of irrelevant attributes and aggregation of other attributes using discretization (Prandini, Campi, Marzolla, & Melis, 2014). Removal of attributes is justified when they show no significant difference in values (e.g., attributes that have same value over 95% of the all instances). When such conditions are seen, attributes can be removed to provide a more compact and manageable data set to work, without impacting the outcome variable. An example of this condition in the research study was the patients' sex, given that over 90% of all breast cancer patients were female. Wrappers search for the best subset of attributes using the classification algorithm(s) being employed to deliver a feature selection specific to the algorithm(s). Wrappers are general considered to deliver better results than filters, but they are more computationally intensive (Dash & Liu, 1997). For this reason, the filters offered by the data mining tool were used.

Finally, another important data transformation step that needs to be taken when using certain classification algorithms is to discretize the data. Discretization takes all possible values of a numeric or string attribute and categorizes then into sub-set states to reduce the total number of values that the attribute can have. Attributes such as age at diagnosis were also discretized. In the research study, discretization was used when employing the Naïve Bayes algorithm. A summary of the steps taken to select, preprocess, and transform the initial raw data to create the control and experimental data sets for mining is shown below in Figure 6. The complete experimental setup, including how the data

mining tool was configured, and how the various filters were enabled to create the control and experimental data sets is described in detail in Appendix C.

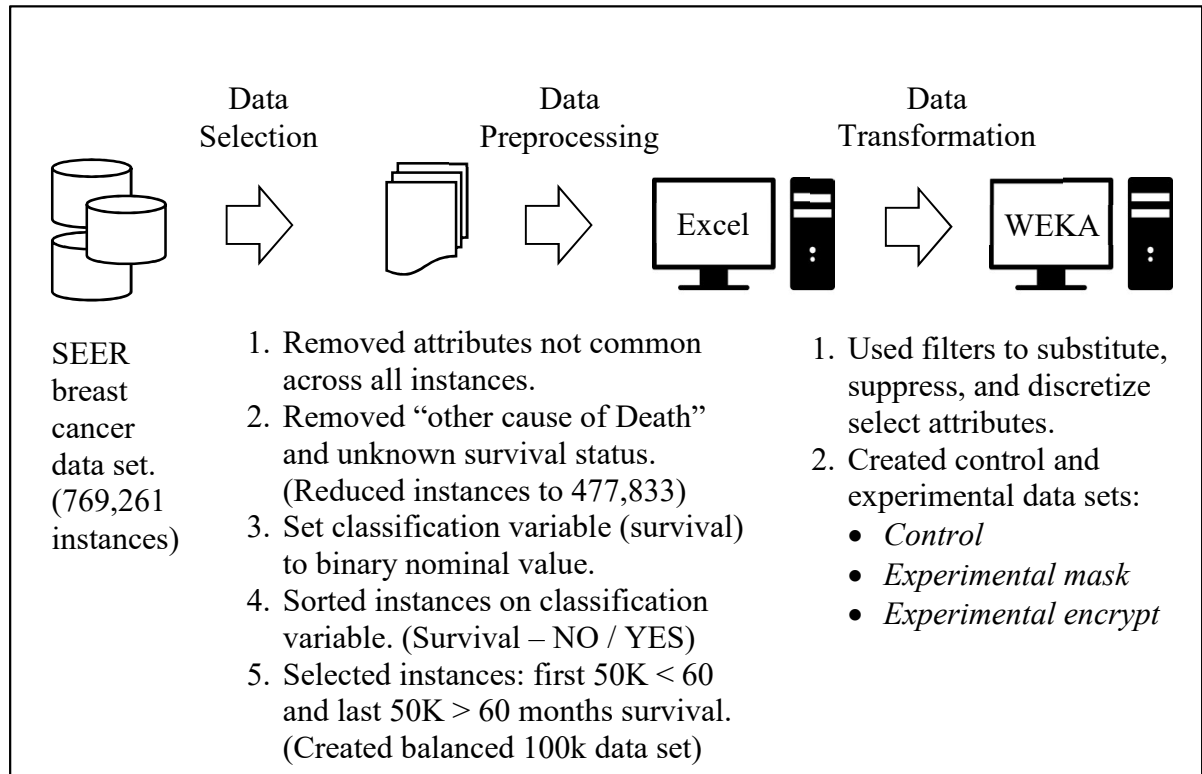


Figure 6. Schematic Representation of Data Preprocessing and Transformation Process

Data Mining

Step four in the KDD process used the data mining tool to run the classifiers on the control and experimental groups. Supervised algorithms were used in the research study. Supervised algorithms are those that calculate an outcome employing a learning or training process (Shmueli, Patel, & Bruce, 2010). The baseline and the four supervised classification algorithms used in the study included:

- ZeroR
- J48
- Naïve Bayes
- Adaptive Boosting
- Random Forest

ZeroR is a simple classifier used to predict the majority category in a data set. In this study, ZeroR was used to establish the balanced samples. The J48 classification algorithm is an implementation of the C4.5 decision tree-based classifiers used in early data mining tools (Quinlan, 1996). Naïve Bayes is a probabilistic classifier, and Adaptive Boosting (AdaBoost) and Random Forest are ensemble meta-learners that uses multiple training inputs to arrive at an aggregate prediction metric. A description of each of the algorithms is included in the Algorithm section in this chapter.

With the data transformation phase completed, the data mining tool was then used to train and cross-validate the algorithms, and to predict patient survival. The configuration of the parameters set for each of the classifiers is shown in Appendix D. The execution of the classifiers across the control and experimental groups produced respective performance parameters used to assess and compare the effect of data masking and

encryption. Values obtained and the corresponding confusion matrixes and ROC/AUC graphs were captured in Chapter 4, and are also presented in Appendix D, F and G.

The concept of information gain, or the amount of useful information contained in the data set was used to select the best attributes. The code that defines the steps in the processes was described by Patil and Sherekar (2013) and is shown in Figure 7.

Information Gain – Steps and Pseudo Code:

1. Check for base cases, and for each attribute a :
2. Find feature that best divides the training data (information gain).
3. Let a_{best} be attribute with highest information gain.
4. Create decision mode that splits on a_{best} .
5. Repeat process on sub-list obtained when splitting a_{best} .
6. Add intermediate nodes and leafs under each instance.
7. Stop when conditions are met and all nodes are classified.

Figure 7. Code Defining Steps Taken to Determine Information Gain

Using this process, attributes that had the most significant impact on the outcome variable (i.e., survival) were identified. The subset of the 12 most influential attributes was then used in the experiment. The complete attribute evaluation and ranking process undertaken is described in Appendix B. By learning how these conditions produce different results, the algorithm was then able to predict resulting values and classification performance. Comparing the performance metrics obtained for each of the algorithms before and after applying treatment, enabled the determination of trade-off between security of the medical records (confidentiality) and the accuracy of the predicted value (utility).

Interpretation and Evaluation

The fifth and final step in the KDD process included the interpretation and evaluation of results of the data mining exercise. The performance values obtained from each of the algorithms was compared and contrasted between the control group and experimental groups. Performing associated algorithmic runs for the control and experimental groups, observations on the impact of attribute treatment were recorded. A profile of the impact brought about by masking and encryption was drawn, and an analysis for statistical significance conducted.

Sample

The data sample used included a cross section of breast cancer patients in the United States collected between 1973 and 2013, making it a representative sub-set of the general population. Sekaran and Bougie (2013) recommended that experimental studies use a minimum sample of the affected population to ensure that cause and effect relationships being studied are widely represented at least 95% of the time. Given that a recognized sample was already contained in the benchmark, a non-probability convenience sampling

approach was used in the research. While convenience sampling does not allow for the inference of result, it is the best method available for gaining an understanding of the dynamics that surround a research question (Sekaran & Bougie, 2013). Convenience sampling uses an accessible subset of the population that is easily reached. Such was the case with the SEER database used. However, since the study employed records already contained in the SEER database, the subjects could not be considered random participants. To balance the sample, the ZeroR classifier was used to determine the split in predicted outcome. Best practices in machine learning call for establishing a baseline before training classifiers and creating the data mining model (Witten, Frank, & Hall, 2011).

Variables

The research study observed the relationship between four nominal independent (predictor) variables and a single nominal binary dependent (outcome) variable. The observations compared performance values obtained between the control group and two experimental groups. The four predictor variables included:

- Marital status at diagnosis
- Race/ethnicity
- Hispanic origin
- Year of birth

These predictor variables were a subset of a larger feature set found in the SEER database which includes additional demographic characteristics. Hispanic origin was derived from the North American Association of Central Cancer Registries (NAACCR) Hispanic Identification Algorithm (NHIA). The dependent/outcome variable was a binary measure

of survival. A schematic representation of the theoretical framework of the research depicting the relationship between the substituted or suppressed attributes and the resulting impact on classification performance as a measure of quality of knowledge discovery is shown in Figure 8.

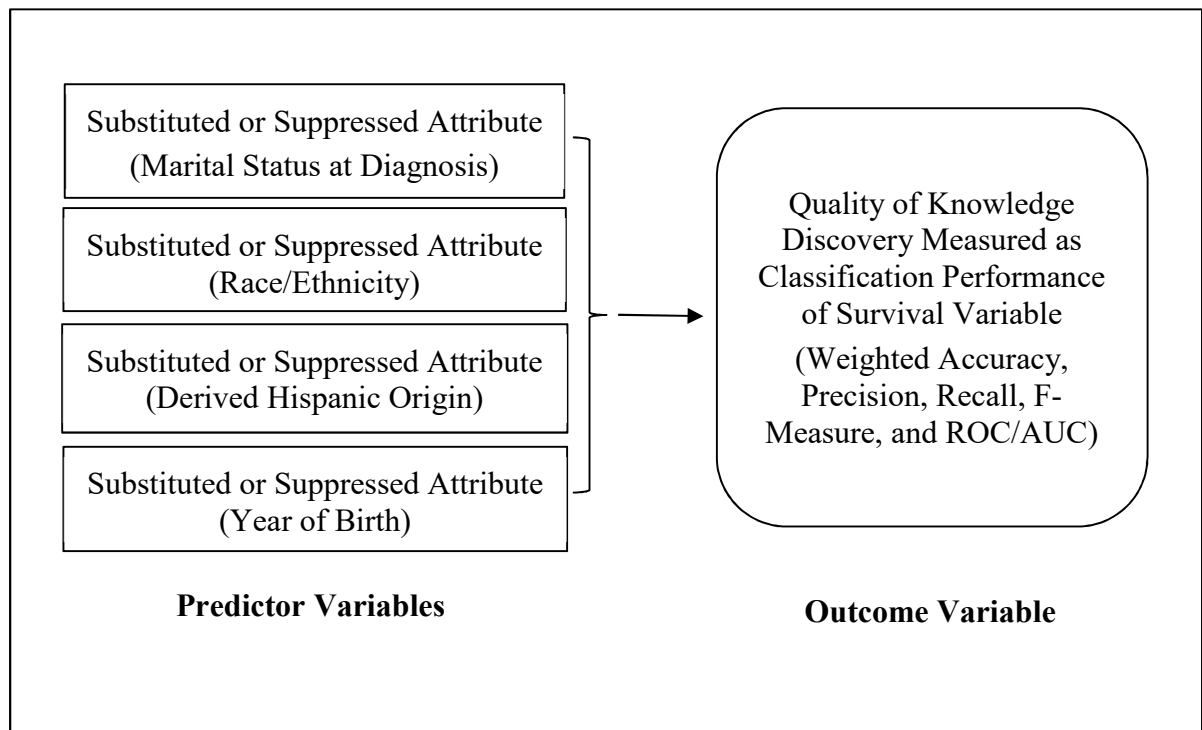


Figure 8. Theoretical Framework of Research Scenario

Research Method

The selection of the appropriate research method took into account how the data was collected and analyzed. Given that a pre-recorded benchmark was used, and that data was not collected from the natural environment, the research was carried out as a laboratory experiment. Laboratory experiments enable researchers to investigate cause and effect relationships between independent and dependent variables in a controlled environment that removes external factors that can skew observations (Sekaran & Bougie, 2013). Since the study used pre-recorded medical records and random selection was not possible under the regime, the quasi-experimental method was employed. Quasi-experiments are used to test descriptive causal hypotheses postulated based on potential causes that can be manipulated. Quasi experiments are unique in the fact that they do not assign units to conditions in a random manner like regular experiments. Research conducted using quasi-experiments typically measures the effect of manipulated causes (Shadish, Cook, & Campbell, 2002).

In the research study, the quasi-experiment subjected the select attributes of the data set to treatment by substituting or suppressing their values. Measuring the cause and effect by comparing results obtained from the control group, where the predictor variables was left untouched, against the results obtained from the experimental groups, where treatment was applied to the selected predictor variables, results were then recorded.

While the SEER data set contained already de-identified data to protect the patients' identities, it still maintained demographic attributes that could be used to re-identify subjects. Because of the commonality of these attributes among the sample population,

they represented useful data points for the classification process, and an ideal schema for the research study. Substitution or suppression of these values had a measurable impact on data mining results.

Internal and External Validity

Validity provides a measure that assesses that an experiment not only measures the concept under study reliably, but also that it indeed measures the right concept (Salkind, 2012). Internal validity denotes whether the treatment of the predictor variable(s) has an impact on the outcome or dependent variable, and whether the results can substantiate the observation through statistical evidence. External validity, on the other hand, refers to the extent to which the causal results obtained from the experiment can also be proven to be true in other independent settings. Laboratory experiments generally yield high internal validity given that they are carried out in a controlled environment. However, this specific characteristic can also make them have less external validity. For this reason, it is important to enable other researchers to have the information needed to be able to repeat experiments using other data sets to corroborate cause and effect relationships. Given that this research was carried out as a laboratory experiment, results were expected to yield high internal validity. To also ensure a high level of external validity, the research method and associated processes and procedures undertaken were carefully documented so they could be easily recreated for validation.

The classification algorithms used in the research study constructed the confusion matrixes from which calculated performance values were derived. Confusion matrixes graphically present the foundational elements defining the performance of classification

models (Tan, Steinbach, & Kumar, 2006). A schematic representation of a representative confusion matrix is shown in Figure 9.

<table border="1"><tr><td>A</td><td>B</td></tr><tr><td>TP</td><td>FP</td></tr><tr><td>FN</td><td>TN</td></tr></table>	A	B	TP	FP	FN	TN	← (classified as) A = No B = Yes
A	B						
TP	FP						
FN	TN						

Figure 9. Confusion Matrix and Precision Calculation

The diagonal sums of the values in a confusion matrix yields the number of correct and incorrect attribute instances (Salama, Abdelhalim, & Zedi, 2012; Patil & Sherekar, 2013). This sum of the true positives TP and true negatives TN divided by the sum of the false negatives FN and false positives FP is used to derive the precision or degree of separation or closeness between the various measures:

$$P = [\sum \text{diagonal elements}] / [\sum \text{relevant column}] \quad 1$$

As defined in Equation 1, confusion matrixes are useful in assessing the precision of the classification process. Using the confusion matrix, the accuracy of the classifier can also be determined by calculating the ratio of correctly classified attributes in a data set (Patil & Sherekar, 2013).

Precision denotes the proportion of predicted positives cases that are correct (Powers, 2011). The metric is often called confidence in data mining. Precision defines the fraction of retrieved instances during a data set search that are relevant, and the closeness between the different classification accuracy measurements recorded to determine how useful they actually are (Patil & Sherekar, 2013). Recall is often referred to as sensitivity in social sciences, and is equivalent to the true positive rate (Hu, Li, Plank, Wang, & Daggard 2006; Kumari & Godara, 2011; Powers, 2011; Tan, Steinbach, & Kumar, 2006). In the study, precision and recall were calculated using Equations 2 and 3.

$$\text{Precision} = TP / (TP + FP) \quad 2$$

$$\text{Recall} = TP / (FN + TP) \quad 3$$

To measure the trade-off between precision and recall, the f-measure or harmonic mean was calculated as shown in Equation 4.

$$F_1 = 2 \times TP / 2 \times TP + FP + FN \quad 4$$

The f-measure is used in machine learning to evaluate the performance of an algorithm when multiple metrics are employed (Powers, 2011). When an algorithmic model is built, it often maximizes one metric over another, and the f-measure attempts to harmonize the two values to deliver a single metric from which the algorithmic model's performance can be assessed. Higher f-measure values are indicative of models with high precision and recall.

Another metric used in the research study to normalize the values obtained for the algorithmic models was weighted accuracy. Weighted accuracy determines the relative weights of the TP, TN, FP, and FN using Equation 5.

$$\text{Weighted Accuracy} = [w_1TP + w_4TN] / [w_1TP + w_2FP + w_3FN + w_4TN] \quad 5$$

Weighted accuracy takes into account the relative weights (w_i) of each of the confusion matrix's components, and delivers a balance value that better represents the conditions observed in the data (Tan, Steinbach, & Kumar, 2006).

To evaluate the relative performance of the data mining classification algorithms used in the study, ROC graphs were built. ROC graphs are a useful tool for visually comparing the performance of different classifiers (Fawcett, 2004). Since accuracy, precision, and recall are not generally cost-sensitive analytical metrics, the use of ROC graphs in the medical sciences has become increasingly popular as a way to substantiate measurements and provide a cost analytical base (Powers, 2011). ROC graph offer visual interpretation of classifier performance, and are particularly useful in diagnostic decision-making (Provost et al., 1998). By plotting classifiers' potential cost or FP rate, measured as the

AUC, against potential benefit or TP rate, interpretation of their suitability for a particular task can be made more convincingly. A typical ROC graph is shown in Figure 10.

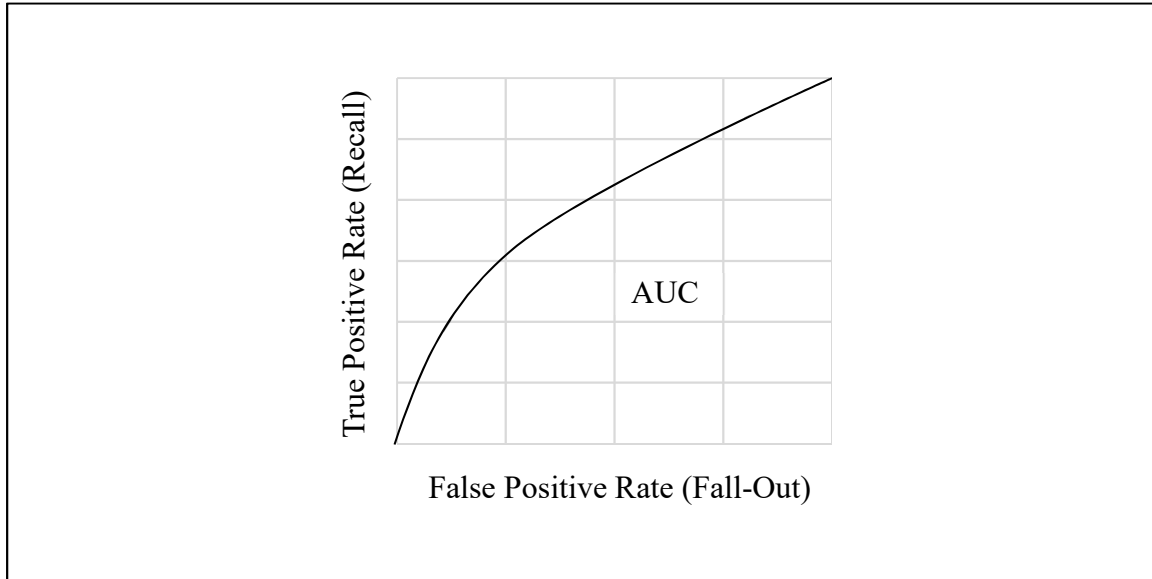


Figure 10. Representative ROC Graph Depicting Potential Costs as AUC

The closer the classifier falls to the lower left of the graph, the origin of the axes, the least FP and least TP predictions they make, and the more risk-averse the classification algorithm. Classifiers that fall on the upper right of the plot take more risks in making predictions. The more cost-effective classifiers will be found in the upper left of the plot (Fawcett, 2004). To compare different classifiers, the AUC is calculated to provide an easy scalar value representing associated performance (Fawcett, 2004; Fogarty, Baker, & Hudson, 2005; Hastie, Tibshirani, & Friedman, 2009). The larger the area under a classifier's curve, the better the classifier is against the other (Tan, Steinback, & Kumar, 2006). In the research study, ROC graphs were built for each of the classifiers employed in the study and overlaid to visualize their relative suitability in predicting the outcome variable. An illustration showing the overlay of ROC graphs for each of the classifiers is shown in Chapter 4. Individual ROC graphs produced by the data mining tool for each classifier are shown in Appendix E.

Algorithms

The research used a simple baseline classifier, ZeroR, and four supervised machine learning algorithms to measure classification performance. J48, Naïve Bayes, AdaBoost Ensemble, and Random Forest have shown to deliver an efficient analytical mechanism to measure classification performance (Al-Bahrani, Agrawal, & Chaudhary; Patil & Sherekar, 2013). The trained classifiers provided an objective metric that could be repeatedly tested with changing conditions, providing an effective way to measure quality of data mining results.

ZeroR is used to predict the majority classification attribute when dealing with nominal values, or the average if dealing with numeric values (Witten, Frank, & Hall,

2010). ZeroR can also be used to establish a baseline (Mazalu, Cechich, & Mart, 2013).

In the research study, the use of ZeroR made a worthwhile addition since it created a reference point from which to compare all other classification algorithms employed.

Sample performance results and associated plotted values obtained when using ZeroR on the control data set are shown in Figure 11.

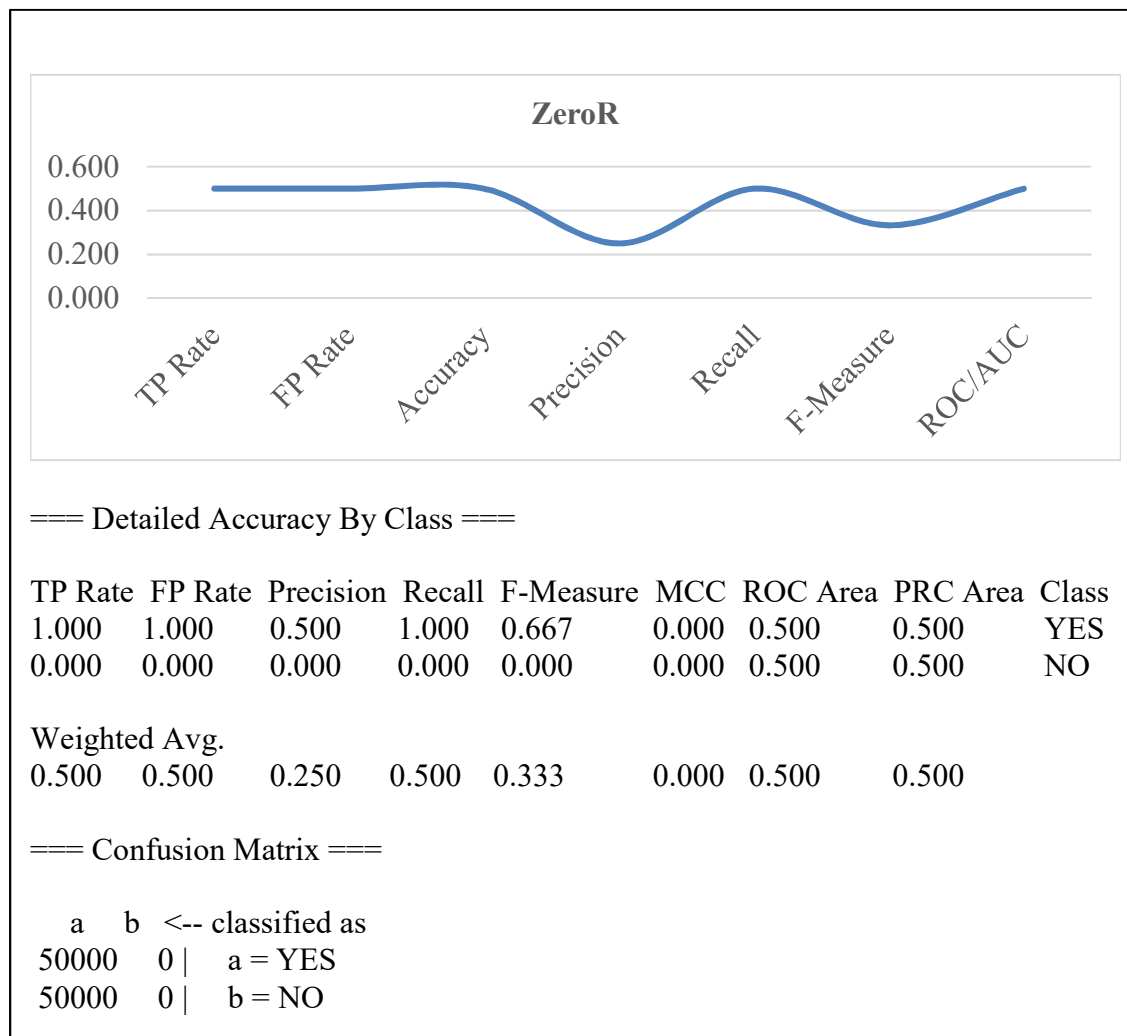


Figure 11. Example of Classification Split and Plotted Values Delivered by ZeroR

The J48 algorithm classified attributes based on their frequency within the binary decision tree construct. The degree of uncertainty of attribute occurrence refers to the overall entropy where each node represented an attribute test. Entropy is therefore the sum of the probabilities of each attribute times the logarithmic probability:

$$E = \sum_{i=1}^c -p_i \log_2 p_i \quad 6$$

The p_i value in Equation 6 represents the frequency of attribute (Bramer, 2007). As a decision tree is built, entropy is determined for each tuple in the data set. The accuracy of the classification process is then measured from the number of correctly classified tuple instances (Fletcher & Islam, 2010). The code that defines the steps in the algorithmic processes undertaken by decision trees was defined by Tan, Steinbach, and Kumar (2006) as shown in Figure 12.

J48 Decision Tree Classifier – Steps and Pseudo Code:

TreeGrowth (E, F)

1. If stopping condition (E, F) = true then
2. $Leaf = \text{Create Node } ()$.
3. $Leaf.label = \text{Classify } (E)$.
4. Return $leaf$.
5. Else
6. $Root = \text{Create Node } ()$.
7. $root.test\ condition = \text{find_best_split } (E, F)$.
8. Let $V = \{v/v \text{ is a possible outcome of } root.test\ condition\}$.
9. For each $v \in V$ do
10. $E_v = \{e \mid root.test\ condition (e) = v \text{ and } e \in E\}$.
11. $Child = \text{Tree Growth } (E_v, F)$.
12. Add $child$ as descendent of $root$ and label the edge ($root \rightarrow child$) as v .
13. End for
14. End if
15. Return $root$.

Figure 12. Framework of Decision Tree Induction Algorithms

Sample performance results and associated plotted values obtained when using J48 on the control data set are shown in Figure 13.

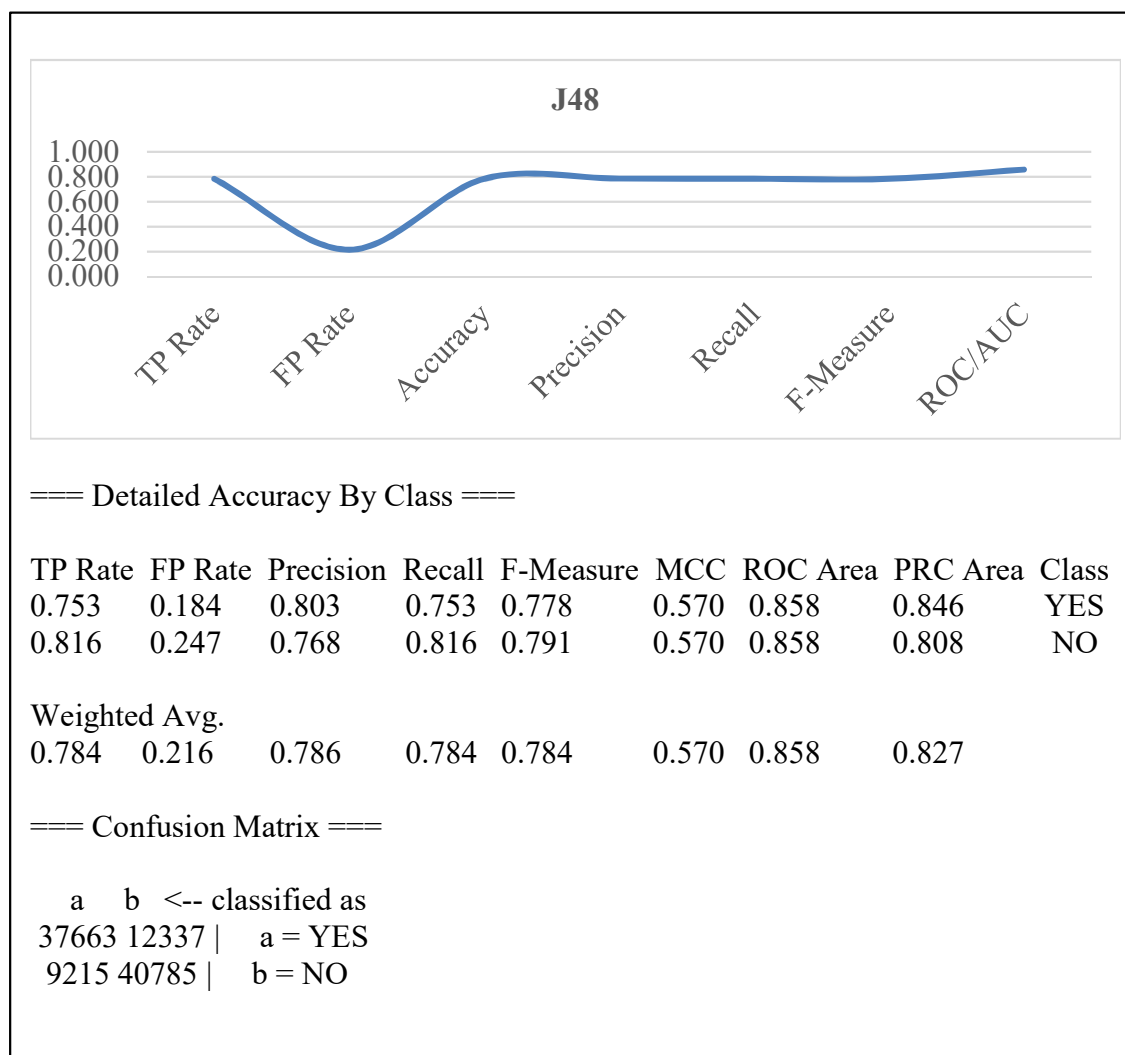


Figure 13. Performance Results and Plotted Values Delivered by J48

Naïve Bayes is grounded on a total probability function based on Baye's theorem that accounts for frequency and value combinations in the data sets. The algorithm makes the independent (naïve) assumption that all attributes contribute equally to a decision (Al-Aidaroos, Bakar, & Othman, 2012). The probability is calculated using the formula:

$$P(h1|xi) = \frac{P(xi|h1)P(h1)}{P(xi)} \quad 7$$

Posterior probability $P(h1|xi)$ in Equation 7 is based on prior probability $P(h1)$, where $h1$ is the hypothesis being postulated, and xi is the instance where it is postulated.

Naïve Bayes' main strengths lie in its ability to perform in noisy environments where there are often missing values. The algorithm has been shown to perform well in medical applications when compared to other algorithms in experimental scenarios (Abraham, Simha, & Iyengar, 2006; Al-Aidaroos, Bakar, & Othman, 2012; Demšar, et al., 2001; Kononenko, Bratko, & Kukar, 1997). Naïve Bayes also takes into account data from all attributes in a data set to predict an outcome variable (Zelič, Kononenko, Lavrač, & Vuga, 1997). As a categorical predictor, in order to enable continuous numeric values and strings to properly work, these are first discretized into categories. Certain data mining tools automatically calculate the mean when numeric values are present and use those values for their prediction.

The code that defines the steps taken by the Naïve Bayes algorithm used in the research was defined by Lowd and Domingos (2005) as shown below in Figure 14.

Naïve Bayes Classifier – Steps and Pseudo Code:

Input: Training set T , hold-out set H , initial number of components k_0 , and convergence thresholds δ_{EM} and δ_{Add} .

1. Initialize M with one component. $k \leftarrow k_0$
2. Repeat
3. Add k new components to M , initialized using k random examples from T .
4. Remove the k initialization examples from T .
5. Repeat
6. *E-step*: Fractionally assign examples in T to mixture components, using M .
7. *M-step*: Compute maximum likelihood parameters for M , using the filled-in data.
8. If $\log P(H|M)$ is best so far, save M in M_{best} .
9. Every 5 cycles, prune low-weight components of M .
10. Until $\log P(H|M)$ fails to improve by ratio δ_{EM} . $M \leftarrow M_{best}$
11. Prune low weight components of M . $k \leftarrow 2k$
12. Until $\log P(H|M)$ fails to improve by ratio δ_{Add} .
13. Execute *E-step* and *M-step* twice more on M_{best} , using both H and T .
14. Return M_{best} .

Figure 14. Code Defining Steps Taken by Naïve Bayes

Sample performance results and associated plotted values obtained when using Naïve Bayes on the control data set are shown in Figure 15.

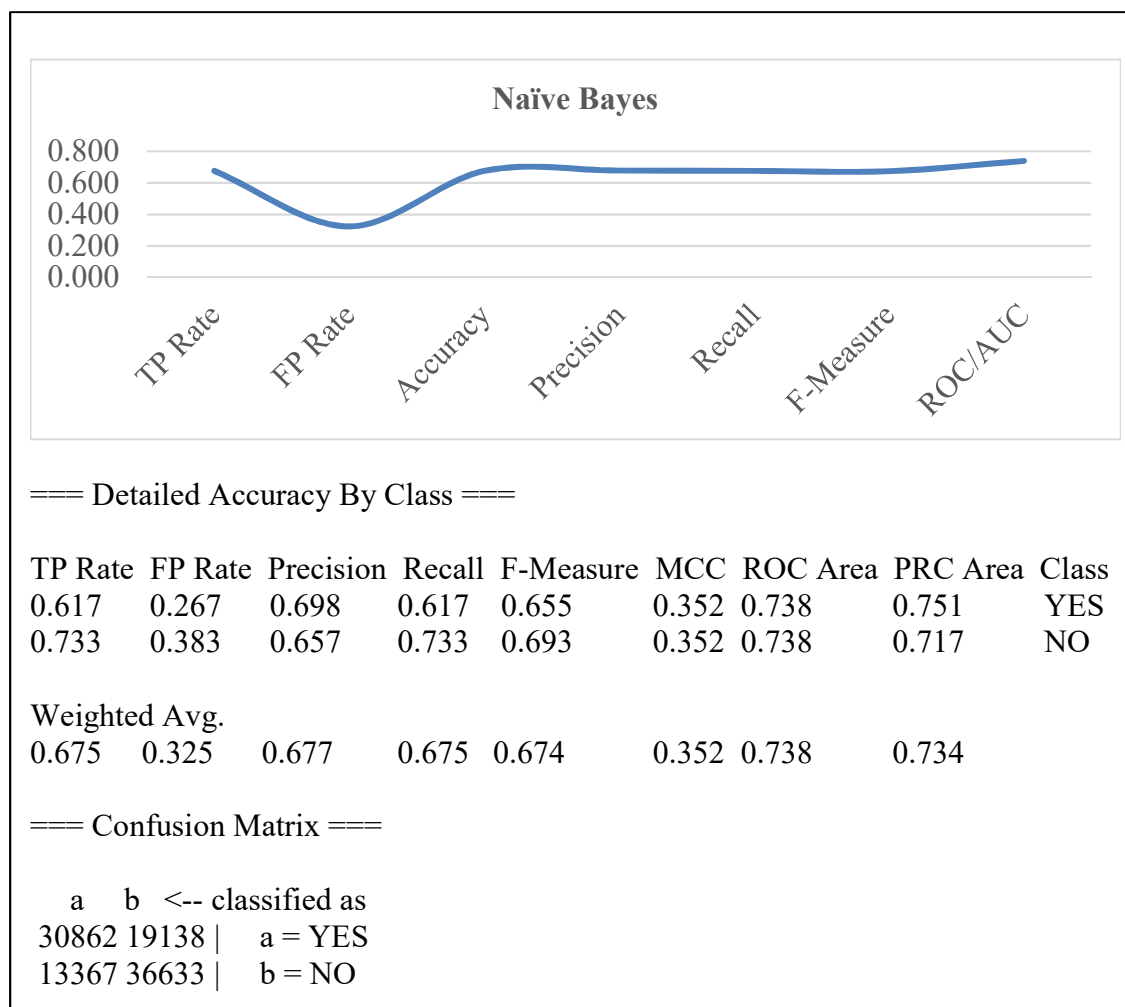


Figure 15. Performance Results and Plotted Values Delivered by Naïve Bayes

The AdaBoost Ensemble meta-learner combines weighted results obtained using different training data and algorithm models. Boosting refers to the process of assessing the impact that different data set instances have on the classifiers' training (Quinlan, 2006). As an ensemble method, the classifier predicts the outcome variable based on the individual predictions made by its component classifiers (Tan, Steinbach, & Kumar, 2006). Given a data set, AdaBoost Ensemble automatically adjusts the distribution of the training samples to force the algorithms to focus on those more difficult to classify. Depending on the degree of difficulty, a weight is assigned to each training iteration (Tan, Steinbach, & Kumar, 2006). The aggregation of independently weighted results, enables random errors to cancel, yielding a classification that more closely represents the correct alternative. The process developed to arrive at the result, essentially creates a new algorithm from the work of multiple ones, building a collection of independent decisions that yields results that are easier to generalize (Freund & Schapire, 1996). Quinlan (2006) found boosting classifiers to yield more accurate results. Caruana, Niculescu-Mizil, Crew, and Ksikes (2004) found ensemble classification models to perform better than other independent classifiers. The code that defines the steps taken by the AdaBoost Ensemble used in the research is shown below in Figure 16 (Freund & Schapire, 1996; Zaki & Meira, 2014).

AdaBoost Ensemble Classifier – Steps and Pseudo Code:

1. Train AdaBoost(D , Base Learn).
2. For each example d_i and D let its weight $w_i = 1/|D|$.
3. Let H be an empty set of hypotheses.
4. For t from 1 to T do:
 5. Learn a hypothesis h_t , from the weighted examples: $h_t = \text{Base Learn}(D)$.
 6. Add h_t to H .
 7. Calculate the error, ϵ_t , of the hypothesis h_t as the total sum weight of the examples that it classifies incorrectly.
 8. If $\epsilon_t > 0.5$ then exit loop, else continue.
 9. Let $\beta_t = \epsilon_t / (1 - \epsilon_t)$.
 10. Multiply the weights of the examples that h_t classifies correctly by β_t .
 11. Rescale weights of all of the examples so that the total sum weight remains 1.
 12. Return H .
13. Test AdaBoost(ex , H).
14. Let each hypothesis, h_t , in H vote for ex 's classification with weight $\log(1 / \beta_t)$.
15. Return the classification with the highest weighted vote total.

Figure 16. Code Defining Steps Taken by AdaBoost Ensemble

Sample performance results and associated plotted values obtained when using AdaBoost Ensemble on the control data set are shown in Figure 17.

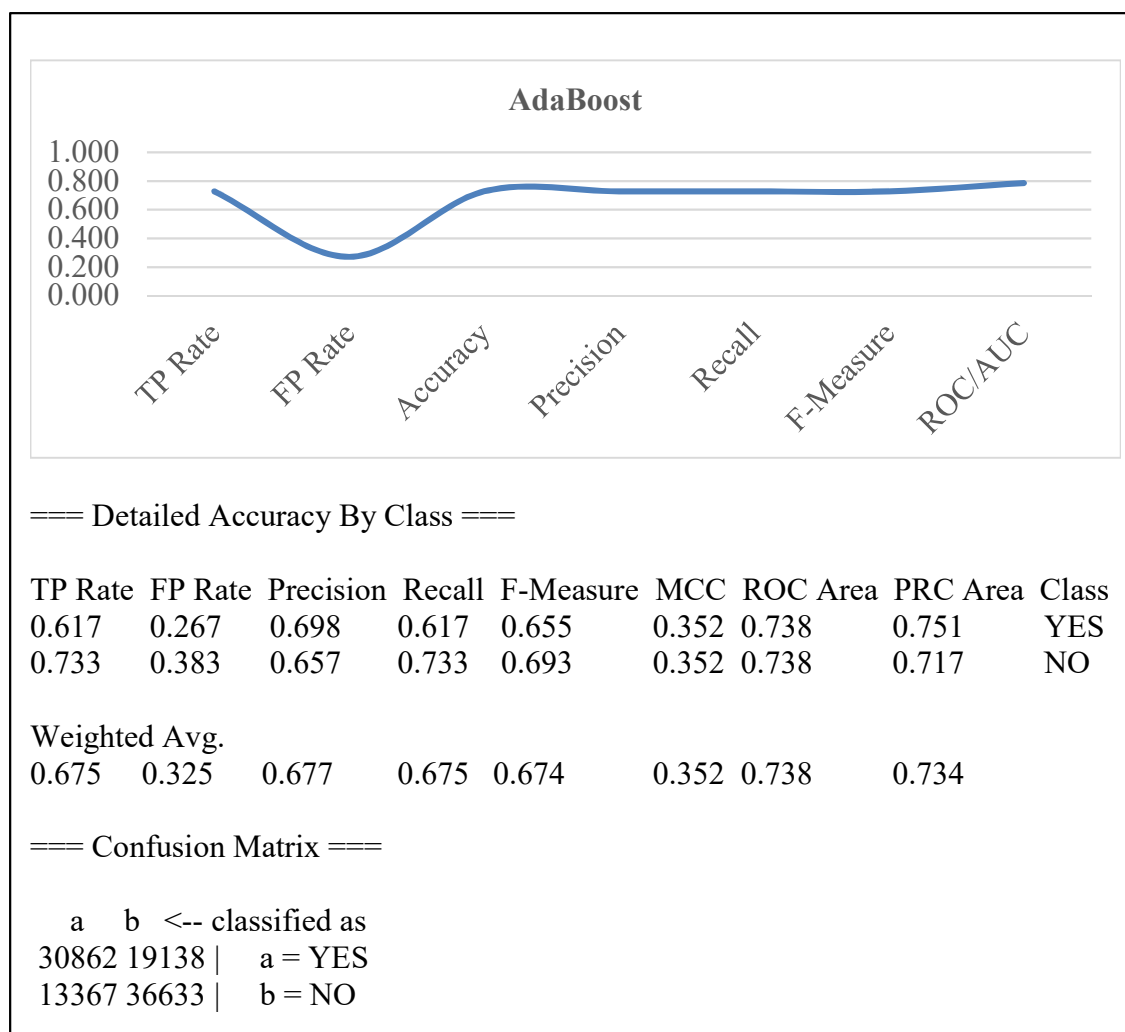


Figure 17. Performance Results and Plotted Values Delivered by AdaBoost

Random Forest is another ensemble method that uses multiple training inputs and employs independent decision trees to arrive at an aggregate performance metric. The code that defines the steps in the algorithmic processes undertaken by this Random Forest ensemble classifiers was defined by Tan, Steinbach, and Kumar (2006) as shown in Figure 18.

Random Forest Ensemble Classifier – Steps and Pseudo Code:

1. Let D denote original training data, k denote number of baseline classifiers, and T be the test data.
2. For $i = 1$ to k do
3. Create training set, D_i from D .
4. Build a base classifier C_i from D_i .
5. End for
6. For each test record $x \in T$ do
7. $C^*(x) = \text{Vote}(C_1(x), C_2(x), \dots, C_k(x))$
8. End for

Figure 18. Framework of Ensemble Algorithms

Sample performance results and associated plotted values obtained when using AdaBoost Ensemble on the control data set are shown in Figure 19.

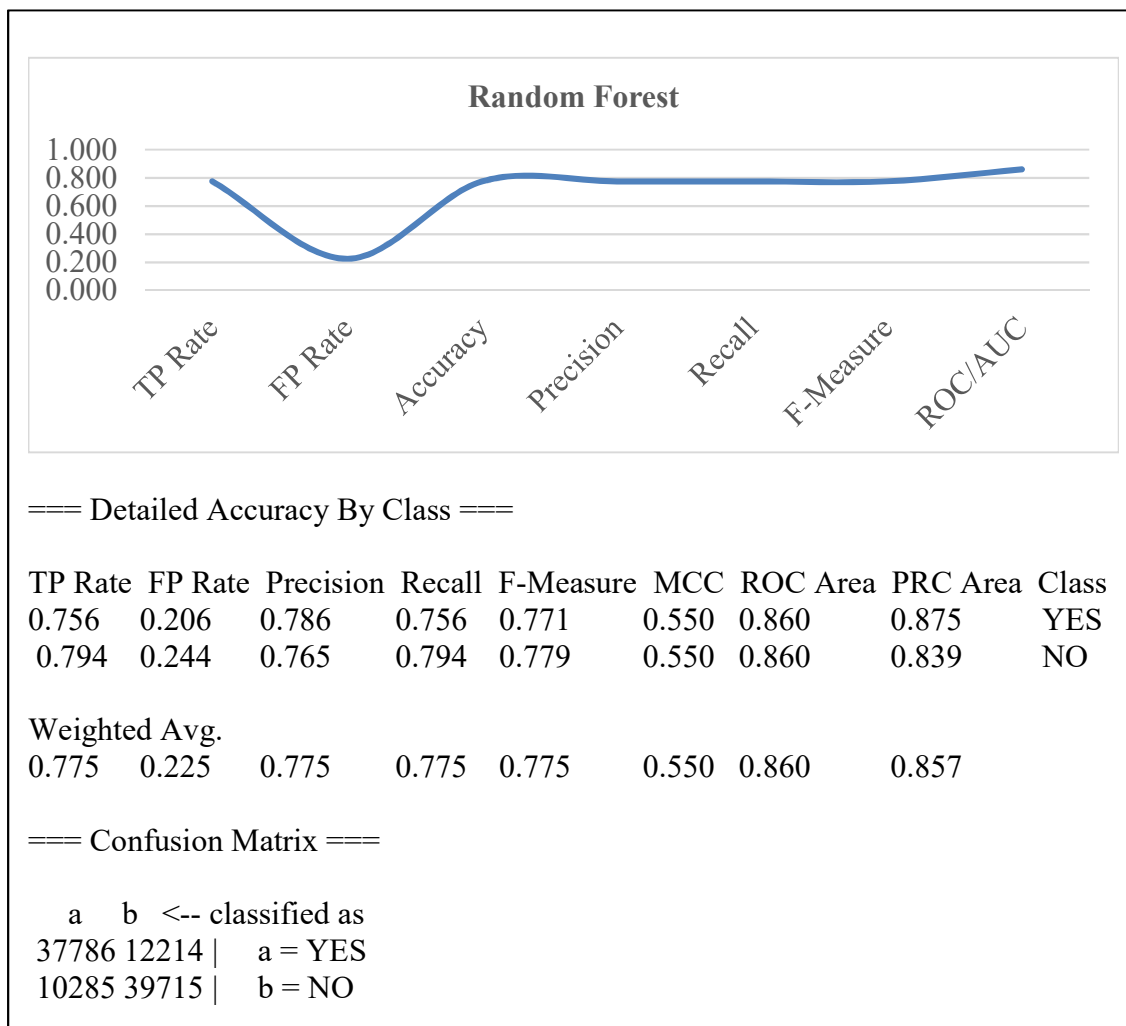


Figure 19. Performance Results and Plotted Values Delivered by Random Forest Ensemble

Experimental Design

In this research study, the Waikato Environment for Knowledge Analysis (WEKA) was used to select the data sets, configure and run the classifiers for data mining, measure their performance, and examine results in text and graphical presentations. Performance measurements associated with the classification process undertaken by the algorithms were collected and analyzed to determine the impact on the resulting knowledge discovery.

Attributes in the data set studied were converted to nominal or categorical values through a process of discretization. Attributes or variables in a data set can be represented in three forms. These can be in a nominal, numeric, or string format. Nominal or categorical variables correspond to specific predefined set values or codes. Numeric or continuous variables consist of real numbers that represent specific measurements. String can display a mix of codes and numbers. Converting the data set attributes to nominal or categorical values ensured that the algorithms used in the experiment would be able to classify the data across an established set of parameters. It was also important to declare the dependent/outcome variable a nominal binary value for the purpose of classification.

With a stable data set as the foundational control group, the tool's data filtering capabilities enabled the substitution and suppression of specific attributes from the data mining process. The feature, allowed select attributes considered to pose a risk to the re-identification of test subjects, to be set to constant values or removed altogether from the mining process. This emulated the effect of data masking and encryption of the select attribute values and created the experimental groups representing each data sets. WEKA enables the use of two kinds of filters; supervised and unsupervised. Supervised filters

screen attributes based on their specific impact on the outcome or classification variable. Unsupervised filters do not consider the influence that the individual attributes may have on results. Screening is performed based on the specific attributes' own characteristics to optimize classification categories. Since filtering was performed prior to algorithm training and mining, unsupervised filters were used. The use of filters enabled the treatment (i.e., modification, substitution, and/or suppression), of the select input variables during the data transformation phase.

Select sensitive attributes (marital status, race/ethnicity, Hispanic origin, and year of birth) were maintained intact in the control group. Correspondingly, the same attributes were masked or encrypted, creating the two experimental groups. The select attributes in the masked group were substituted with constant values already existing in the attribute value set. The same select attributes in the encrypted group were removed from the data set. The resulting control and experimental data sets included:

- *breast_cancer_100k_control.csv*
- *breast_cancer_100k_exp_mask.csv*
- *breast_cancer_100k_exp_encrypt.csv*

Using the four classification algorithms, their measured performance parameters were then recorded and compared to determine the impact of masking and encryption.

For each of the data mining classification algorithms employed in the research, specific configurations were selected to ensure they delivered the best possible results for performance comparison. Settings such as batch size, confidence factor, discretization, and number of iterations during execution were determined. The configuration of the parameters selected for each classifier are shown in Appendix D.

Before the algorithms were used for classification purposes, they were trained to perform within the parameters of the data sets in question. Percentage split and cross-validation are two of the preferred methods for training algorithms. Although percentage split performs the work with less computational resources (Witten, Frank, & Hall, 2011), 10-fold cross-validation was chosen for its completeness. Cross-validation is a technique that swaps the roles of training and testing data subsets (Tan, Steinbach, Kumar, 2006). The technique iteratively divides the data into n subsets or folds. Using the first $n-1$ folds to train the classifier, it then employs the last one to test for its accuracy in predicting the outcome variable. The process is repeated n times by swapping the roles of the training and testing folds, and calculating the average value to determine the overall algorithm accuracy. Cross-validation reduces the effects of bias that can be introduced with random sampling (Kohavi, 1995). According to Witten, Frank, and Hall (2011) the optimum number of folds needed to minimize error estimation in cross-validation is 10. The 10-fold cross-validation process automatically controls overfitting by gauging the amount of data used by the prediction model, and ensuring it works for a broad set of conditions where prediction of patient survival is desired. Overfitting occurs when a large amount of similar data is used to train an algorithm. This leads to a very precise prediction model for a narrow spectrum of possibilities, but one that fails when attempting to predict the outcome of unknown conditions. The cross-validation step ensured that the algorithm training produces an accurate prediction model across a wide range of conditions. According to Wilson and Rosen (2003), 10-fold cross-validation not only provides a robust means of measuring classification accuracy, but does so in a statistically sound approach.

The classification performance of the J48, Naïve Bayes, AdaBoost Ensemble, and Random Forest classifiers was then independently calculated and compared. A schematic representation of the data loading and mining process performed is shown in Figure 20. A description of how the process was set up and carried out using WEKA's "Explorer" and "Knowledge Flow" interfaces is described in Appendix C and D.

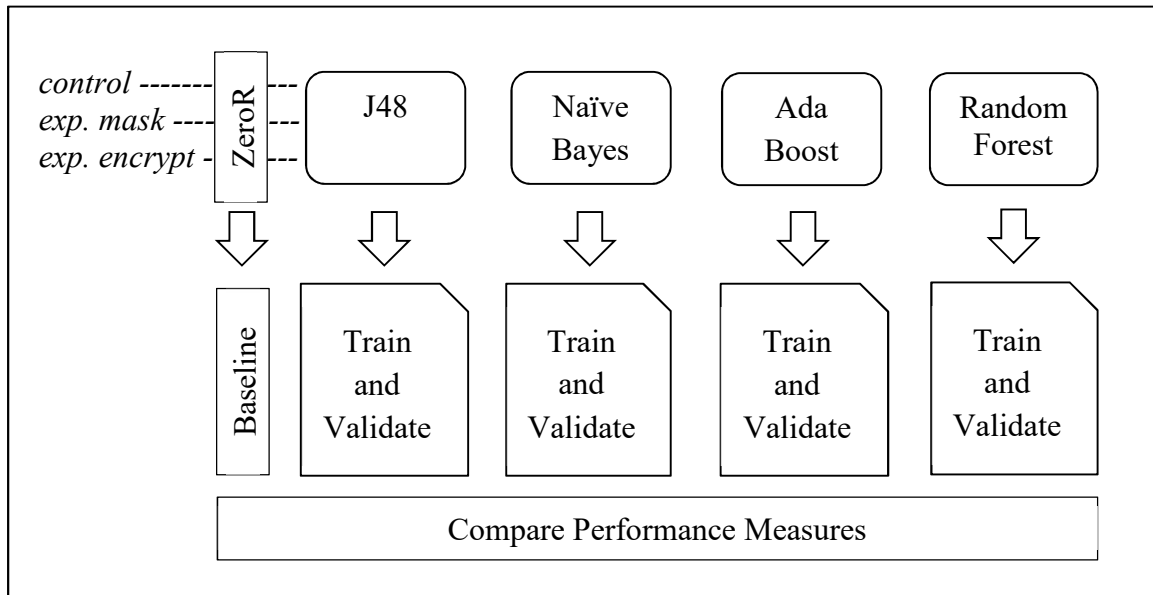


Figure 20. Algorithm Training and Validation Process

Resources Used

The research study utilized the hardware, software, and data resources outlined below.

Hardware

Hardware resources include a personal computer platform and an external hard disk. A Dell Latitude E7470 Ultrabook computer platform with 16 GB of RAM and a Toshiba 160 GB expanded hard drive were used to download and store the raw and preprocessed data sets and to run the analytical, data processing, and graphic design software.

Software

The hardware platform employed Microsoft Windows® 10 Pro 64-bit operating system and Java 8 general purpose programming language. The data mining application software included WEKA version 3.8.1, with Java runtime version 1.8.0_112-b15 (Hall, et al., 2009). WEKA is maintained by the open source community and was used under a general public license. The software was used to train and to validate the inductive learning classification algorithms employed in the study. Installation of the software followed standard procedures outlined by the University of Waikato. Best practices already employed by previous researchers, including Ahmadi and Abadi (2013) and Iyer (2013) were also followed. Ahmadi and Abadi (2013) used WEKA's association rule mining capability to uncover relationships between records in data sets and to measure related quality of the relationships. Iyer (2013) used WEKA's forward error correction and decision tree building capabilities to filter out non-relevant attributes of streaming data to improve quality. As an open source machine learning tool, WEKA has been used for over 20 years and is widely accepted in academia and the business community. WEKA has been extensively used in scientific research and in enterprises including

hospital information systems (Murugan & Kannan, 2013). WEKA is used for data mining simulations to assess data quality dimensions. Tiwari, Jha, and Yadav (2012) used WEKA to gauge performance of data mining algorithms and found that the nature of data sets and volume of instances played a significant part in how different algorithms performed. Zlotnik, Gallardo-Antolín, and Martínez (2015) further used WEKA to determine predictive probabilities and focused on calibration as a critical component of the classification problem.

Microsoft Excel[®] was used for data formatting and statistical analysis, and other Microsoft Office[®] Suite products were employed to create the documentation and graphics. The laboratory environment was created by the researcher within available private space.

Summary

The goal of the research outlined herein was to define cause and effect. The specific objective was to experimentally determine if the implementation of masking and encryption techniques impacted the quality of knowledge discovery. By its nature, the research did not lend itself to be carried out in the natural environment using real-world data. Doing so would have limited how data could have been treated to test the different scenarios under study, and would have introduced external variables that would have limited the generalizability of findings. For these reasons, a controlled quasi-experimental approach was employed to conduct the study. The quasi-experimental approach tested the impact of treatment on a control and experimental group, and determined the associated performance of four different classifiers.

Chapter 4

Results

Introduction

The quasi experiment enabled the measurement of classification performance values by different algorithms on a control group data set and two experimental group data sets. Results displayed general variations in performance parameters. The performance measured by J48, Naïve Bayes, and the AdaBoost and Random Forest Ensemble classifiers consistently showed superior values over those delivered by the baseline ZeroR classifier. A steady decline in performance values was also observed when data masking and encryption were used to protect the select attributes in the data sets.

Information gain rankings of the attributes in the initial data set, performed as part of the data preprocessing and transformation phase, had previously confirmed that the select attributes were among the most influential in predicting the value of the outcome variable (i.e., patient survival). Results of the classification performance test supported the premise that the use of data masking and encryption can have a measured effect on the quality of data mining outcomes, and potentially impact decision-making and patient treatment protocols.

Findings

Comparing the resulting performance metric values: classification accuracy, precision, recall, f-measure, and ROC/AUC delivered by the four algorithms, it was found that they all figured within a percentage point of each other when measured within the control group. J48 and Random Forest consistently outperformed the other three algorithms

across all groups. The spread of performance figures were also observed to be tighter together within the experimental groups. Figure across the experimental masking and encryption data sets varied on average less than 5% to 6% between the algorithms in the same data group, but nearly doubled to 9% to 10% when compared with results obtained from the control and experimental groups.

All four algorithms and the baseline deliver similar rankings in performance values across the control and experimental groups. J48 performed best, followed by Random Forest. Naïve Bayes and AdaBoost switched ranking positions depending on the groups. Two sets of performance metrics, corresponding to the different filtering techniques employed, were calculated when using Naïve Bayes. However, performance metrics obtained using unsupervised numeric to nominal filtering closely overlapped the values obtained using supervised discretization in all three groups. Of the four classifiers used, AdaBoost Ensemble delivered the lowest performance values observed above the baseline.

Observations also showed that data masking and encryption, on average, delivered matching classification performance metrics. This was a surprising finding. All four algorithmic models, including the average value comparison between the control and experimental data sets, showed a higher ROC/AUC value over the other performance metrics. The larger the value of the AUC, the better the classifier is at measuring the outcome (Tan, Steinback, & Kumar, 2006). Comparing the AUC values of the four algorithms used, J48 and Random Forest delivered the higher values in all three groups. The performance metrics measured by each classification algorithms are compiled in Table 1-3. A graphical illustration of each of these is shown in Figures 21-24. Confusion

matrixes built by the data mining tool to produce the results listed in all four tables, as well as the ROC graphs providing visual representation of the cost/benefit analyses were also recorded and are included in Appendix D.

Control Group

Table 1 lists weighted performance metrics achieved by each of the classification algorithms used to predict the outcome variable within the control group data set. Values are weighted because they are based on the set of attributes determined to carry highest influence on the outcome variable that they were used to predict. J48 and Random Forest delivered the highest values with almost matching 78% accuracy and precision. This was followed by AdaBoost Ensemble at 73%. Naïve Bayes delivered the lowest performance metrics at 68%. Values obtained when attributes were converted from numeric to nominal using an unsupervised filter, matched values obtained when the same attributes were subjected to supervised discretization.

Table 1. Control Group Weighted Results

Number of Instances: 100,000						
Number of Attributes: 13						
Algorithm	ZeroR	J48	Naïve Bayes (Non-Sup)	Naïve Bayes (Sup Dis)	Ada Boost	Random Forest
TP Rate	0.500	0.784	0.676	0.675	0.728	0.775
FP Rate	0.500	0.216	0.324	0.325	0.272	0.225
Accuracy	0.500	0.784	0.676	0.675	0.728	0.775
Precision	0.250	0.786	0.678	0.677	0.728	0.775
Recall	0.500	0.784	0.676	0.675	0.728	0.775
F-Measure	0.333	0.784	0.674	0.674	0.728	0.775
ROC/AUC	0.500	0.858	0.739	0.738	0.786	0.860

Plotting the metrics obtained for each of the classification algorithms allowed a more practical way to view the relative performance between classifiers. A schematic representation is shown in Figure 21.

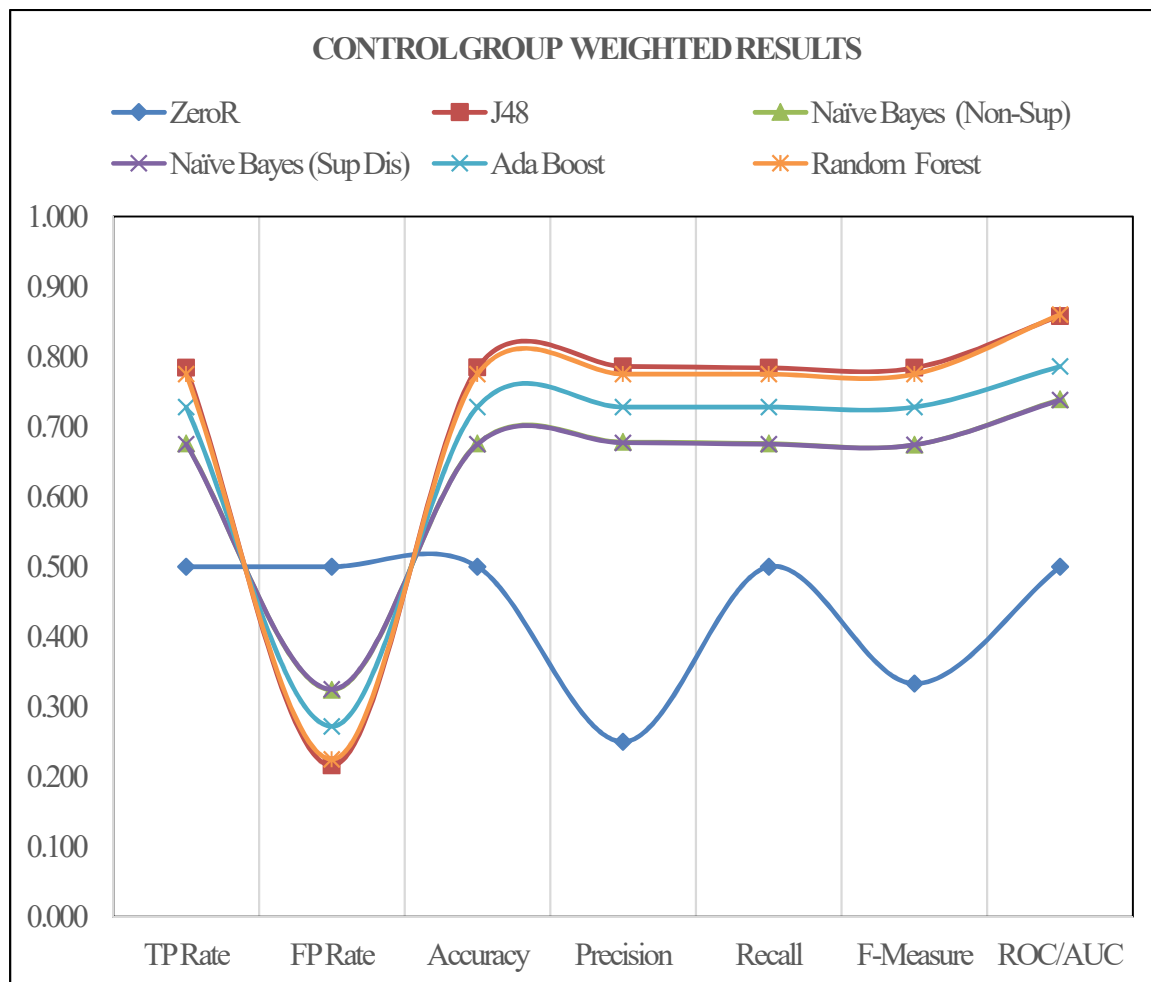


Figure 21. Graphical Representation of Control Group Performance Metrics Delivered by J48, Naïve Bayes, and AdaBoost and Random Forest Ensemble Algorithms Against Baseline ZeroR Classifier

Plotting the associated ROC graphs showing the AUC for each of the classification algorithms also provided an easy way to compare the classifiers relative performance. Overlaid ROC graph produced for J48, Naïve Bayes, AdaBoost Ensemble, and Random Forest classifiers against the ZeroR baseline are shown in Figure 22.

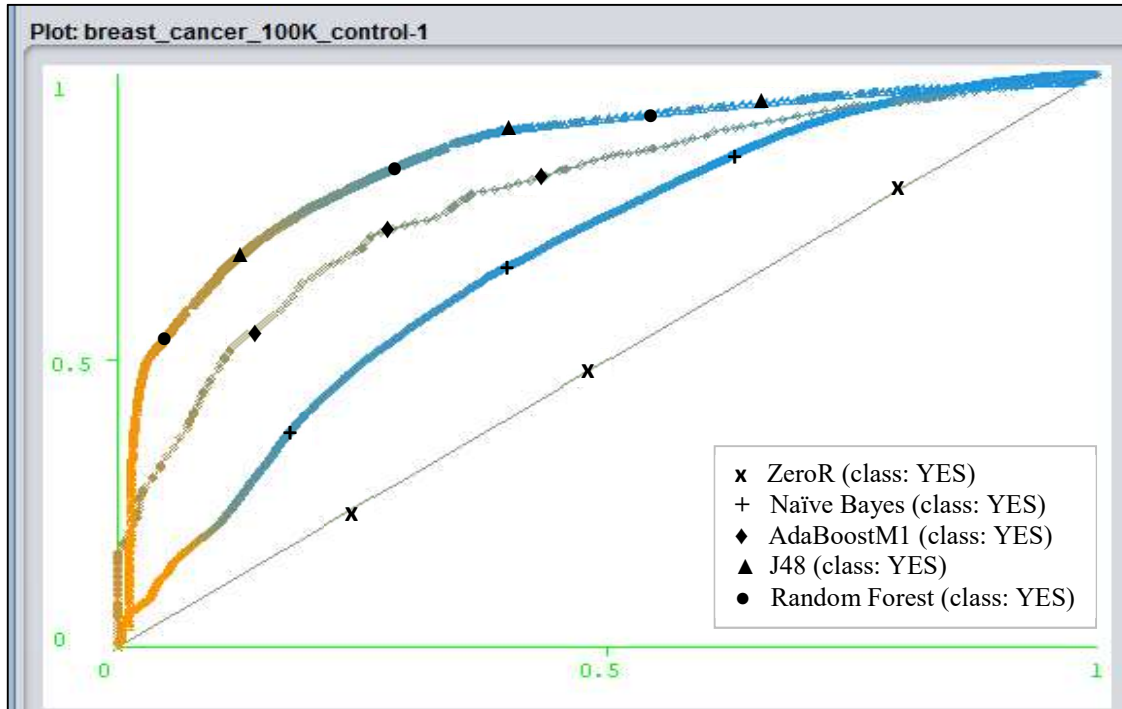


Figure 22. Overlaid ROC Graph Produced by Classifiers for the Control Group Data Set

Experimental Group – Data Masking

Weighted performance metrics achieved by each of the classification algorithms used to predict the outcome variable within the first experimental group when select attributes were masked, are listed in Table 2. Again, J48 and Random Forest performed better than the rest of the classifiers with 65% and 63% accuracy and precision values respectively. Naïve Bayes came in third with 63% accuracy and precision when both unsupervised filtering and supervised discretization were used, followed by AdaBoost Ensemble with 60% and 61% accuracy and precision values respectively.

Attribute masking had two noticeable effect on the classifiers' behavior. First, the spread between performance metrics measured by each of them was much tighter. Second, Naïve Bayes delivered better performance than AdaBoost Ensemble.

Table 2. Experimental Group Weighted Results Using Data Masking

Number of Instances: 100,000						
Number of Attributes: 13						
Algorithm	ZeroR	J48	Naïve Bayes (Non-Sup)	Naïve Bayes (Sup Dis)	Ada Boost	Random Forest
TP Rate	0.500	0.652	0.630	0.678	0.602	0.632
FP Rate	0.500	0.348	0.370	0.372	0.398	0.368
Accuracy	0.500	0.652	0.630	0.628	0.602	0.632
Precision	0.250	0.653	0.630	0.628	0.613	0.632
Recall	0.500	0.652	0.630	0.628	0.602	0.632
F-Measure	0.333	0.652	0.630	0.628	0.592	0.632
ROC/AUC	0.500	0.707	0.691	0.690	0.667	0.684

The closeness of resulting performance metrics is illustrated by overlaying each of the classifiers' measured results as shown in Figure 23.

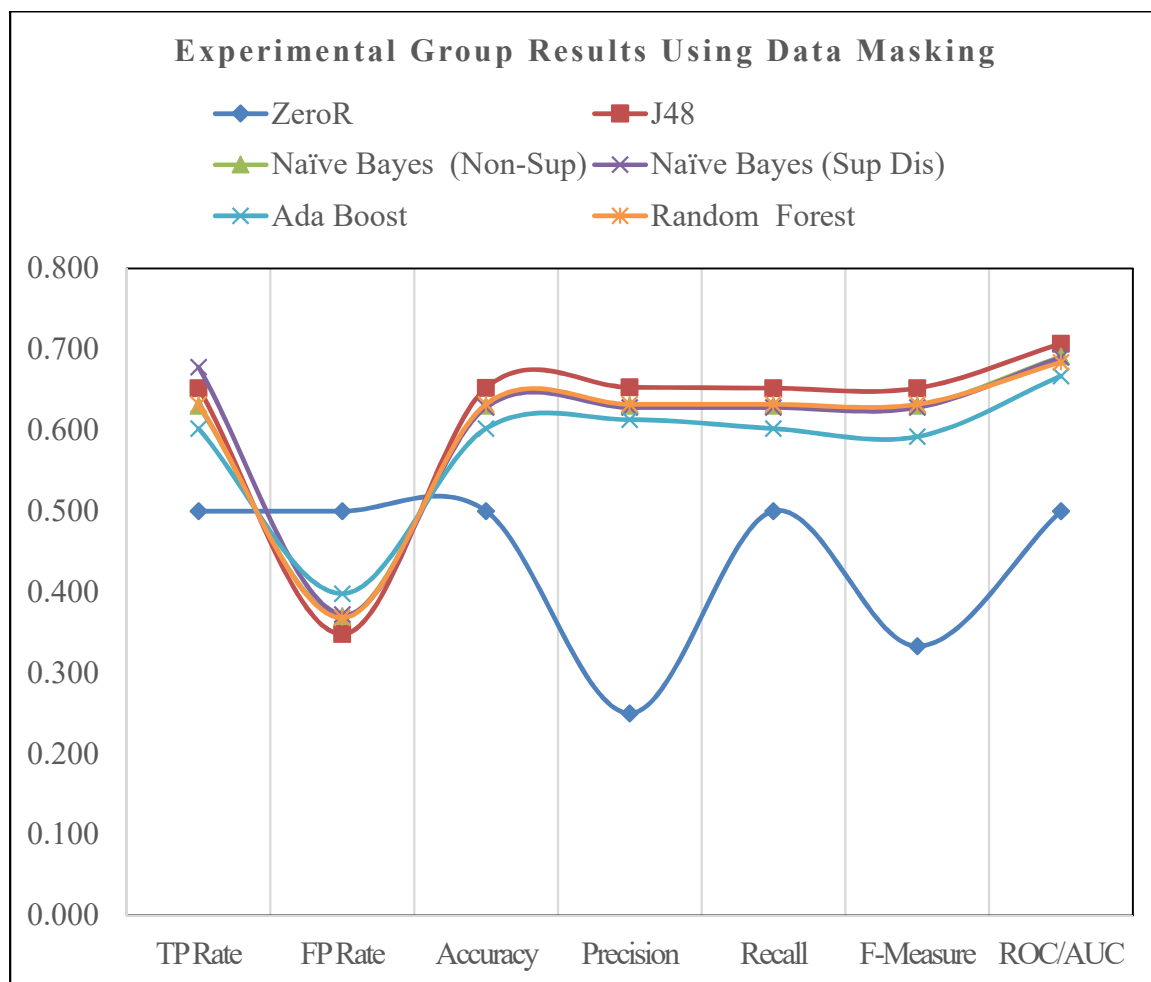


Figure 23. Graphical Representation of Experimental Group Performance Metrics Delivered by J48, Naïve Bayes, and AdaBoost and Random Forest Ensemble Algorithms Against Baseline ZeroR Classifier when Data Masking was Used

A plot of associated ROC graphs for the experimental data masking group is shown in Figure 24.

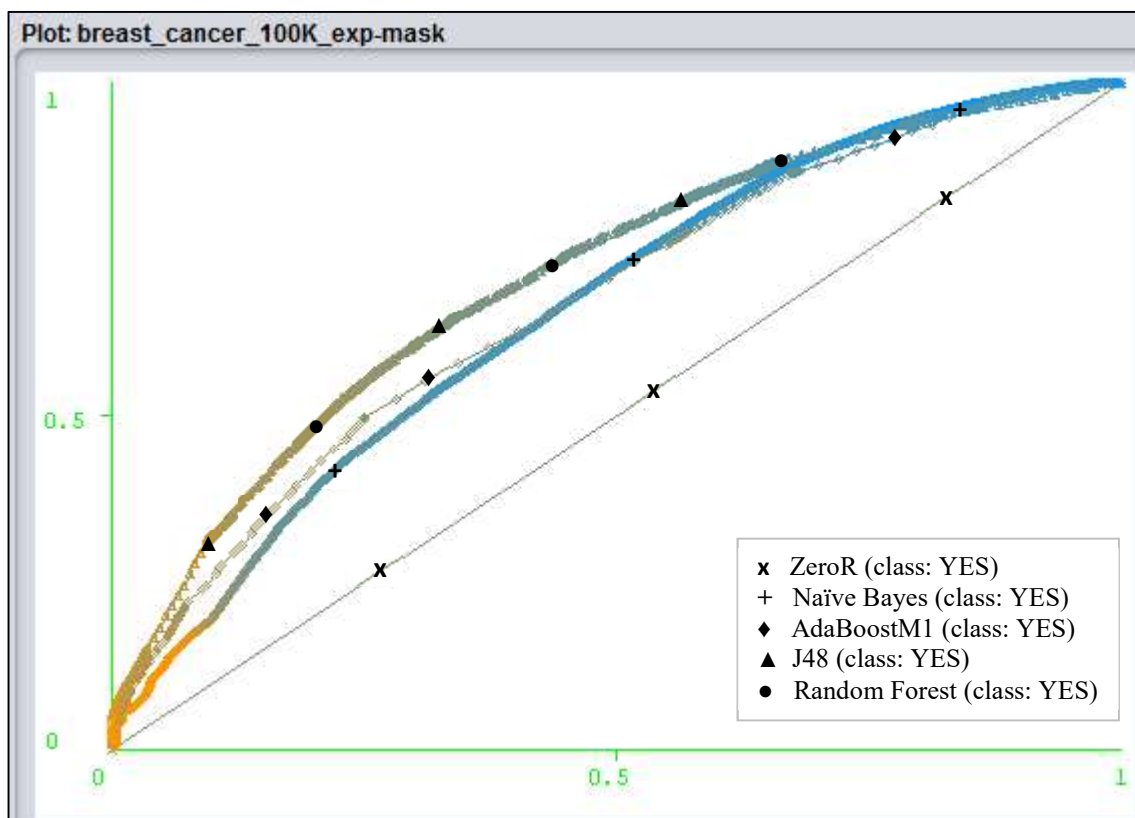


Figure 24. Overlaid ROC Graph Produced by Classifiers for the Experimental Data Masking Group Data Set

Experimental Group – Data Encryption

Weighted performance metrics achieved by each of the classification algorithms when select attributes were suppressed in the experimental encrypted data set are listed in Table 3. Once again, J48 and Random Forest delivered the highest accuracy and precision with 65% and 63% respectively. Naïve Bayes came in third also with 63% accuracy and precision when numeric values were transformed to nominal, and 62% when they were discretized using the supervised filter. AdaBoost Ensemble followed with 60% and 61% accuracy and precision respectively. The effect of attribute encryption on classification performance closely matched the results obtained when attributes were masked.

Table 3. Experimental Group Weighted Results Using Data Encryption

Number of Instances: 100,000						
Number of Attributes: 9						
Algorithm	ZeroR	J48	Naïve Bayes (Non-Sup)	Naïve Bayes (Sup Dis)	Ada Boost	Random Forest
TP Rate	0.500	0.652	0.630	0.616	0.602	0.632
FP Rate	0.500	0.348	0.370	0.384	0.398	0.368
Accuracy	0.500	0.652	0.630	0.616	0.602	0.632
Precision	0.250	0.653	0.630	0.616	0.613	0.633
Recall	0.500	0.652	0.630	0.616	0.602	0.632
F-Measure	0.333	0.652	0.630	0.616	0.592	0.632
ROC/AUC	0.500	0.707	0.691	0.658	0.667	0.684

Associated performance metrics measured by each of the classification algorithms are graphically illustrated in Figure 25.

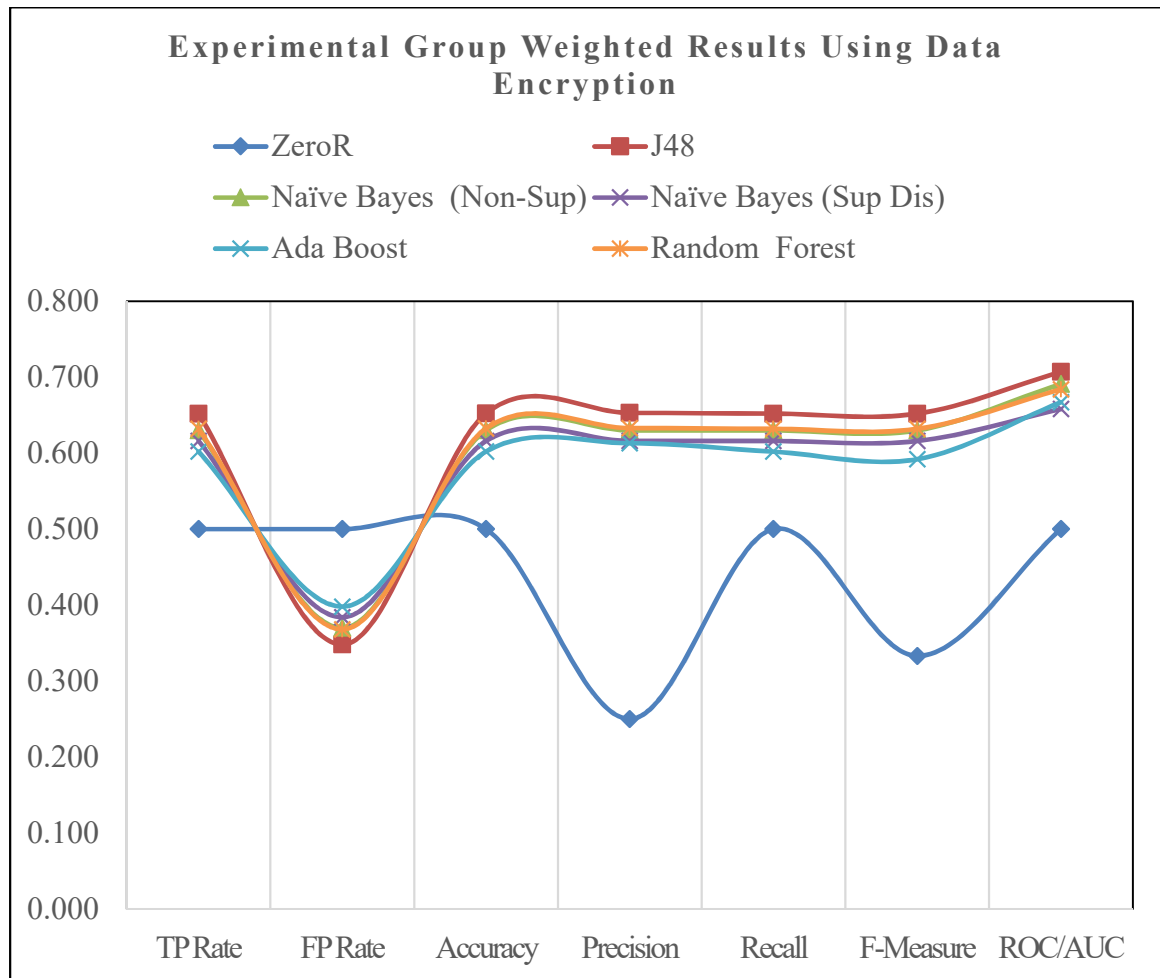


Figure 25. Graphical Representation of Experimental Group Performance Metrics Delivered by J48, Naïve Bayes, and AdaBoost and Random Forest Ensemble Algorithms against Baseline ZeroR Classifier when Data Encryption was Used

A plot of associated ROC graphs for the experimental data encryption group is shown in Figure 26.

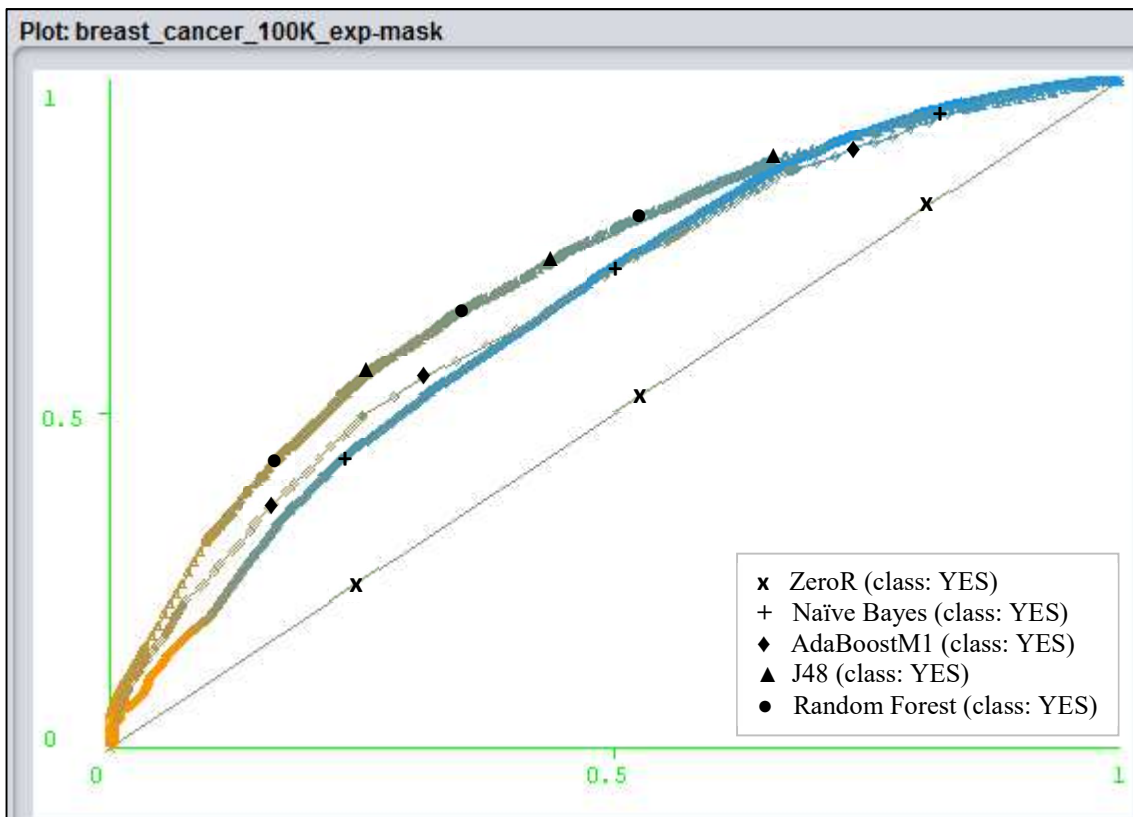


Figure 26. Overlaid ROC Graph Produced by Classifiers for the Experimental Data Encryption Group Data Set

To compare the performance of each classification algorithms across the control and experimental groups, average results of each of the seven metrics values were calculated. Tabulated results are shown in Table 4. Average performance measures across control and experimental groups are listed in Table 5. The variance observed for each measured parameter between groups are shown in Table 6.

Table 4. Relative Classifier Performance Measures Between Groups

Cross Validation: 10-Fold							
Control Group							
<i>breast_cancer_100k_control.csv</i>							
Algorithm	TP	FP	Accuracy	Precision	Recall	F-Measure	ROC/AUC
J48	0.784	0.216	0.784	0.786	0.784	0.784	0.858
Naïve Bayes (Non-Sup)	0.676	0.324	0.676	0.678	0.676	0.674	0.739
Naïve Bayes (Sup Dis)	0.675	0.325	0.675	0.677	0.675	0.674	0.378
AdaBoost	0.728	0.272	0.728	0.728	0.728	0.728	0.786
Random Forest	0.775	0.225	0.775	0.775	0.775	0.775	0.860
Average	0.728	0.272	0.728	0.729	0.728	0.727	0.724
Experimental Group Using Data Masking							
<i>breast_cancer_100k_exp_mask.csv</i>							
J48	0.652	0.348	0.652	0.653	0.652	0.652	0.707
Naïve Bayes (Non-Sup)	0.630	0.370	0.630	0.630	0.630	0.630	0.691
Naïve Bayes (Sup Dis)	0.628	0.372	0.628	0.628	0.628	0.628	0.690
AdaBoost	0.602	0.398	0.602	0.613	0.602	0.592	0.667
Random Forest	0.632	0.368	0.632	0.632	0.632	0.632	0.684
Average	0.629	0.371	0.629	0.631	0.629	0.627	0.688
Experimental Group Using Data Encryption							
<i>breast_cancer_100k_exp_encrypt.csv</i>							
J48	0.652	0.348	0.652	0.653	0.652	0.652	0.707
Naïve Bayes (Non-Sup)	0.630	0.370	0.630	0.630	0.630	0.630	0.691
Naïve Bayes (Sup Dis)	0.616	0.384	0.616	0.616	0.616	0.616	0.658
AdaBoost	0.602	0.398	0.602	0.613	0.602	0.592	0.667
Random Forest	0.632	0.368	0.632	0.633	0.632	0.632	0.684
Average	0.626	0.374	0.626	0.629	0.626	0.624	0.681

Table 5. Average Performance Measured Across Control and Experimental Groups

Group	TP	FP	Accuracy	Precision	Recall	F-Measure	ROC/AUC
Control	0.728	0.272	0.728	0.729	0.728	0.727	0.724
Exp. Mask	0.629	0.371	0.629	0.631	0.629	0.627	0.688
Exp. Encrypt	0.626	0.374	0.626	0.629	0.626	0.624	0.681

Table 6. Variance of Average Performance Metrics Measured Between Groups

Variance	TP	FP	Accuracy	Precision	Recall	F-Measure	ROC/AUC
Between Control and Experimental Mask Data Set	0.099	0.099	0.099	0.098	0.099	0.100	0.036
Between Control and Experimental Encrypt Data Set	0.101	0.101	0.101	0.100	0.101	0.103	0.043

When comparing performance metrics obtained between the control and experimental groups, there was a clear decrease in values obtained when masking and encrypting the select attributes (i.e., marital status at diagnosis, race/ethnicity, Hispanic origin, and year of birth). Having built four different classification models, the ROC/AUC was used to determine which model made the best predictions on the outcome variable given the impact of treatment on the select input variables. J48 showed a more significant increase in the AUC value when treatment was applied, with 21% increase over the baseline value, followed by Random Forest with 18%. The AUC value represents the probability that one of the randomly selected instances from the experimental groups (i.e., patient from the data sets where attributes were treated) had a higher rate of survival than any other randomly selected patient from the control group. The higher the AUC calculated by a classifier in the experimental groups, the more significant the effect that data masking and encryption of select attributes had on the outcome variable. An illustration of how the average performance measures compared between the control group and the two experimental data sets is shown in Figure 27.

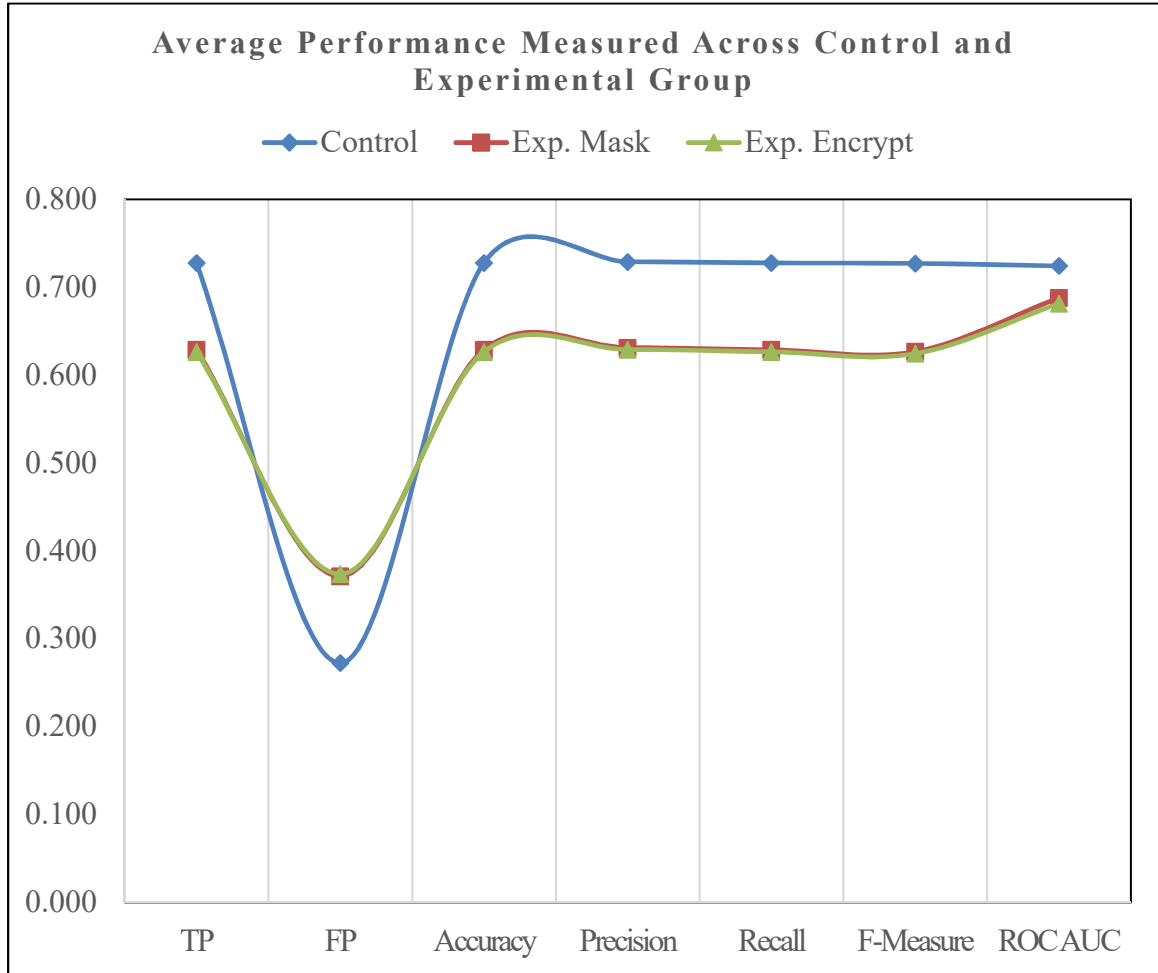


Figure 27. Graphical Representation Comparing Combined Performance Variance Between Control and Experimental Groups

A graphical representation of the variance observed for the average performance metrics between the control group and the experimental mask group, and between the control group and the experimental encrypt group is shown in Figure 28.

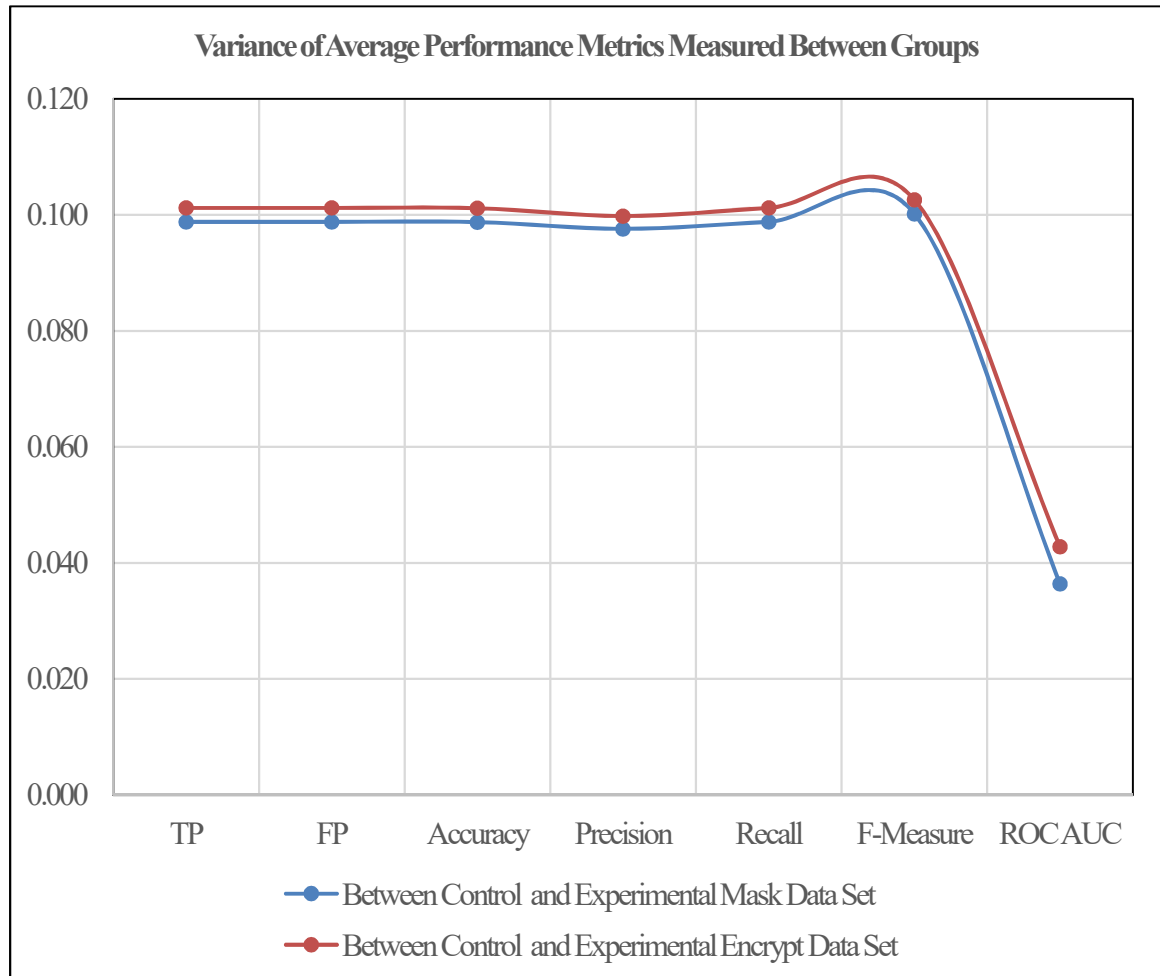


Figure 28. Average Performance Variance Between Control and Experimental Groups

Data Analysis

Results obtained from the measured performance parameters across a control group and two experimental groups, provided evidence of a general impact on classification performance when masking and encrypting select predictor variables. The single control group applied no treatment to the predictor variables. The two experimental groups applied different types of treatment to the same predictor variables. In the experimental masking group, the values of the predictor variables were substituted with like values found within the data set. In the experimental encrypt group, the same predictor variables were suppressed.

After conducting associated algorithmic runs on the control and experimental groups, classification performance was observed to vary between the groups. The algorithms employed measured the impact of applied treatment to various degrees. The baseline established using ZeroR on the balanced sample, yielded the expected 50% accuracy across all three data sets. The absence of FPs and FNs in the control group met expectations and offered a high level of confidence in the model. FPs and FNs are generally indicative of errors, and while misclassification can be expected in any model, the less these occur is a sign of accuracy.

An interpretation of results and an assessment of the representative impact of data masking and encryption on the utility of the data was then made comparing the values by group and data protection mechanism. This enabled the determination of the trade-off between security of sensitive medical records and the usefulness of predicted values for more accurate decision-making in healthcare.

Statistical Significance

To corroborate the statistical significance of the cause and effect relationship and test the hypothesis that data masking and encryption have an impact on classification performance, a statistical analysis was performed. Statistical analyses are used to validate hypotheses on the basis of the available experimental data used for testing. An analysis of variance (ANOVA) was carried out to determine if the differences in measured classification values, obtained using the various classifiers across the control and experimental groups, were substantial enough to be noteworthy between and within one another. Data mining studies found in the literature have used ANOVA to test differences between related sample data (Gao, 2015; Wilson & Rosen, 2003). To statistically test the hypothesis on the basis of the experimental data obtained, two-factor ANOVA with replication was used. Since algorithm performance measurements were obtained from multiple groups (i.e., one control group and two experimental groups), and multiple algorithms were used to measure performance parameters, two-factor ANOVA with replication offered a way to assess the significance of these relationships since it tests for differences in means between two or more groups of measurements. Given that the experiment measured the performance of four algorithms' ability to predict the patients' survival across a control group and two experimental groups, two-factor ANOVA was the most appropriate test for data analysis. Assessing the effect that the treatment of the four predictor variables had on the single dependent/outcome variable (patient survival), and determining whether there was an interacting effect between these, two-factor ANOVA with replication determined whether there was convergence of results and statistical relevancy.

Statistical differences between the results obtained in experiments are generally considered significant when they vary by more than 5% (Witten & Frank, 2005). Using 0.05 as the α value, the results of the ANOVA test enabled the rejection of the null hypothesis. Starting from the base null hypothesis that masking and encryption have no effect on classification performance, the results of the ANOVA test showed statistical significance at the sample level. This level was representative of the three states of the data set used in the experiment; untreated control group data, treated masked data, and treated encrypted data. Table 7 presents the variability among values collected from the experiment.

Table 7. Variability Among Values Collected from Experiment

Group / Classifier	Accuracy	Precision	Recall	F-Measure	ROC/AUC
CONTROL					
J48	0.78448	0.786	0.784	0.784	0.858
Naïve Bayes (Non-Sup)	0.67570	0.678	0.676	0.674	0.739
Naïve Bayes (Sup Dis)	0.67495	0.677	0.675	0.674	0.378
Ada Boost	0.72809	0.728	0.728	0.728	0.786
Random Forest	0.77501	0.775	0.775	0.775	0.860
MASK					
J48	0.65214	0.653	0.652	0.652	0.707
Naïve Bayes (Non-Sup)	0.62999	0.630	0.630	0.630	0.691
Naïve Bayes (Sup Dis)	0.62825	0.628	0.628	0.628	0.690
Ada Boost	0.60200	0.613	0.602	0.592	0.667
Random Forest	0.63208	0.632	0.632	0.632	0.684
ENCRYPT					
J48	0.65214	0.653	0.652	0.652	0.707
Naïve Bayes (Non-Sup)	0.62999	0.630	0.630	0.630	0.691
Naïve Bayes (Sup Dis)	0.61593	0.616	0.616	0.616	0.658
Ada Boost	0.60200	0.613	0.602	0.592	0.667
Random Forest	0.63240	0.633	0.632	0.632	0.684

When the source of variation was measured at the sample level representative of the three states of the data set, the F value was found to be greater than the F critical value, and the p value was observed to be less than the significance level α , which had been set at a value of 0.05. The set value of α represented the predisposition to accept at least five erroneous classifications every 100 times the data mining test is performed.

The lack of statistical significance when the source of variation was measured at the column, or classification metric level, as shown in Table 7, was indicative of the consistency of the measured values by each of the classifiers (i.e., accuracy, precision, recall, f-measure, and ROC/AUC). Since no single metric deviated in value from the others measured parameters, no statistical significance was observed at the column level. The interaction between the column corresponding to the individual metrics, and the rows corresponding to the individual classifiers used to measure the performance values, was also found not to be statistically significant. A reasons why the analysis may have failed to show statistical significance at the column level may had been that data masking and encryption had no impact on classification performance when compared as individual accuracy, precision, recall, f-measure, and ROC/AUC. Another reason may be that the research failed to collect enough data to provide sufficient evidence. According to Tan, Steinbach, and Kumar (2006), when comparing the performance of different classifiers, the variations observed may not always be statistically significant depending on the size of the sample. However, since the experiment used a sample size of 100,000 instances, this factor is not believed to be the case. Results of the two-factor ANOVA with replication test are shown in Table 8.

Table 8. Statistical Significance of Classification Parameters

ANOVA						
<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Sample	0.12899	2	0.06449	17.80935	8.477E-07	3.15041
Columns	0.01627	4	0.00407	1.123413	0.3540183	2.52522
Interaction	0.00978	8	0.00122	0.337504	0.9479039	2.09697
Within	0.21728	60	0.00362			
Total	0.37232	74				

Results obtained from the two-factor ANOVA with replication testing revealed that a significant differences existed in the measured classification performance between the control and experimental groups. However, the results of the test did not assess which specific treatment group (i.e., the experimental mask or experimental encrypt), presented the significant difference(s). Therefore, to assess which experimental data set carried the most significant variation, a multi comparison post hoc Tukey honest significance difference (HSD) test was performed.

The Tukey HSD test is generally performed to confirm where variations arise between control and treatment groups (Wilson & Rosen, 2003). Tukey HSD test are carried out when an overall statistically significant difference in group means is found and the null hypothesis has been rejected. In this experiment, the Tukey HSD test complemented two-way ANOVA with replication to determine which pairwise results produced the most significant differences in observed mean values. A re-configuration of Table 7 combined the variability of results by groups to enable the execution of the Tukey HSD test. The variability of results by groups is presented in Table 9.

Table 9. Variability of Measured Results by Group

CONTROL	MASK	ENCRYPT
0.7845	0.6521	0.6521
0.7860	0.6530	0.6530
0.7840	0.6520	0.6520
0.7840	0.6520	0.6520
0.8580	0.7070	0.7070
0.6757	0.6300	0.6300
0.6780	0.6300	0.6300
0.6760	0.6300	0.6300
0.6740	0.6300	0.6300
0.7390	0.6910	0.6910
0.6750	0.6283	0.6159
0.6770	0.6280	0.6160
0.6750	0.6280	0.6160
0.6740	0.6280	0.6160
0.3780	0.6900	0.6580
0.7281	0.6020	0.6020
0.7280	0.6130	0.6130
0.7280	0.6020	0.6020
0.7280	0.5920	0.5920
0.7860	0.6670	0.6670
0.7750	0.6321	0.6324
0.7750	0.6320	0.6330
0.7750	0.6320	0.6320
0.7750	0.6320	0.6320
0.8600	0.6840	0.6840

As part of the Tukey HSD test, the absolute difference in the means of observed values by group were calculated, and the critical range determined and then compared. Observing the absolute difference against the critical range for each of the data set comparisons (i.e., control versus experimental mask, control versus experimental encrypt, and experimental mask versus experimental encrypt), values showing higher absolute difference corresponding to the comparison between control and experimental mask, and control and experimental encrypt, were determined to be the ones showing significantly different results. No significant difference was found between the experimental mask and experimental encrypt groups. Results of the test are presented in Table 10. The complete arrangement and results of the two-factor ANOVA with replication and Tukey HSD tests, including the associated studentized range distribution table are included in Appendix F and G.

Table 10. Multiple Comparison of Statistical Significance

Tukey Multiple Comparison			
	Q_u	3.384	
	<i>Numerator df</i>	3	<i>Denominator df</i> 72
<i>Comparison</i>	<i>Absolute Comparison</i>	<i>Critical Range</i>	<i>Result</i>
Control to Mask	0.08635	0.0393455	Significantly Different
Control to Encrypt	0.08951	0.0393455	Significantly Different
Mask to Encrypt	0.00316	0.0393455	Not Significantly Different

Given that the experiment sought out to investigate the causal effect of masking and encryption, the measured variations between the samples or states of the data sets represented by the control and experimental groups, was the focus for validation of statistical significance. Results of this test enabled the rejection of the null hypothesis that stated that data masking and encryption of the predictor variables in the data set had no effect on classification performance, and validated that treatment of these attributes in the form of data masking and encryption, can indeed impact the dependent/outcome variable and the quality of data mining results.

Summary

Experimental results obtained when masking and encrypting potentially sensitive demographic attributes in the data set showed evidence of statistically significant impact on predicted patient survival. The observed 9-10% impact was indicative of the relationship between the weight and ranking of the demographic attributes with respect to their influence on the patient survival and the extent of the effect on knowledge discovery. In practice, this was representative of the risk of basing treatment decisions using data sets where attributes may often be masked or encrypted for patient privacy and security concerns.

Using data mining tools to develop applications that will automatically rank attributes' relative information gain and alert clinicians to the impact that masked and/or encrypted attributes may have on the quality of data mining results use to base their treatment decisions, is a subject for further research and potential software development.

Chapter 5

Conclusions, Implications, Recommendations, and Summary

Introduction

The adoption of data mining technology continues to grow, but at the same time more private and sensitive information is also being protected using cryptographic techniques. As data mining becomes more prevalent as a decision-making tool, its trustworthiness become critically important. This research study examined the effect of data masking and encryption on the quality of data mining results measured as classification performance.

Data masking and encryption are two commonly used techniques employed to protect the confidentiality and integrity of private and sensitive data. Measuring the effect they had on data mining algorithms' classification performance, provided a metric to assess the impact on the quality of data mining results and their ability to harness the power of information to assist decision-making.

The research study used a comprehensive medical benchmark data set representative of the general U.S. population, with an extensive number of instances and attributes. As previous researchers had found when using WEKA to gauge algorithm performance, significant raw data preprocessing is required for proper experimentation (Ahmadi & Abadi, 2013; Blake & Mangiameli, 2011; Farhangfar, et al., 2008; Tiwari, Jha, & Yadav, 2012). Given that a large number of instances and a wide set of attributes were originally present in the raw data set, in order to focus on the objective of measuring classification performance parameters, significant data preprocessing and transformation was necessary. The process included the reduction in the number of instances and attributes.

Previous studies on classification performance had also shown that reduced number of attributes can improve classification accuracy among other performance parameters (Villacampa, 2015, Wilson & Rosen, 2003).

Developing a control group representative of instances that shared common attributes that ranked high on information gain value and influence they had on the outcome variable, two experimental groups were then derived. The derivative experimental masking group, applied treatment by replacing select attributes to emulate the effect of data masking. The experimental encryption group, applied treatment by suppressing values of the select attributes to emulate the effect of data encryption.

Conclusions

The methodology outlined in this research report was validated to be sound and provide a repeatable means by which classification performance could be measured across sets of algorithms. Results of testing revealed that classification performance parameters, obtained after training and cross validating the experimental groups, were lower on average than the same metrics calculated after training and cross validating the control group. These results were indicative of a higher number of correctly classified instances in the data sets where attributes were not substituted or suppressed through representative masking and encryption techniques. Drawing a parallel to findings made by Farhangfar, et al. (2008), which showed that classification with imputed values was more accurate than classification with missing values, the classification error rate, and therefore the classification accuracy, was found in this study to decrease, as treatment was applied to data sets and sensitive attributes were suppressed through masking and encryption.

Implications

The results of this research study provided an initial assessment that confirmed that data masking and encryption can impact the performance of classification algorithms in a statistically significant manner. As the use of data mining continues to increase as a decision-support tool in critical applications such as healthcare, its trustworthiness must be assured. While more personal and sensitive information is secured through masking and encryption to protect individual privacy, awareness of the effects that such data protection techniques can have on the dependability of data mining results is essential.

Recommendations

Based on the results obtained from this investigation, and the implications that the problem studied can have as data mining technologies see increasing adoption, additional research is recommended in the area of knowledge discovery quality metrics. Further research is necessary in fields such as zero-knowledge computing that enables extraction of insight from protected data without compromising confidentiality or integrity of the original sensitive information. Research into applications that can also map the degree to which protected attributes in data sets can potentially degrade derived knowledge is needed to fully capitalize on the potential of big data analytics. The main focus of future research should be to further the understanding of the interactions between data mining technology, analytics, and established and evolving data protection techniques such as masking, encryption, and new technologies such as blockchain.

Summary

Results obtained from this research study provided empirical indication that data masking and encryption can impact classification performance in a statistically

significant manner, and therefore affect the trustworthiness of data mining results. Performance parameters measured by four different classifiers delivered sizeable variations between the control group, where the data set attributes were untouched, and the two experimental groups where select attributes were substituted or suppressed to simulate the effects of data masking and encryption. The findings led to the rejection of the null hypothesis and the notion that the use of data masking and encryption do not necessarily have a detrimental effects on data mining algorithms' ability to extract valuable insight from large data sets.

Appendix A

SEER Data Record Description Summary

Table below lists all 121 attributes contained in original SEER breast cancer data set, including NAACCR name, item number, variable name, year, position and field length.

NAACCR Name	NAACCR Item #	SAS Variable Name	Applicable Years	Position	Length
Patient ID number	20	PUBCSNUM		1-8	8
Registry ID	40	REG		9-18	10
Marital Status at DX	150	MAR_STAT		19	1
Race/Ethnicity	160	RACE1V		20-21	2
NHIA Derived Hispanic Origin	191	NHIADE		23	1
Sex	220	SEX		24	1
Age at diagnosis	230	AGE_DX		25-27	3
Year of Birth	240	YR_BRTH		28-31	4
Sequence Number—Central	380	SEQ_NUM		35-36	2
Month of diagnosis	390	MDXRECOMP		37-38	2
Year of diagnosis	390	YEAR_DX		39-42	4
Primary Site	400	PRIMSITE		43-46	4
Laterality	410	LATERAL		47	1
Histology (92-00) ICD-O-2	420	HISTO2V		48-51	4
Behavior (92-00) ICD-O-2	430	BEHO2V		52	1
Histologic Type ICD-O-3	522	HISTO3V		53-56	4
Behavior Code ICD-O-3	523	BEHO3V		57	1
Grade	440	GRADE		58	1
Diagnostic Confirmation	490	DX_CONF		59	1
Type of Reporting Source	500	REPT_SRC		60	1
EOD—Tumor Size	780	EOD10_SZ	1988-2003	61-63	3
EOD—Extension	790	EOD10_EX	1988-2003	64-65	2
EOD—Extension Prost Path	800	EOD10_PE	1985-2003	66-67	2
EOD—Lymph Node Involv	810	EOD10_ND	1988-2003	68	1
Regional Nodes Positive	820	EOD10_PN	1988+	69-70	2
Regional Nodes Examined	830	EOD10_NE	1988+	71-72	2
EOD—Old 13 Digit	840	EOD13	1973-1982	73-85	13
EOD—Old 2 Digit	850	EOD2	1973-1982	86-87	2
EOD—Old 4 Digit	860	EOD4	1983-1987	88-91	4
Coding System for EOD	870	EOD_CODE	1973-2003	92	1
Tumor Marker 1	1150	TUMOR_1V	1990-2003	93	1
Tumor Marker 2	1160	TUMOR_2V	1990-2003	94	1
Tumor Marker 3	1170	TUMOR_3V	1998-2003	95	1
CS Tumor Size	2800	CSTUMSIZ	2004+	96-98	3
CS Extension	2810	CSEXTEN	2004+	99-101	3
CS Lymph Nodes	2830	CSLYMPHN	2004+	102-104	3
CS Mets at Dx	2850	CSMETSIX	2004+	105-106	2
CS Site-Specific Factor 1	2880	CS1SITE	2004+*	107-109	3
CS Site-Specific Factor 2	2890	CS2SITE	2004+*	110-112	3
CS Site-Specific Factor 3	2900	CS3SITE	2004+*	113-115	3
CS Site-Specific Factor 4	2910	CS4SITE	2004+*	116-118	3
CS Site-Specific Factor 5	2920	CS5SITE	2004+*	119-121	3

NAACCR Name	NAACCR Item #	SAS Variable Name	Applicable Years	Position	Length
CS Site-Specific Factor 6	2930	CS6SITE	2004+*	122-124	3
CS Site-Specific Factor 25	2879	CS25SITE	2004+*	125-127	3
Derived AJCC T	2940	DAJCCT	2004+	128-129	2
Derived AJCC N	2960	DAJCCN	2004+	130-131	2
Derived AJCC M	2980	DAJCCM	2004+	132-133	2
Derived AJCC Stage Group	3000	DAJCCSTG	2004+	134-135	2
Derived SS1977	3010	DSS1977S	2004+	136	1
Derived SS2000	3020	DSS2000S	2004+	137	1
Derived AJCC—Flag	3030	DAJCCFL	2004+	138	1
CS Version Input Original	2935	CSVFIRST	2004+	141-146	6
CS Version Derived	2936	CSVLATES	2004+	147-152	6
CS Version Input Current	2937	CSVCURRENT	2004+	153-158	6
RX Summ—Surg Prim Site	1290	SURGPRIF	1998+	159-160	2
RX Summ—Scope Reg LN Sur	1292	SURGSCOF	2003+	161	1
RX Summ—Surg Oth Reg/Dis	1294	SURGSITF	2003+	162	1
RX Summ—Reg LN Examined	1296	NUMNODES	1998-2002	163-164	2
Reason for no surgery	1340	NO_SURG		166	1
RX Summ—Radiation	1360	RADIATN		167	1
RX Summ—Rad to CNS	1370	RAD_BRN	1988-1997	168	1
RX Summ—Surg / Rad Seq	1380	RAD_SURG		169	1
RX Summ—Surgery Type	1640	SS_SURG	1973-1997	170-171	2
RX Summ—Scope Reg 98-02	1647	SURGSCOP	1998-2002	174	1
RX Summ—Surg Oth 98-02	1648	SURGSITE	1998-2002	175	1
SEER Record Number	2190	REC_NO		176-177	2
SEER Type of Follow-up	2180	TYPE_FU		191	1
Age Recode <1 Year olds	N/A	AGE_1REC		192-193	2
Site Recode ICD-O-3/WHO 2008	N/A	SITERWHO		199-203	5
Recode ICD-O-2 to 9	N/A	ICDOTO9V		204-207	4
Recode ICD-O-2 to 10	N/A	ICDOT10V		208-211	4
ICCC site recode ICD-O-3/WHO 2008	N/A	ICCC3WHO		218-220	3
ICCC site rec extended ICD-O-3/WHO 2008	N/A	ICCC3XWHO		221-223	3
Behavior Recode for Analysis	N/A	BEHTREND		224	1
Histology Recode—Broad Groupings	N/A	HISTREC		226-227	2
Histology Recode—Brain Groupings	N/A	HISTRECB		228-229	2
CS Schema v0204+	N/A	cs0204schema		230-232	3
Race recode (White, Black, Other)	N/A	RAC_RECA		233	1
Race recode (W, B, AI, API)	N/A	RAC_RECY		234	1
Origin recode NHIA (Hispanic, Non-Hisp)	N/A	ORIGRECB		235	1
SEER historic stage A	N/A	HST_STGA		236	1
AJCC stage 3 rd edition (1988-2003)	N/A	AJCC_STG		237-238	2

NAACCR Name	NAACCR Item #	SAS Variable Name	Applicable Years	Position	Length
SEER modified AJCC Stage 3 rd ed (1988-2003)	N/A	AJ_3SEER		239-240	2
SEER Summary Stage 1977 (1995-2000)	N/A	SSS77VZ	1995-2000	241	1
SEER Summary Stage 2000 (2001-2003)	N/A	SSSM2KPZ	2001-2003	242	1
First malignant primary indicator	N/A	FIRSTPRM		245	1
State-county recode	N/A	ST_CNTY		246-250	5
Cause of Death to SEER site recode	N/A	CODPUB		255-259	5
COD to site rec KM	N/A	CODPUBKM		260-264	5
Vital Status recode	N/A	STAT_REC		265	1
IHS Link	192	IHSLINK		266	1
Summary stage 2000 (1998+)	N/A	SUMM2K	1998+	267	1
AYA site recode/WHO 2008	N/A	AYASITERWHO		268-269	2
Lymphoma subtype recode/WHO 2008	N/A	LYMSUBRWHO		270-271	2
SEER Cause-Specific Death Classification	N/A	VSRTSADX		272	1
SEER Other Cause of Death Classification	N/A	ODTHCLASS		273	1
CS Tumor Size/Ext Eval	2820	CSTSEVAL	2004+	274	1
CS Lymph Nodes Eval	2840	CSRGEVAL	2004+	275	1
CS Mets Eval	2860	CSMTEVAL	2004+	276	1
Primary by international rules	N/A	intprim		277	1
ER Status Recode Breast Cancer (1990+)	N/A	erstatus	1990+	278	1
PR Status Recode Breast Cancer (1990+)	N/A	prstatus	1990+	279	1
CS Schema -AJCC 6 th ed (previously called v1)	N/A	csschema		280-281	2
CS Site-Specific Factor 8	2862	CS8SITE	2004+*	282-284	3
CS Site-Specific Factor 10	2864	CS10SITE	2004+*	285-287	3
CS Site-Specific Factor 11	2865	CS11SITE	2004+*	288-290	3
CS Site-Specific Factor 13	2867	CS13SITE	2004+*	291-293	3
CS Site-Specific Factor 15	2869	CS15SITE	2004+*	294-296	3
CS Site-Specific Factor 16	2870	CS16SITE	2004+*	297-299	3
Lymph vascular invasion	1182	VASINV	2004+*	300	1
Survival months	N/A	srv_time_mon		301-304	4
Survival months flag	N/A	srv_time_mon_flag		305	1
Insurance recode (2007+)	N/A	INSREC_PUB	2007+	311	1
Derived AJCC-7 T	3400	DAJCC7T	2010+	312-314	3
Derived AJCC-7 N	3410	DAJCC7N	2010+	315-317	3
Derived AJCC-7 M	3420	DAJCC7M	2010+	318-320	3
Derived AJCC-7 Stage Grp	3430	DAJCC7STG	2010+	321-323	3
Breast Adjusted AJCC 6 th T (1988+)	N/A	ADJTM_6VALUE	1988+	324-325	2
Breast Adjusted AJCC 6 th N (1988+)	N/A	ADJNM_6VALUE	1988+	326-327	2
Breast Adjusted AJCC 6 th M (1988+)	N/A	ADJTM_6VALUE	1988+	328-329	2
Breast Adjusted AJCC 6 th Stage (1988+)	N/A	ADJAJCCSTG	1988+	330-331	2

NAACCR Name	NAACCR Item #	SAS Variable Name	Applicable Years	Position	Length
CS Site-Specific Factor 7	2861	CS7SITE	2004+*	332-334	3
CS Site-Specific Factor 9	2863	CS9SITE	2004+*	335-337	3
CS Site-Specific Factor 12	2866	CS12SITE	2004+*	338-340	3
Derived HER2 Recode (2010+)	N/A	her2	2010+	341	1
Breast Subtype (2010+)	N/A	brst_sub	2010+	342	1
Lymphomas: Ann Arbor Staging (1983+)	N/A	ANNARBOR	1983+	348	1
CS Mets at Dx-Bone	2851	CSMETSDXB_PUB	2010+	349	1
CS Mets at Dx-Brain	2852	CSMETSDXBR_PUB	2010+	350	1
CS Mets at Dx-Liver	2853	CSMETSDXLIV_PUB	2010+	351	1
CS Mets at Dx-Lung	2854	CSMETSDXLUNG_PUB	2010+	352	1
T value - based on AJCC 3rd (1988-2003)	N/A	T_VALUE	1988-2003	353	2
N value - based on AJCC 3rd (1988-2003)	N/A	N_VALUE	1988-2003	355	2
M value - based on AJCC 3rd (1988-2003)	N/A	M_VALUE	1988-2003	357	2

*2004+ varying by schema

Appendix B

Attribute Evaluation and Ranking Results

Of the total 19 attributes included in the preprocessed breast cancer data set (18 plus the classification variable), the 12 highest ranking ones for information gain value were considered. The selected subset ensured that only attributes that significantly impacted the outcome variable (survival) were used in the experiment. Employing a method similar to the one used by Al-Bahrani, Agrawal, and Choudhary (2013), and Bellaachia and Guven (2006), the attributes' relative information gain were determined using "InfoGain AttributeEval" from the "Explorer" tab for "Select Attributes" in WEKA, and "Ranker" as the search method.

==== Run information ====

Evaluator: weka.attributeSelection.InfoGainAttributeEval
Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation: breast_cancer_100K
Instances: 100000
Attributes: 19
Evaluation mode: evaluate on all training data

==== Attribute Selection on all input data ====

Search Method: Attribute ranking.
Attribute Evaluator (supervised, Class (nominal): 19 SURVIVAL): Information
Gain Ranking Filter

Ranked attributes:

0.057939 6 YEAR OF BIRTH
0.049508 1 MARITAL STATUS AT DX
0.020809 14 RESON FOR NO SURGERY
0.019207 11 HISTOLOGIC TYPE ICD-O-3
0.016646 9 HISTOLOGY (92-00) ICD-O-2
0.01562 7 PRIMARY SITE
0.010745 15 RX SUMM-RADIATION
0.010725 17 SEER HISTORIC STAGE A
0.009326 16 RX SUMM-SURG/RAD SEQ
0.009017 2 RACE/ETHNICITY
0.005044 3 NHIA DERIVED HISPANIC ORIGIN
0.004056 5 AGE AT DIAGNOSIS
0.003196 10 BEHAVIOR (92-00) ICD-O-2
0.003196 12 BEHAVIOR CODE ICD-O-3
0.001368 8 LATERALITY
0.001227 13 DIAGNOSTIC CONFIRMATION
0.00083 18 FIRST MALIGNANT PRIMARY INDICATOR
0.000121 4 SEX

Appendix C

Experimental Setup

To initially create the control and experimental groups, the SEER Program Coding and Staging Manual (Adamo, Dickie, Ruhl, (2015) was used to determine the record format and the attribute headers. Once attribute headers were added, the raw breast cancer data set had to be preprocessed to conform to format requirements to enable the data mining tool to be used. A method modeled on one employed by Amado, Dickie, and Ruhl (2015) was followed. Since the SEER database comprises patient data across a 40 year period, different attributes had been added over time to the data set. It was therefore important to select only common attributes for the research. Downloading the breast cancer data set into Microsoft Excel[®], a comma separated value (CSV) file was created to initially preprocess and removed all attributes not common across the entire sample period.

Opening *breast_cancer_complete.csv*, the value for the survival attribute initially downloaded as a numeric value in months, was changed to a categorical value using a two-step process. First, an IF function was used to replace the value for each instance in the attribute column: a zero for all values < 60 months and a one for all values > 60 months. Step two used the “search and replace” feature to find and substitute all zero values for NO and all one values for YES. This change converted the dependent/ outcome variable to a binary nominal variable, in line with the objective of predicting accuracy of patient survival beyond 60 months from initial diagnosis. In order to facilitate the classification process, the dependent classification variable (survival in the case of

this experiment) must be set to a binary nominal value with only two potential outcomes (Tan, Steinbach, & Kumar, 2006). Using the SEER “other cause of death” classification, records of patients that died of reasons other than their cancer were also filtered out of the sample. For this purpose, instances with attribute code other than zero (indicating cancer) were deleted from the data set.

Following the framework of the KDD model, WEKA was used to further preprocess *breast_cancer_complete.csv* to create the control group data set, and to apply treatment to the predictor variables to create the experimental groups. Once preprocessing and transformation created the control and experimental data sets, WEKA was used run the four classification algorithms in sequence. Doing this first for the control and then for the experimental groups enabled the performance analysis to be conducted in an orderly manner, and results to be compiled and recorded for comparison.

To create the control and experimental data sets with the associated preprocessed data, the sample *breast_cancer_100k.csv* was loaded in WEKA using the “Explorer” interface. With the data set loaded, filters were applied to substitute or suppress select attributes to create the experimental group data sets. Since the filtering was performed prior to the classification process, unsupervised filters were used. The steps taken in WEKA to create the control data set are listed below:

1. Using “Unsupervised/Attribute/Remove,” attributes not falling in the top 12 ranking were removed from the data set. Attribute indices 4, 8, 10, 12, 13, and 18 were selected and removed.
2. Using “Unsupervised/Attribute/Numeric to Nominal,” numeric attributes in the data set were transformed into nominal values. Attribute indices 1-5 and 7-12

were selected and changed from numeric to nominal values. This was necessary since values represented a coding map and therefore could not be continuous.

Attributes included marital status at diagnosis, race/ethnicity, NHIA derived Hispanic origin, age at diagnosis, year of birth, histology (92-00) ICD-0-2, histologic type ICD-02-3, reason for no surgery, rx summary radiation, rx summary-surgery/radiation sequence, SEER historic stage A.

3. Using “Unsupervised/Attribute/Replace Missing Values,” absent nominal values in the data set were replaced with modes from the training data. Having completed the transformation process, the derived file was saved as the control data set:

breast_cancer_100k_control.csv.

To create the two experimental data sets, further transformation of the control data set was necessary using additional unsupervised filters to emulate the effects of data masking and encryption. The steps taken to create the two experimental data sets are listed below:

1. Using “Unsupervised/Attribute/Replace with Missing Value,” attribute indices 1-3 and 5 were selected and the probability changed from default 0.1 to 1.0. This replaced existing values for marital status at diagnosis, race/ethnicity, NHIA derived Hispanic origin, and year of birth with blank values. With these attribute values erased, “Unsupervised/Attribute/Replace Missing with User Constant” was selected, and again attribute indices 1-3 and 5 were specified one at a time. This replaced the missing values with randomly selected user-supplied nominal constant values already present in the control data set (i.e., 1 for single marital status, 01 for race white, 5 for other Hispanic origin, and 1972 for birth year). The changes represented the effect of masking the attributes. Having completed this

step, the experimental masked data set was finalized and the file saved as:

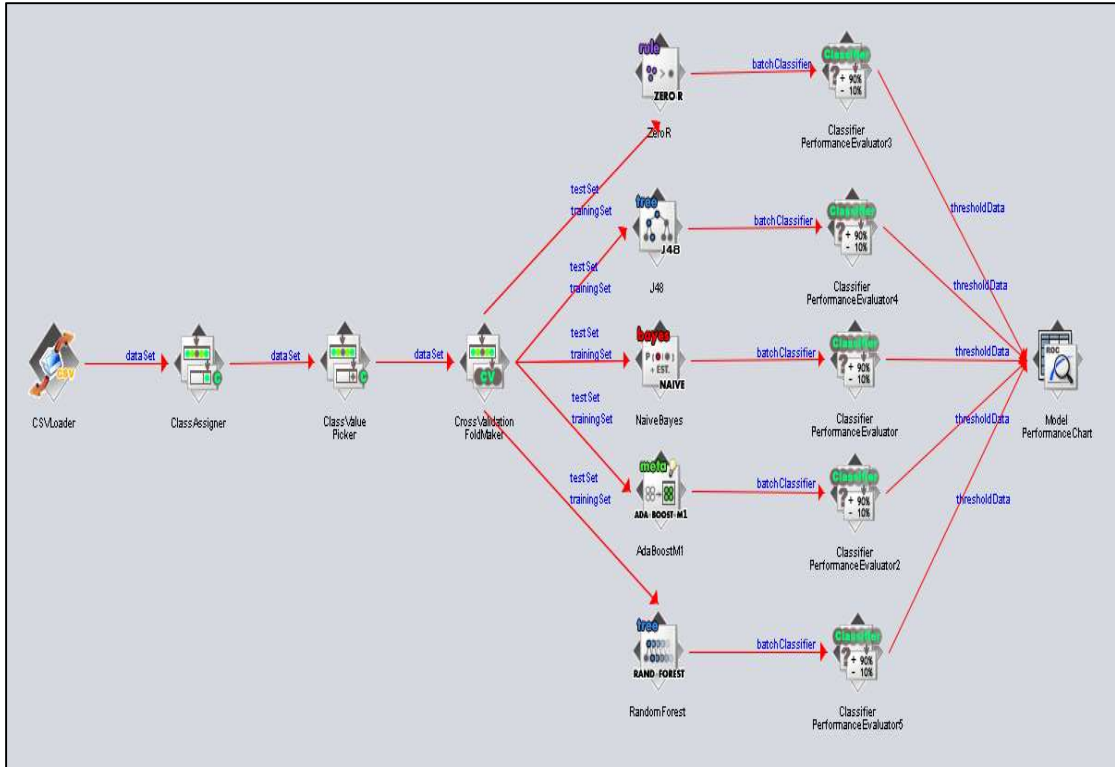
breast_cancer_100k_exp_mask.csv.

2. Using the filter “Unsupervised/Attribute/Replace with Missing Value,” missing values were introduced in the data set to emulate the effect of encryption.

Attribute indices 1-3 and 5 representing marital status at diagnosis, race/ethnicity, NHIA derived Hispanic origin, and year of birth were selected and the probability set to one to suppress the attribute values from the data set. Once this final step was completed, the experimental encrypted data set was finalized and the file saved as: *breast_cancer_100k_exp_encrypt.csv*.

Using WEKA’s “Explorer” interface, each of the classification algorithms were then executed on all three data sets. Associated metrics for weighted accuracy, precision, recall, and f-measure were calculated, and ROC graphs produced. Each data set was loaded from “Open File” under the “Preprocess” tab. To reduce the number of leaves in the decision trees, “MinNumObj” corresponding to the minimum number of instances considered per leaf in the tree, was increased from the default value of two. The higher the minimum number object, the smaller the tree. Recording test result, WEKA then computed the average value for algorithm’s accuracy.

To plot relative classifier performance values across the control group, experimental mask, and experimental encrypt groups, WEKA’s “Knowledge Flow” interface was used. An illustration of the process map showing each of the steps taken to overlay the ROC graphs produced by each of the classifiers is shown below.



Appendix D

Configuration of Classification Algorithms

Configuration Parameters Used for each of the Algorithms Employed in the Experiment

Parameter	ZeroR	J48	Naïve Bayes	AdaBoost	Random Forest
Batch Size	100	100	100	100	100
Binary Split		False			
Collapse Tree		True			
Confidence Factor		0.25			
Debug	False	False	False	False	False
Do Not Check Capabilities	False	False	False	False	False
Do Not Make Split Point Actual Value		False	False		
Minimum Number of Objects		2			
Minimum Decimal Places	2	2	2	2	2
Number of Folds		3			
Reduce Error Pruning		False			
Save Instance Data		False			
Use Kernel Estimator			False		
Use Supervised Discretization			False		
Number of Iterations				10	100
Seed		1		1	1
Use Reshaping				False	
Weight Threshold				100	

Appendix E

Classification Results

WEKA Results *breast_cancer_100k_control.csv* Data Set Using ZeroR

```

==== Run information ====
Scheme:   weka.classifiers.rules.ZeroR
Relation: breast_cancer_100K_control
Instances: 100000
Attributes: 13
Test mode: 10-fold cross-validation

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

ZeroR predicts class value: YES

Time taken to build model: 0.01 seconds

==== Stratified cross-validation ====
==== Summary ====
Correctly Classified Instances   50000      50   %
Incorrectly Classified Instances 50000      50   %
Kappa statistic                  0
Mean absolute error              0.5
Root mean squared error          0.5
Relative absolute error          100   %
Root relative squared error      100   %
Total Number of Instances       100000

==== Detailed Accuracy By Class ====
TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
1.000    1.000    0.500     1.000  0.667     0.000  0.500    0.500    YES
0.000    0.000    0.000     0.000  0.000     0.000  0.500    0.500    NO
Weighted Avg.
0.500    0.500    0.250     0.500  0.333     0.000  0.500    0.500

==== Confusion Matrix ====
  a  b <-- classified as
50000  0 | a = YES
50000  0 | b = NO

```

Associated cost/benefit ROC graphs for ZeroR

Weka Classifier: Cost/Benefit Analysis - rules.ZeroR (class = NO)

X: Sample Size (Num) Y: True Positive Rate (Num) X: Sample Size (Num) Y: Cost/Benefit (Num)
 Colour: Threshold (Num) Select Instance Colour: Threshold (Num) Select Instance

Res... Clear Open Save Jitter Res... Clear Open Save Jitter

Plot: ThresholdCurve

Plot: Cost/Benefit Curve

Threshold

% of Population % of Target (recall) Score Threshold

% of Population: 0 % of Target: 0
 Score Threshold: 0.5

Confusion Matrix

	Predicted (a)	Predicted (b)	
Actual (a): NO	0 0%	50000 50%	
Actual (b): YES	0 0%	50000 50%	

Classification Accuracy: 50%

Cost Matrix

	Predicted (a)	Predicted (b)	Actual (a)
	0.0	1.0	
	1.0	0.0	

Actual (a)
Actual (b)

Total Population: 100000

Cost: 50000
 Random: 50000
 Gain: 0

Maximize Cost/Benefit
 Minimize Cost/Benefit

Cost Benefit

WEKA Results *breast_cancer_100k_control.csv* Data Set Using J48

```
==== Run information ====
```

```
Scheme:   weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: breast_cancer_100K_control
Instances: 100000
Attributes: 13
Test mode: 10-fold cross-validation
```

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

```
J48 pruned tree
```

```
Number of Leaves : 2303
```

```
Size of the tree : 4136
```

```
Time taken to build model: 19.37 seconds
```

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

```
==== Summary ====
```

```
Correctly Classified Instances 78448 78.448 %
Incorrectly Classified Instances 21552 21.552 %
Kappa statistic 0.569
Mean absolute error 0.2701
Root mean squared error 0.3894
Relative absolute error 54.0106 %
Root relative squared error 77.8867 %
Total Number of Instances 100000
```

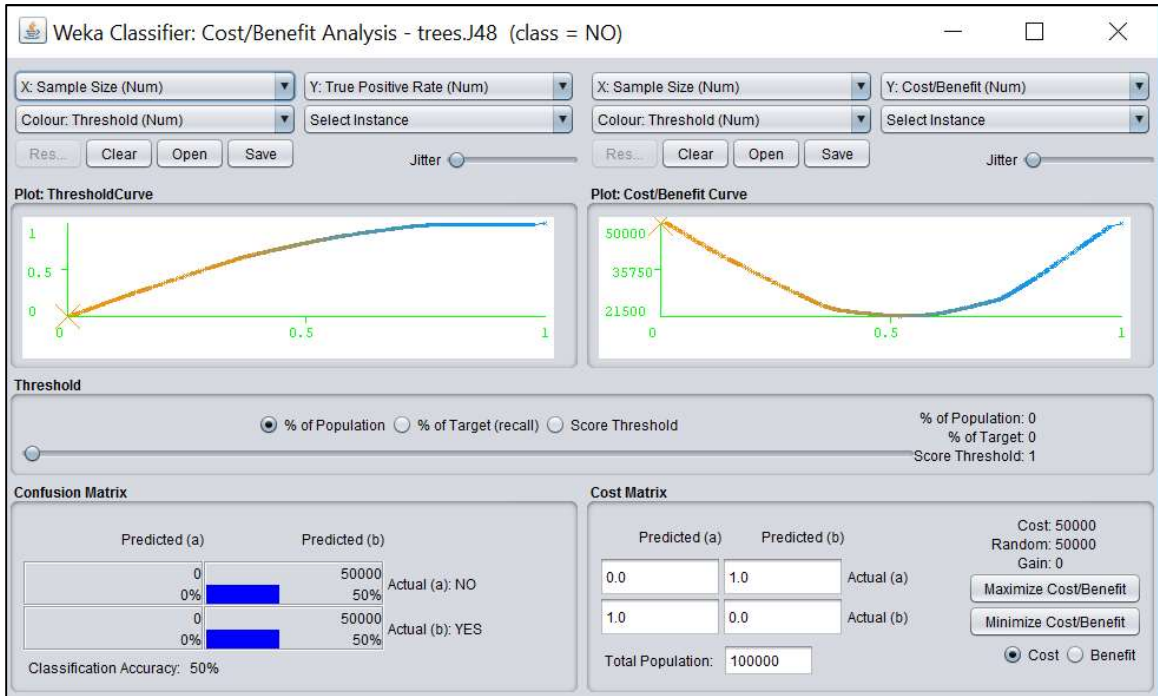
```
==== Detailed Accuracy By Class ====
```

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.753	0.184	0.803	0.753	0.778	0.570	0.858	0.846	YES
0.816	0.247	0.768	0.816	0.791	0.570	0.858	0.808	NO
Weighted Avg.								
0.784	0.216	0.786	0.784	0.784	0.570	0.858	0.827	

```
==== Confusion Matrix ====
```

```
  a  b <-- classified as
37663 12337 |  a = YES
9215 40785 |  b = NO
```

Associated cost/benefit ROC graphs for J48



WEKA Results *breast_cancer_100k_control.csv* Data Set Using Naïve Bayes

(Using Supervised Filter for Attribute Discretization)

==== Run information ====

Scheme: weka.classifiers.bayes.NaiveBayes
 Relation: breast_cancer_100K_control-
 weka.filters.supervised.attribute.Discretize-R1-5,7-12-precision6
 Instances: 100000
 Attributes: 13
 Test mode: 10-fold cross-validation

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

Naive Bayes ClassifierTime taken to build model: 0.01 seconds

==== Stratified cross-validation ====

==== Summary ====

Correctly Classified Instances	67495	67.495 %
Incorrectly Classified Instances	32505	32.505 %
Kappa statistic	0.3499	
Mean absolute error	0.3926	
Root mean squared error	0.457	
Relative absolute error	78.5149 %	
Root relative squared error	91.399 %	
Total Number of Instances	100000	

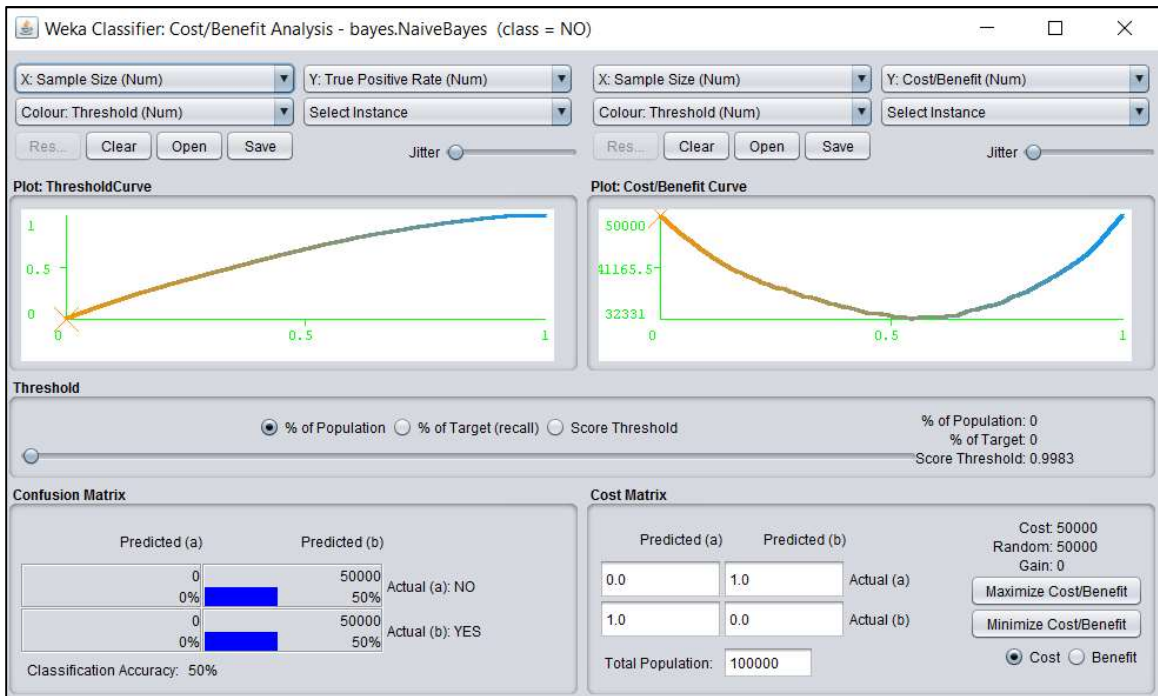
==== Detailed Accuracy By Class ====

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.617	0.267	0.698	0.617	0.655	0.352	0.738	0.751	YES
0.733	0.383	0.657	0.733	0.693	0.352	0.738	0.717	NO
Weighted Avg.								
0.675	0.325	0.677	0.675	0.674	0.352	0.738	0.734	

==== Confusion Matrix ====

a	b	<-- classified as
30862	19138	a = YES
13367	36633	b = NO

Associated cost/benefit ROC graphs for Naïve Bayes



(Using Unsupervised Filter for Attribute Transformation from Numeric to Nominal)

```

==== Run information ====
Scheme:   weka.classifiers.bayes.NaiveBayes
Relation: breast_cancer_100K_control-
weka.filters.unsupervised.attribute.NumericToNominal-R1-5,7-12-
weka.filters.supervised.attribute.Discretize-Rfirst-last-precision6
Instances: 100000
Attributes: 13
Test mode: 10-fold cross-validation

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

Naive Bayes Classifier
Time taken to build model: 0.04 seconds

==== Stratified cross-validation ====
==== Summary ====
Correctly Classified Instances   67570           67.57 %
Incorrectly Classified Instances  32430           32.43 %
Kappa statistic                  0.3514
Mean absolute error              0.3914
Root mean squared error          0.4568
Relative absolute error          78.2782 %
Root relative squared error      91.3584 %
Total Number of Instances       100000

==== Detailed Accuracy By Class ====
TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
0.613    0.262    0.701     0.613   0.654     0.354  0.739    0.753     YES
0.738    0.387    0.656     0.738   0.695     0.354  0.739    0.717     NO
Weighted Avg.
0.676    0.324    0.678     0.676   0.674     0.354  0.739    0.735

==== Confusion Matrix ====
  a  b <-- classified as
30667 19333 |  a = YES
13097 36903 |  b = NO

```

Associated cost/benefit ROC graphs for Naïve Bayes

Weka Classifier: Cost/Benefit Analysis - bayes.NaiveBayes (class = NO)

X: Sample Size (Num) Y: True Positive Rate (Num) X: Sample Size (Num) Y: Cost/Benefit (Num)

Colour: Threshold (Num) Select Instance Colour: Threshold (Num) Select Instance

Res... Clear Open Save Jitter Res... Clear Open Save Jitter

Plot: ThresholdCurve

Plot: Cost/Benefit Curve

Threshold

% of Population % of Target (recall) Score Threshold

% of Population: 0 % of Target: 0 Score Threshold: 0.9993

Confusion Matrix

		Predicted (a)	Predicted (b)	
Actual (a): NO	0		50000	50%
	0%			
Actual (b): YES	0		50000	50%
	0%			

Classification Accuracy: 50%

Cost Matrix

		Predicted (a)	Predicted (b)	
Actual (a)	0.0			
	1.0			
Actual (b)	0.0			
	1.0			

Total Population: 100000

Cost: 50000
Random: 50000
Gain: 0

Maximize Cost/Benefit
Minimize Cost/Benefit

Cost Benefit

WEKA Results *breast_cancer_100k_control.csv* Data Set Using AdaBoost

```
==== Run information ====
```

```
Scheme: weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 10 -W
weka.classifiers.trees.DecisionStump
Relation: breast_cancer_100K_control
Instances: 100000
Attributes: 13
Test mode: 10-fold cross-validation
```

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

```
AdaBoostM1: Base classifiers and their weights:
```

```
Number of performed Iterations: 10
```

```
Time taken to build model: 3.01 seconds
```

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

```
==== Summary ====
```

Correctly Classified Instances	72809	72.809 %
Incorrectly Classified Instances	27191	27.191 %
Kappa statistic	0.4562	
Mean absolute error	0.4245	
Root mean squared error	0.4452	
Relative absolute error	84.9052 %	
Root relative squared error	89.0478 %	
Total Number of Instances	100000	

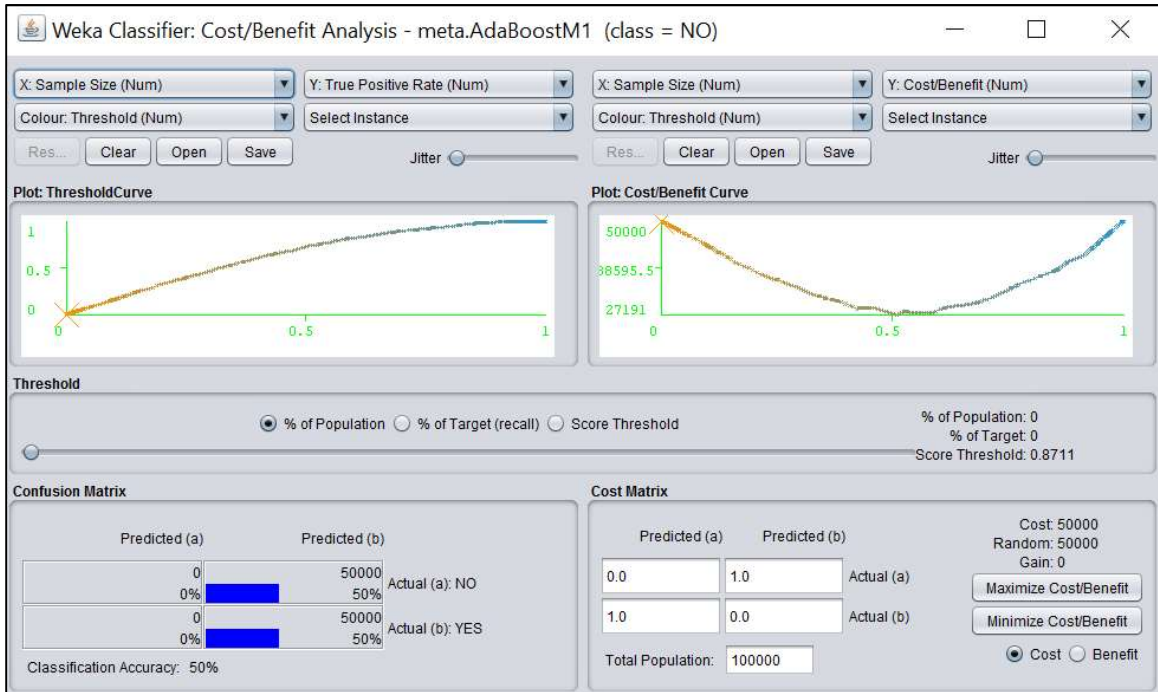
```
==== Detailed Accuracy By Class ====
```

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.718	0.261	0.733	0.718	0.725	0.456	0.786	0.799	YES
0.739	0.282	0.723	0.739	0.731	0.456	0.786	0.742	NO
Weighted Avg.								
0.728	0.272	0.728	0.728	0.728	0.456	0.786	0.770	

```
==== Confusion Matrix ====
```

```
  a  b <-- classified as
35879 14121 |  a = YES
13070 36930 |  b = NO
```

Associated cost/benefit ROC graphs for AdaBoost



WEKA Results *breast_cancer_100k_control.csv* Data Set Using Random Forest

```

=== Run information ===
Scheme:   weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -
M 1.0 -V 0.001 -S 1
Relation: breast_cancer_100K_control
Instances: 100000
Attributes: 13
Test mode: 10-fold cross-validation

```

```

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

```

```

RandomForest

```

```

Bagging with 100 iterations and base learner

```

```

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

```

```

Time taken to build model: 44.5 seconds

```

```

=== Stratified cross-validation ===

```

```

=== Summary ===

```

Correctly Classified Instances	77501	77.501 %
Incorrectly Classified Instances	22499	22.499 %
Kappa statistic	0.55	
Mean absolute error	0.2711	
Root mean squared error	0.3942	
Relative absolute error	54.2228 %	
Root relative squared error	78.8323 %	
Total Number of Instances	100000	

```

=== Detailed Accuracy By Class ===

```

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.756	0.206	0.786	0.756	0.771	0.550	0.860	0.875	YES
0.794	0.244	0.765	0.794	0.779	0.550	0.860	0.839	NO
Weighted Avg.								
0.775	0.225	0.775	0.775	0.775	0.550	0.860	0.857	

```

=== Confusion Matrix ===

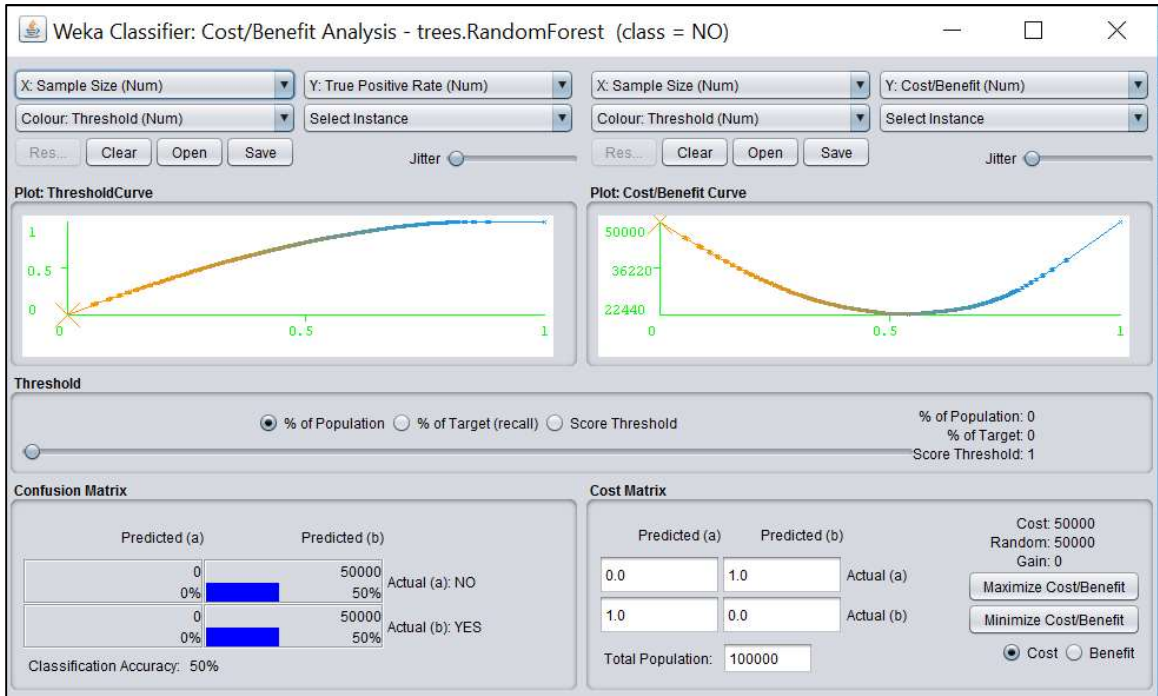
```

```

  a   b  <-- classified as
37786 12214 |  a = YES
10285 39715 |  b = NO

```

Associated cost/benefit ROC graphs for Random Forest



WEKA Results *breast_cancer_100k_exp_mask.csv* Data Set Using ZeroR

```

==== Run information ====
Scheme:   weka.classifiers.rules.ZeroR
Relation: breast_cancer_100K_exp-mask
Instances: 100000
Attributes: 13
Test mode: 10-fold cross-validation

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

ZeroR predicts class value: YES

Time taken to build model: 0.01 seconds

==== Stratified cross-validation ====
==== Summary ====
Correctly Classified Instances   50000      50  %
Incorrectly Classified Instances 50000      50  %
Kappa statistic                  0
Mean absolute error              0.5
Root mean squared error          0.5
Relative absolute error          100  %
Root relative squared error      100  %
Total Number of Instances       100000

==== Detailed Accuracy By Class ====
TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
1.000    1.000    0.500     1.000   0.667     0.000  0.500    0.500    YES
0.000    0.000    0.000     0.000   0.000     0.000  0.500    0.500    NO
Weighted Avg.
0.500    0.500    0.250     0.500   0.333     0.000  0.500    0.500

==== Confusion Matrix ====
  a  b <-- classified as
50000  0 |  a = YES
50000  0 |  b = NO


```

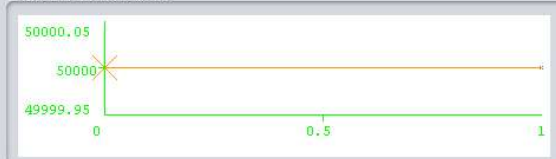
Associated cost/benefit ROC graphs for ZeroR

Weka Classifier: Cost/Benefit Analysis - rules.ZeroR (class = NO)

X: Sample Size (Num) Y: True Positive Rate (Num) X: Sample Size (Num) Y: Cost/Benefit (Num)
 Colour: Threshold (Num) Select Instance Colour: Threshold (Num) Select Instance

Res... Clear Open Save Jitter Res... Clear Open Save Jitter

Plot: ThresholdCurve


Plot: Cost/Benefit Curve


Threshold
 % of Population % of Target (recall) Score Threshold
 % of Population: 0 % of Target: 0 Score Threshold: 0.5

Confusion Matrix

		Predicted (a)	Predicted (b)	
Actual (a): NO	0	0%	50000	
	0%	0%	50%	
Actual (b): YES	0	0%	50000	
	0%	0%	50%	

Classification Accuracy: 50%

Cost Matrix

		Predicted (a)	Predicted (b)	
Actual (a)	0.0	1.0		
	1.0	0.0		
Actual (b)				

Total Population: 100000

Cost: 50000
 Random: 50000
 Gain: 0

Maximize Cost/Benefit
 Minimize Cost/Benefit

Cost Benefit

WEKA Results *breast_cancer_100k_exp_mask.csv* Data Set Using J48

```
==== Run information ====
```

```
Scheme:   weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: breast_cancer_100K_exp-mask
Instances: 100000
Attributes: 13
```

```
Test mode: 10-fold cross-validation
```

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

```
J48 pruned tree
```

```
Number of Leaves : 415
```

```
Size of the tree : 745
```

```
Time taken to build model: 6.47 seconds
```

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

```
==== Summary ====
```

Correctly Classified Instances	65214	65.214 %
Incorrectly Classified Instances	34786	34.786 %
Kappa statistic	0.3043	
Mean absolute error	0.4288	
Root mean squared error	0.466	
Relative absolute error	85.7604 %	
Root relative squared error	93.2005 %	
Total Number of Instances	100000	

```
==== Detailed Accuracy By Class ====
```

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.620	0.315	0.663	0.620	0.640	0.305	0.707	0.703	YES
0.685	0.380	0.643	0.685	0.663	0.305	0.707	0.682	NO
Weighted Avg.								
0.652	0.348	0.653	0.652	0.652	0.305	0.707	0.693	

```
==== Confusion Matrix ====
```

```
  a  b <-- classified as
30983 19017 |  a = YES
15769 34231 |  b = NO
```

Associated cost/benefit ROC graphs for J48

Weka Classifier: Cost/Benefit Analysis - trees.J48 (class = NO)

X: Sample Size (Num) Y: True Positive Rate (Num) X: Sample Size (Num) Y: Cost/Benefit (Num)

Colour: Threshold (Num) Select Instance Colour: Threshold (Num) Select Instance

Res... Clear Open Save Jitter Res... Clear Open Save Jitter

Plot: ThresholdCurve

Plot: Cost/Benefit Curve

Threshold

% of Population % of Target (recall) Score Threshold

% of Population: 0 % of Target: 0
Score Threshold: 1

Confusion Matrix

		Predicted (a)	Predicted (b)	
Actual (a): NO	0	0	50000	
	0%		50%	
Actual (b): YES	0	0	50000	
	0%		50%	

Classification Accuracy: 50%

Cost Matrix

		Predicted (a)	Predicted (b)	
Actual (a)	0.0			
	1.0			
Actual (b)	0.0			
	1.0			

Total Population: 100000

Cost: 50000
Random: 50000
Gain: 0

Maximize Cost/Benefit
Minimize Cost/Benefit

Cost Benefit

WEKA Results *breast_cancer_100k_exp_mask.csv* Data Set Using Naïve Bayes

(Using Supervised Filter for Attribute Discretization)

```

=== Run information ===
Scheme:   weka.classifiers.bayes.NaiveBayes
Relation: breast_cancer_100K_exp-mask-
weka.filters.supervised.attribute.Discretize-R1-5,7-12-precision6
Instances: 100000
Attributes: 13
Test mode: 10-fold cross-validation

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

Naive Bayes Classifier
Time taken to build model: 0.02 seconds

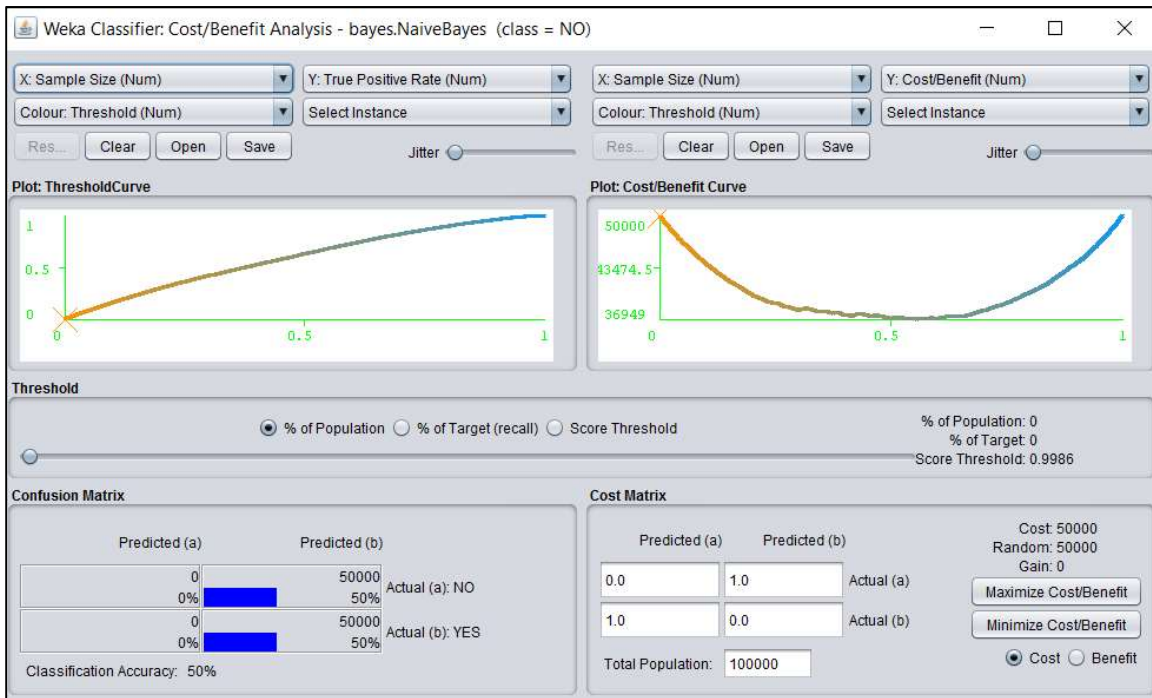
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances   62825           62.825 %
Incorrectly Classified Instances 37175           37.175 %
Kappa statistic                  0.2565
Mean absolute error              0.4199
Root mean squared error          0.4769
Relative absolute error          83.9861 %
Root relative squared error      95.3704 %
Total Number of Instances       100000

=== Detailed Accuracy By Class ===
TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
0.620    0.364    0.630     0.620  0.625     0.257 0.690    0.685    YES
0.636    0.380    0.626     0.636  0.631     0.257 0.690    0.682    NO
Weighted Avg.
0.628    0.372    0.628     0.628  0.628     0.257 0.690    0.684

=== Confusion Matrix ===
  a  b <-- classified as
31011 18989 |  a = YES
18186 31814 |  b = NO

```

Associated cost/benefit ROC graphs for Naïve Bayes



(Using Unsupervised Filter for Attribute Transformation from Numeric to Nominal)

```

==== Run information ====
Scheme:   weka.classifiers.bayes.NaiveBayes
Relation: breast_cancer_100K_exp-mask-
weka.filters.unsupervised.attribute.NumericToNominal-R1-5,7-12
Instances: 100000
Attributes: 13

Test mode: 10-fold cross-validation

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

Naive Bayes Classifier
Time taken to build model: 0.03 seconds

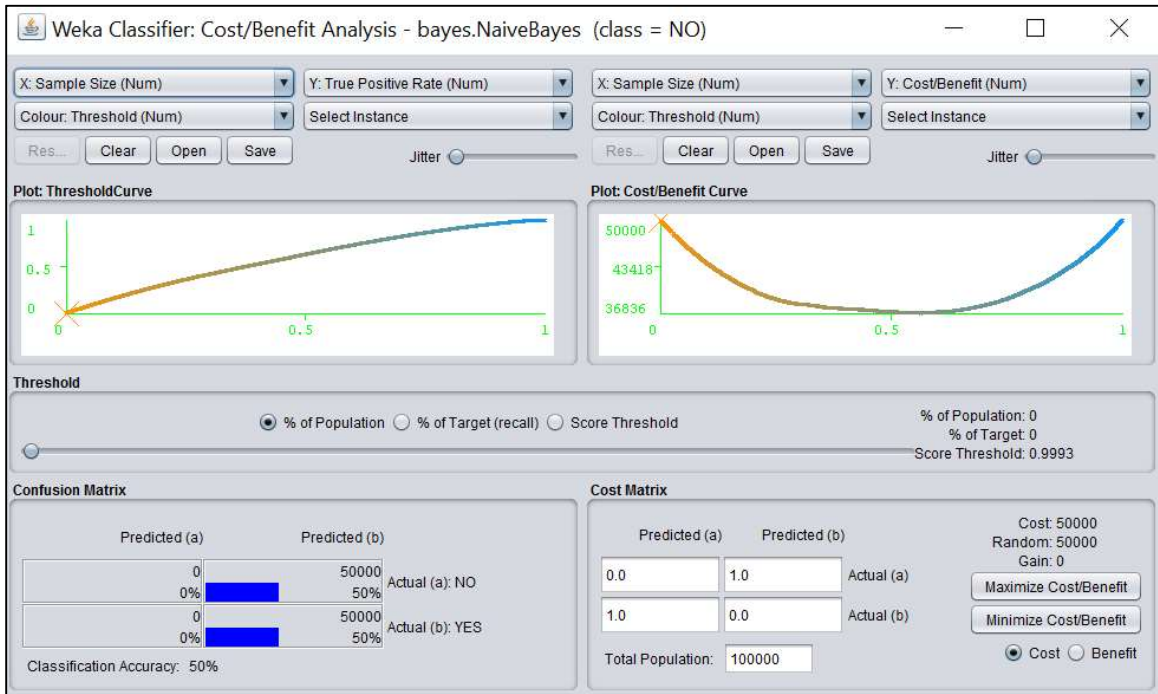
==== Stratified cross-validation ====
==== Summary ====
Correctly Classified Instances   62999           62.999 %
Incorrectly Classified Instances 37001           37.001 %
Kappa statistic                  0.26
Mean absolute error              0.4187
Root mean squared error          0.4765
Relative absolute error          83.7481 %
Root relative squared error      95.2915 %
Total Number of Instances       100000

==== Detailed Accuracy By Class ====
TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
0.621    0.361    0.632     0.621  0.626     0.260 0.691    0.686    YES
0.639    0.379    0.628     0.639  0.633     0.260 0.691    0.684    NO
Weighted Avg.
0.630    0.370    0.630     0.630  0.630     0.260 0.691    0.685

==== Confusion Matrix ====
  a   b  <-- classified as
31032 18968 |   a = YES
18033 31967 |   b = NO

```

Associated cost/benefit ROC graphs for Naïve Bayes



WEKA Results *breast_cancer_100k_exp_mask.csv* Data Set Using AdaBoost

```
==== Run information ====
```

```
Scheme:   weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 10 -W
weka.classifiers.trees.DecisionStump
Relation:  breast_cancer_100K_exp-mask
Instances: 100000
Attributes: 13
```

```
Test mode: 10-fold cross-validation
```

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

```
AdaBoostM1: Base classifiers and their weights:
```

```
Decision StumpNumber of performed Iterations: 10
```

```
Time taken to build model: 2.69 seconds
```

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

```
==== Summary ====
```

```
Correctly Classified Instances   60200      60.2 %
Incorrectly Classified Instances  39800      39.8 %
Kappa statistic                  0.204
Mean absolute error              0.4601
Root mean squared error          0.4784
Relative absolute error          92.0204 %
Root relative squared error      95.6769 %
Total Number of Instances       100000
```

```
==== Detailed Accuracy By Class ====
```

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.755	0.551	0.578	0.755	0.655	0.214	0.667	0.659	YES
0.449	0.245	0.647	0.449	0.530	0.214	0.667	0.654	NO
Weighted Avg.								
0.602	0.398	0.613	0.602	0.592	0.214	0.667	0.656	

```
==== Confusion Matrix ====
```

```
  a  b  <-- classified as
37758 12242 |  a = YES
27558 22442 |  b = NO
```

Associated cost/benefit ROC graphs for AdaBoost

Weka Classifier: Cost/Benefit Analysis - meta.AdaBoostM1 (class = NO)

X: Sample Size (Num) Y: True Positive Rate (Num) X: Sample Size (Num) Y: Cost/Benefit (Num)

Colour: Threshold (Num) Select Instance Colour: Threshold (Num) Select Instance

Res... Clear Open Save Jitter Res... Clear Open Save Jitter

Plot: ThresholdCurve

Plot: Cost/Benefit Curve

Threshold

% of Population % of Target (recall) Score Threshold

% of Population: 0
% of Target: 0
Score Threshold: 0.7842

Confusion Matrix

		Predicted (a)	Predicted (b)	
Actual (a): NO	0	0%	50000	
	0%		50%	
Actual (b): YES	0	0%	50000	
	0%		50%	

Classification Accuracy: 50%

Cost Matrix

		Predicted (a)	Predicted (b)	
Actual (a)	0.0		1.0	
	1.0		0.0	
Actual (b)	0.0		1.0	
	1.0		0.0	

Total Population: 100000

Cost: 50000
Random: 50000
Gain: 0

Maximize Cost/Benefit
Minimize Cost/Benefit

Cost Benefit

WEKA Results *breast_cancer_100k_exp_mask.csv* Data Set Using Random Forest

==== Run information ====

Scheme: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1

Relation: breast_cancer_100K_exp-mask

Instances: 100000

Attributes: 13

Test mode: 10-fold cross-validation

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

RandomForest

Bagging with 100 iterations and base learner

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

Time taken to build model: 46.23 seconds

==== Stratified cross-validation ====

==== Summary ====

Correctly Classified Instances	63208	63.208 %
Incorrectly Classified Instances	36792	36.792 %
Kappa statistic	0.2642	
Mean absolute error	0.417	
Root mean squared error	0.4825	
Relative absolute error	83.4027 %	
Root relative squared error	96.4907 %	
Total Number of Instances	100000	

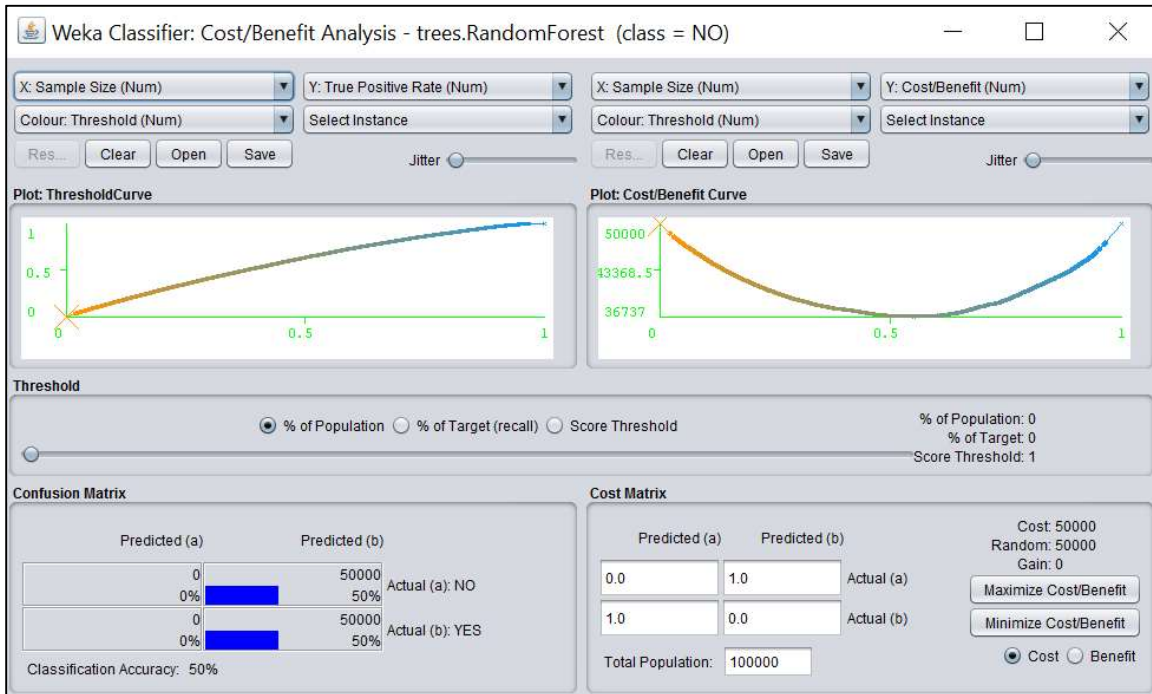
==== Detailed Accuracy By Class ====

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.617	0.353	0.636	0.617	0.627	0.264	0.684	0.688	YES
0.647	0.383	0.628	0.647	0.637	0.264	0.684	0.665	NO
Weighted Avg.								
0.632	0.368	0.632	0.632	0.632	0.264	0.684	0.676	

==== Confusion Matrix ====

a	b	<-- classified as
30870	19130	a = YES
17662	32338	b = NO

Associated cost/benefit ROC graphs for Random Forest



WEKA Results *breast_cancer_100k_exp_encrypt.csv* Data Set Using ZeroR

```
==== Run information ====
```

```
Scheme:   weka.classifiers.rules.ZeroR
Relation: breast_cancer_100K_exp_encrypt-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1
Instances: 100000
Attributes: 9
```

```
Test mode: 10-fold cross-validation
```

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

```
ZeroR predicts class value: YES
```

```
Time taken to build model: 0.01 seconds
```

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

```
==== Summary ====
```

Correctly Classified Instances	50000	50	%
Incorrectly Classified Instances	50000	50	%
Kappa statistic	0		
Mean absolute error	0.5		
Root mean squared error	0.5		
Relative absolute error	100	%	
Root relative squared error	100	%	
Total Number of Instances	100000		

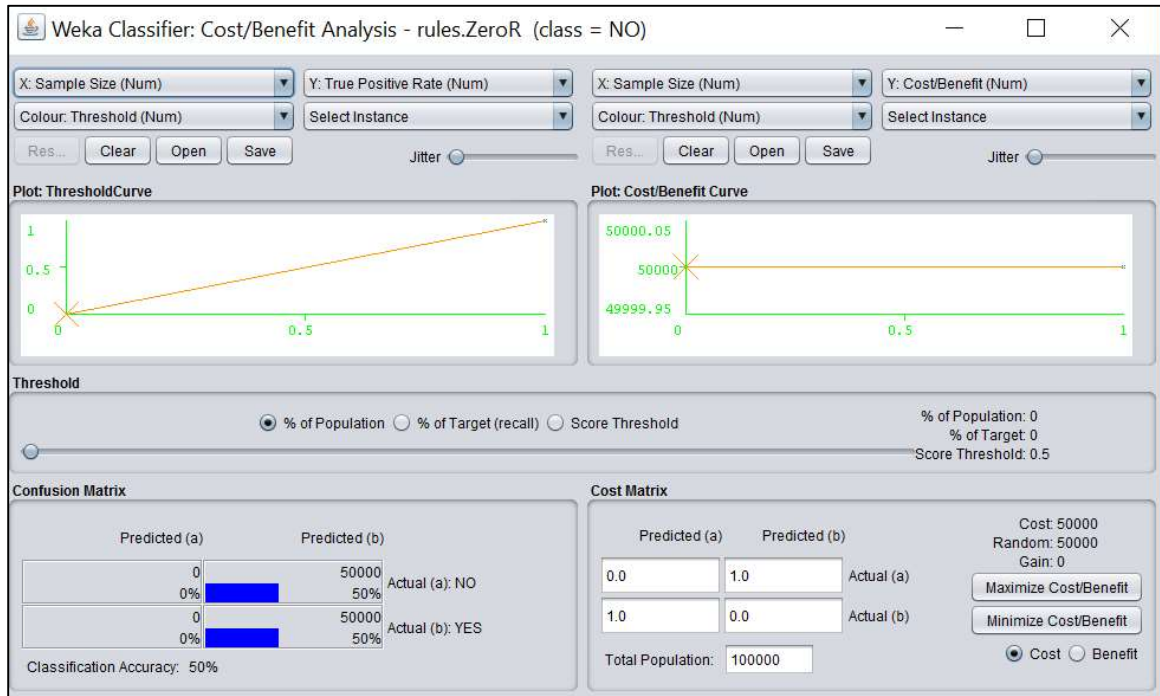
```
==== Detailed Accuracy By Class ====
```

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
1.000	1.000	0.500	1.000	0.667	0.000	0.500	0.500	YES
0.000	0.000	0.000	0.000	0.000	0.000	0.500	0.500	NO
Weighted Avg.								
0.500	0.500	0.250	0.500	0.333	0.000	0.500	0.500	

```
==== Confusion Matrix ====
```

```
  a  b <-- classified as
50000  0 | a = YES
50000  0 | b = NO
```

Associated cost/benefit ROC graphs for ZeroR



WEKA Results *breast_cancer_100k_exp_encrypt.csv* Data Set Using J48

==== Run information ====

Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: breast_cancer_100K_exp_encrypt-

weka.filters.unsupervised.attribute.Remove-R1-

weka.filters.unsupervised.attribute.Remove-R1-

weka.filters.unsupervised.attribute.Remove-R1-

weka.filters.unsupervised.attribute.Remove-R1

Instances: 100000

Attributes: 9

Test mode: 10-fold cross-validation

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

J48 pruned tree

Number of Leaves : 415

Size of the tree : 745

Time taken to build model: 5.66 seconds

==== Stratified cross-validation ====

==== Summary ====

Correctly Classified Instances 65214 65.214 %

Incorrectly Classified Instances 34786 34.786 %

Kappa statistic 0.3043

Mean absolute error 0.4288

Root mean squared error 0.466

Relative absolute error 85.7604 %

Root relative squared error 93.2005 %

Total Number of Instances 100000

==== Detailed Accuracy By Class ====

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.620	0.315	0.663	0.620	0.640	0.305	0.707	0.703	YES
0.685	0.380	0.643	0.685	0.663	0.305	0.707	0.682	NO
Weighted Avg.								
0.652	0.348	0.653	0.652	0.652	0.305	0.707	0.693	

==== Confusion Matrix ====

a b <-- classified as

30983 19017 | a = YES

15769 34231 | b = NO

Associated cost/benefit ROC graphs for J48

Weka Classifier: Cost/Benefit Analysis - trees.J48 (class = NO)

X: Sample Size (Num) Y: True Positive Rate (Num) X: Sample Size (Num) Y: Cost/Benefit (Num)

Colour: Threshold (Num) Select Instance Colour: Threshold (Num) Select Instance

Res... Clear Open Save Jitter Res... Clear Open Save Jitter

Plot: ThresholdCurve

Plot: Cost/Benefit Curve

Threshold

% of Population % of Target (recall) Score Threshold

% of Population: 0 % of Target: 0
Score Threshold: 1

Confusion Matrix

		Predicted (a)	Predicted (b)	
Actual (a): NO	0	0%	50000	
	0%		50%	
Actual (b): YES	0	0%	50000	
	0%		50%	

Classification Accuracy: 50%

Cost Matrix

		Predicted (a)	Predicted (b)	
Actual (a)	0.0		1.0	
	1.0		0.0	
Actual (b)	0.0		1.0	
	1.0		0.0	

Total Population: 100000

Cost: 50000
Random: 50000
Gain: 0

Maximize Cost/Benefit
Minimize Cost/Benefit

Cost Benefit

WEKA Results *breast_cancer_100k_exp_encrypt.csv* Data Set Using Naïve Bayes

(Using Supervised Filter for Attribute Discretization)

==== Run information ====

Scheme: weka.classifiers.bayes.NaiveBayes
 Relation: breast_cancer_100K_control-
 weka.filters.supervised.attribute.Discretize-Rfirst-last-precision6-
 weka.filters.unsupervised.attribute.ReplaceWithMissingValue-R1-3,5-S1-P1.0
 Instances: 100000
 Attributes: 13
 Test mode: 10-fold cross-validation

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

Naive Bayes ClassifierTime taken to build model: 0.04 seconds

==== Stratified cross-validation ====

==== Summary ====

Correctly Classified Instances	61593	61.593 %
Incorrectly Classified Instances	38407	38.407 %
Kappa statistic	0.2319	
Mean absolute error	0.4441	
Root mean squared error	0.4838	
Relative absolute error	88.8229 %	
Root relative squared error	96.7636 %	
Total Number of Instances	100000	

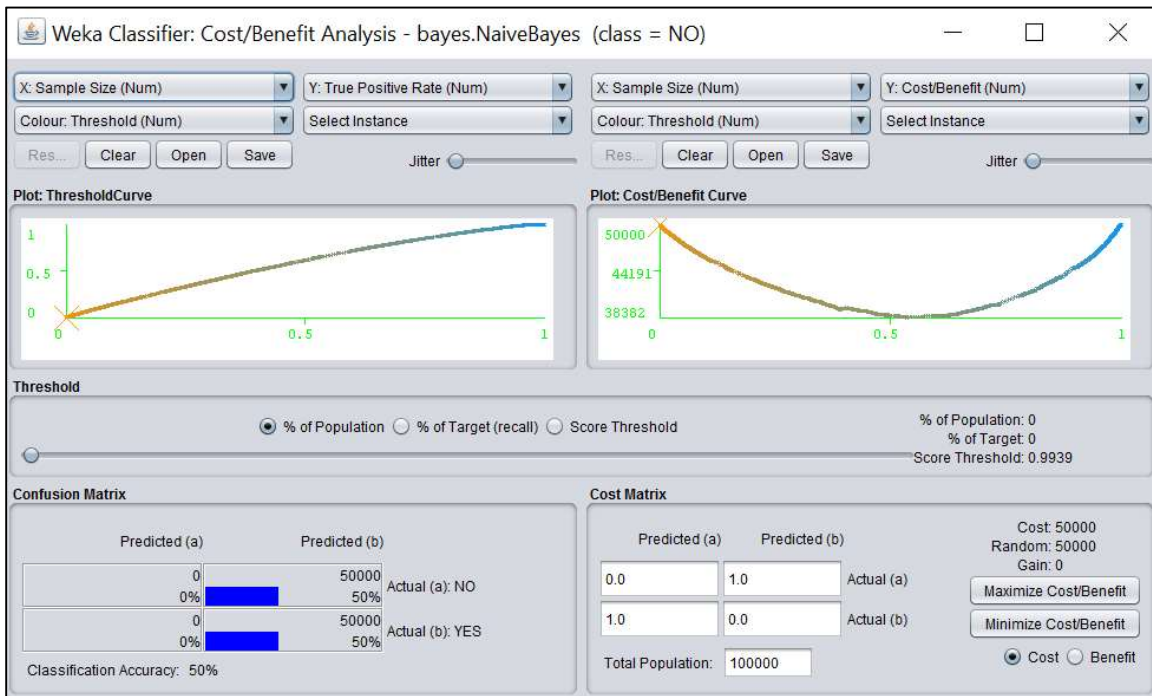
==== Detailed Accuracy By Class ====

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.585	0.353	0.624	0.585	0.604	0.232	0.658	0.662	YES
0.647	0.415	0.609	0.647	0.627	0.232	0.658	0.635	NO
Weighted Avg.								
0.616	0.384	0.616	0.616	0.616	0.232	0.658	0.648	

==== Confusion Matrix ====

a	b	<-- classified as
29244	20756	a = YES
17651	32349	b = NO

Associated cost/benefit ROC graphs for Naïve Bayes



(Using Unsupervised Filter for Attribute Transformation from Numeric to Nominal)

=== Run information ===

Scheme: weka.classifiers.bayes.NaiveBayes
 Relation: breast_cancer_100K_exp_encrypt-
 weka.filters.unsupervised.attribute.Remove-R1-
 weka.filters.unsupervised.attribute.Remove-R1-
 weka.filters.unsupervised.attribute.Remove-R1-
 weka.filters.unsupervised.attribute.Remove-R1-
 weka.filters.unsupervised.attribute.NumericToNominal-R1,3-8
 Instances: 100000
 Attributes: 9

Test mode: 10-fold cross-validation

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

Naive Bayes Classifier

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	62999	62.999 %
Incorrectly Classified Instances	37001	37.001 %
Kappa statistic	0.26	
Mean absolute error	0.4187	
Root mean squared error	0.4765	
Relative absolute error	83.7481 %	
Root relative squared error	95.2915 %	
Total Number of Instances	100000	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.621	0.361	0.632	0.621	0.626	0.260	0.691	0.686	YES
0.639	0.379	0.628	0.639	0.633	0.260	0.691	0.684	NO
Weighted Avg.								
0.630	0.370	0.630	0.630	0.630	0.260	0.691	0.685	

=== Confusion Matrix ===

a	b	<-- classified as
31032	18968	a = YES
18033	31967	b = NO

Associated cost/benefit ROC graphs for Naïve Bayes

Weka Classifier: Cost/Benefit Analysis - bayes.NaiveBayes (class = NO)

X: Sample Size (Num) Y: True Positive Rate (Num) X: Sample Size (Num) Y: Cost/Benefit (Num)

Colour: Threshold (Num) Select Instance Colour: Threshold (Num) Select Instance

Res... Clear Open Save Jitter Res... Clear Open Save Jitter

Plot: ThresholdCurve

Plot: Cost/Benefit Curve

Threshold

% of Population % of Target (recall) Score Threshold

% of Population: 0 % of Target: 0
Score Threshold: 0.9993

Confusion Matrix

		Predicted (a)	Predicted (b)	
Actual (a): NO	0		50000	
	0%		50%	
Actual (b): YES	0		50000	
	0%		50%	

Classification Accuracy: 50%

Cost Matrix

		Predicted (a)	Predicted (b)	
Actual (a)	0.0			
	1.0			
Actual (b)	0.0			
	1.0			

Total Population: 100000

Cost: 50000
Random: 50000
Gain: 0

Maximize Cost/Benefit
Minimize Cost/Benefit

Cost Benefit

WEKA Results *breast_cancer_100k_exp_encrypt.csv* Data Set Using AdaBoost

==== Run information ====

Scheme: weka.classifiers.meta.AdaBoostM1 -P 100 -S 1 -I 10 -W
weka.classifiers.trees.DecisionStump

Relation: breast_cancer_100K_exp_encrypt-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1

Instances: 100000

Attributes: 9

Test mode: 10-fold cross-validation

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

AdaBoostM1: Base classifiers and their weights:

Decision Stump

Number of performed Iterations: 10

Time taken to build model: 2.12 seconds

==== Stratified cross-validation ====

==== Summary ====

Correctly Classified Instances	60200	60.2 %
Incorrectly Classified Instances	39800	39.8 %
Kappa statistic	0.204	
Mean absolute error	0.4601	
Root mean squared error	0.4784	
Relative absolute error	92.0204 %	
Root relative squared error	95.6769 %	
Total Number of Instances	100000	

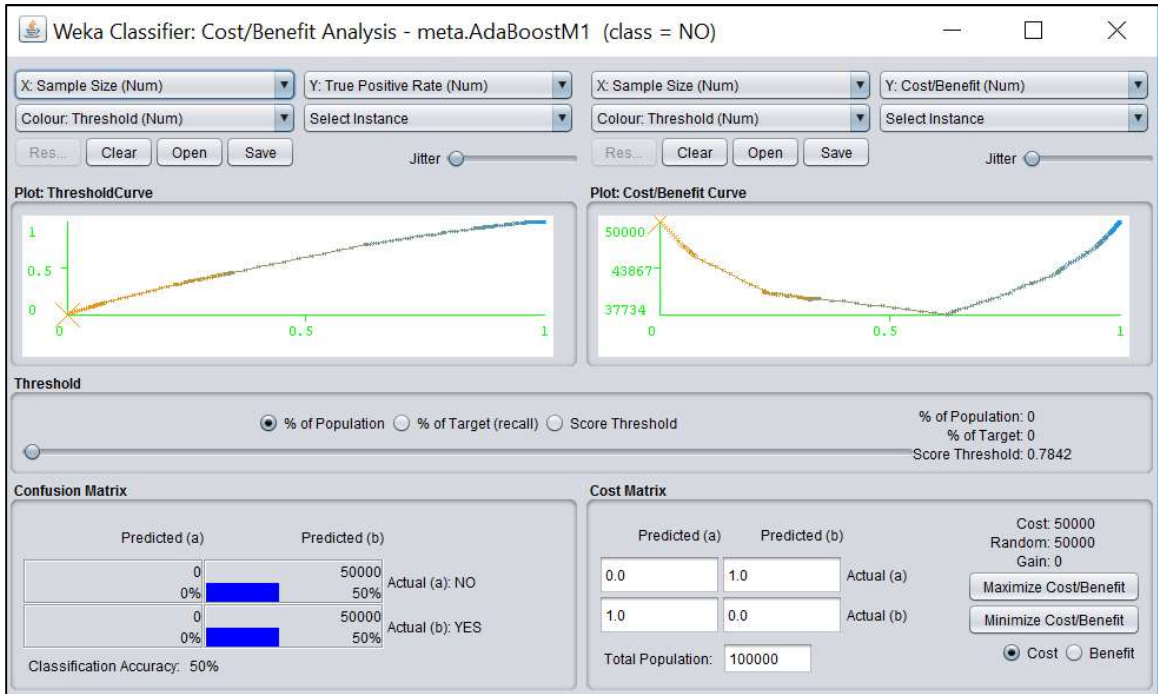
==== Detailed Accuracy By Class ====

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.755	0.551	0.578	0.755	0.655	0.214	0.667	0.659	YES
0.449	0.245	0.647	0.449	0.530	0.214	0.667	0.654	NO
Weighted Avg.								
0.602	0.398	0.613	0.602	0.592	0.214	0.667	0.656	

==== Confusion Matrix ====

a	b	<-- classified as
37758	12242	a = YES
27558	22442	b = NO

Associated cost/benefit ROC graphs for AdaBoost



WEKA Results *breast_cancer_100k_exp_encrypt.csv* Data Set Using Random Forest

```
=== Run information ===
```

```
Scheme: weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1
```

```
Relation: breast_cancer_100K_exp_encrypt-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1
```

```
Instances: 100000
```

```
Attributes: 9
```

```
Test mode: 10-fold cross-validation
```

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

```
RandomForest
```

```
Bagging with 100 iterations and base learner
```

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

```
Time taken to build model: 40.84 seconds
```

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

```
=== Summary ===
```

Correctly Classified Instances	63240	63.24 %
Incorrectly Classified Instances	36760	36.76 %
Kappa statistic	0.2648	
Mean absolute error	0.4172	
Root mean squared error	0.4822	
Relative absolute error	83.4497 %	
Root relative squared error	96.4401 %	
Total Number of Instances	100000	

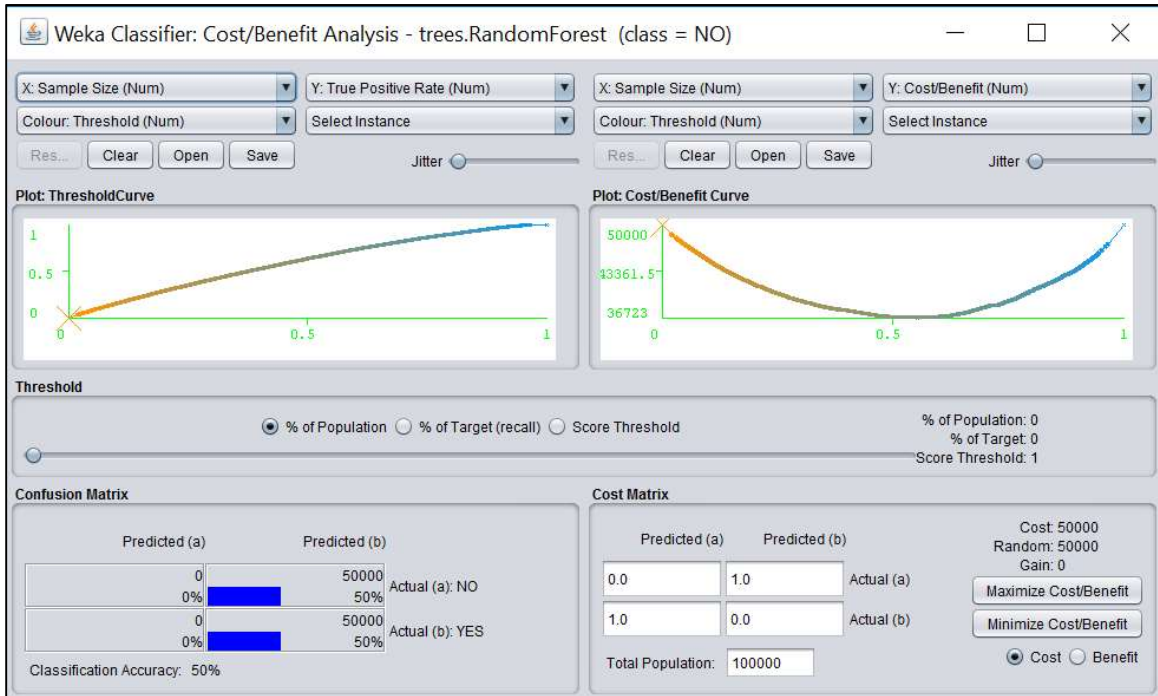
```
=== Detailed Accuracy By Class ===
```

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.617	0.353	0.636	0.617	0.627	0.265	0.684	0.688	YES
0.647	0.383	0.629	0.647	0.638	0.265	0.684	0.665	NO
Weighted Avg.								
0.632	0.368	0.633	0.632	0.632	0.265	0.684	0.677	

```
=== Confusion Matrix ===
```

```
  a  b <-- classified as
30872 19128 |  a = YES
17632 32368 |  b = NO
```

Associated cost/benefit ROC graphs for Random Forest



Appendix F

Data Analysis

Statistical Significance (Using 12 Most Influential Attributes)					
Group / Classifier	Accuracy	Precision	Recall	F-Measure	ROC/AUC
CONTROL					
J48	0.78448	0.786	0.784	0.784	0.858
Naïve Bayes (Non-Sup)	0.67570	0.678	0.676	0.674	0.739
Naïve Bayes (Sup Dis)	0.67495	0.677	0.675	0.674	0.378
Ada Boost	0.72809	0.728	0.728	0.728	0.786
Random Forest	0.77501	0.775	0.775	0.775	0.860
MASK					
J48	0.65214	0.653	0.652	0.652	0.707
Naïve Bayes (Non-Sup)	0.62999	0.630	0.630	0.630	0.691
Naïve Bayes (Sup Dis)	0.62825	0.628	0.628	0.628	0.690
Ada Boost	0.60200	0.613	0.602	0.592	0.667
Random Forest	0.63208	0.632	0.632	0.632	0.684
ENCRYPT					
J48	0.65214	0.653	0.652	0.652	0.707
Naïve Bayes (Non-Sup)	0.62999	0.630	0.630	0.630	0.691
Naïve Bayes (Sup Dis)	0.61593	0.616	0.616	0.616	0.658
Ada Boost	0.60200	0.613	0.602	0.592	0.667
Random Forest	0.63240	0.633	0.632	0.632	0.684

ANOVA: Two-Factor With Replication

SUMMARY	Accuracy	Precision	Recall	F-Measure	ROC/AUC	Total
<i>CONTROL</i>						
Count	5	5	5	5	5	25
Sum	3.63823	3.644	3.638	3.635	3.621	18.1762
Average	0.72765	0.7288	0.7276	0.727	0.7242	0.72705
Variance	0.00274	0.00267	0.00271	0.002793	0.0400592	0.0085
<i>MASK</i>						
Count	5	5	5	5	5	25
Sum	3.14446	3.156	3.144	3.134	3.439	16.0175
Average	0.62889	0.6312	0.6288	0.6268	0.6878	0.6407
Variance	0.00032	0.0002	0.00032	0.000471	0.0002077	0.00083
<i>ENCRYPT</i>						
Count	5	5	5	5	5	25
Sum	3.13246	3.145	3.132	3.122	3.407	15.9385
Average	0.62649	0.629	0.6264	0.6244	0.6814	0.63754
Variance	0.00035	0.00025	0.00035	0.000493	0.0003773	0.00081
<i>Total</i>						
Count	15	15	15	15	15	
Sum	9.91515	9.945	9.914	9.891	10.467	
Average	0.66101	0.663	0.66093	0.6594	0.6978	
Variance	0.00335	0.00321	0.00335	0.003523	0.0119933	
<i>ANOVA</i>						
<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Sample	0.12899	2	0.06449	17.80935	8.477E-07	3.15041
Columns	0.01627	4	0.00407	1.123413	0.3540183	2.52522
Interaction	0.00978	8	0.00122	0.337504	0.9479039	2.09697
Within	0.21728	60	0.00362			
Total	0.37232	74				

Variability or Measured Results by Group			
	CONTROL	MASK	ENCRYPT
	0.7845	0.6521	0.6521
	0.7860	0.6530	0.6530
	0.7840	0.6520	0.6520
	0.7840	0.6520	0.6520
	0.8580	0.7070	0.7070
	0.6757	0.6300	0.6300
	0.6780	0.6300	0.6300
	0.6760	0.6300	0.6300
	0.6740	0.6300	0.6300
	0.7390	0.6910	0.6910
	0.6750	0.6283	0.6159
	0.6770	0.6280	0.6160
	0.6750	0.6280	0.6160
	0.6740	0.6280	0.6160
	0.3780	0.6900	0.6580
	0.7281	0.6020	0.6020
	0.7280	0.6130	0.6130
	0.7280	0.6020	0.6020
	0.7280	0.5920	0.5920
	0.7860	0.6670	0.6670
	0.7750	0.6321	0.6324
	0.7750	0.6320	0.6330
	0.7750	0.6320	0.6320
	0.7750	0.6320	0.6320
	0.8600	0.6840	0.6840
<i>Sum</i>	18.1762	16.0175	15.9385
<i>Mean</i>	0.7270	0.6407	0.6375
<i>Variance</i>	0.0085	0.0008	0.0008

Tukey Multiple Comparison			
	Q_u	3.384	
	<i>Numerator df</i>	3	<i>Denominator df</i> 72
<i>Comparison</i>	<i>Absolute Comparison</i>	<i>Critical Range</i>	<i>Result</i>
Control to Mask	0.08635	0.0393455	Significantly Different
Control to Encrypt	0.08951	0.0393455	Significantly Different
Mask to Encrypt	0.00316	0.0393455	Not Significantly Different

<i>Factor Levels:</i>	3	Total # Groups
<i>n</i>	75	Total # Observations
<i>n.</i>	25	# of Observations in One Particular Group
Q_u	3.384	From Studentized Range Distribution Table
s^2_{pooled}	0.0034	Average Variance Across Groups Equals ANOVA <i>MS</i> Within Groups
<i>Critical Range</i>	0.03934548	$Q_u \sqrt{(s^2_{pooled} / n.)}$

Appendix G

Critical Value Table

The studentized range distribution table defines the value for the basic statistic Q_u used in the Tukey HSD multiple comparison test (Student, 1927).

<i>Critical Values of Studentized Range Distribution(q) for Familywise ALPHA = .05.</i>								
Denominator	Number of Groups (a.k.a. Treatments)							
DF	3	4	5	6	7	8	9	10
1	26.976	32.819	37.081	40.407	43.118	45.397	47.356	49.070
2	8.331	9.798	10.881	11.734	12.434	13.027	13.538	13.987
3	5.910	6.825	7.502	8.037	8.478	8.852	9.177	9.462
4	5.040	5.757	6.287	6.706	7.053	7.347	7.602	7.826
5	4.602	5.218	5.673	6.033	6.330	6.582	6.801	6.995
6	4.339	4.896	5.305	5.629	5.895	6.122	6.319	6.493
7	4.165	4.681	5.060	5.359	5.606	5.815	5.997	6.158
8	4.041	4.529	4.886	5.167	5.399	5.596	5.767	5.918
9	3.948	4.415	4.755	5.024	5.244	5.432	5.595	5.738
10	3.877	4.327	4.654	4.912	5.124	5.304	5.460	5.598
11	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.486
12	3.773	4.199	4.508	4.748	4.947	5.116	5.262	5.395
13	3.734	4.151	4.453	4.690	4.884	5.049	5.192	5.318
14	3.701	4.111	4.407	4.639	4.829	4.990	5.130	5.253
15	3.673	4.076	4.367	4.595	4.782	4.940	5.077	5.198
16	3.649	4.046	4.333	4.557	4.741	4.896	5.031	5.150
17	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108
18	3.609	3.997	4.276	4.494	4.673	4.824	4.955	5.071
19	3.593	3.977	4.253	4.468	4.645	4.794	4.924	5.037
20	3.578	3.958	4.232	4.445	4.620	4.768	4.895	5.008
21	3.565	3.942	4.213	4.424	4.597	4.743	4.870	4.981
22	3.553	3.927	4.196	4.405	4.577	4.722	4.847	4.957
23	3.542	3.914	4.180	4.388	4.558	4.702	4.826	4.935
24	3.532	3.901	4.166	4.373	4.541	4.684	4.807	4.915
25	3.523	3.890	4.153	4.358	4.526	4.667	4.789	4.897
26	3.514	3.880	4.141	4.345	4.511	4.652	4.773	4.880
27	3.506	3.870	4.130	4.333	4.498	4.638	4.758	4.864
28	3.499	3.861	4.120	4.322	4.486	4.625	4.745	4.850
29	3.493	3.853	4.111	4.311	4.475	4.613	4.732	4.837
30	3.487	3.845	4.102	4.301	4.464	4.601	4.720	4.824
31	3.481	3.838	4.094	4.292	4.454	4.591	4.709	4.813
32	3.475	3.832	4.086	4.284	4.445	4.581	4.698	4.802
33	3.470	3.825	4.079	4.276	4.436	4.572	4.689	4.791
34	3.465	3.820	4.072	4.268	4.428	4.563	4.680	4.782
35	3.461	3.814	4.066	4.261	4.421	4.555	4.671	4.773
36	3.457	3.809	4.060	4.255	4.414	4.547	4.663	4.764
37	3.453	3.804	4.054	4.249	4.407	4.540	4.655	4.756
38	3.449	3.799	4.049	4.243	4.400	4.533	4.648	4.749
39	3.445	3.795	4.044	4.237	4.394	4.527	4.641	4.741
40	3.442	3.791	4.039	4.232	4.388	4.521	4.634	4.735
41	3.439	3.787	4.035	4.227	4.383	4.515	4.628	4.728
42	3.436	3.783	4.030	4.222	4.378	4.509	4.622	4.722
43	3.433	3.779	4.026	4.217	4.373	4.504	4.617	4.716
44	3.430	3.776	4.022	4.213	4.368	4.499	4.611	4.710
45	3.428	3.773	4.018	4.209	4.364	4.494	4.606	4.705
46	3.425	3.770	4.015	4.205	4.359	4.489	4.601	4.700
47	3.423	3.767	4.011	4.201	4.355	4.485	4.597	4.695
48	3.420	3.764	4.008	4.197	4.351	4.481	4.592	4.690
49	3.418	3.761	4.005	4.194	4.347	4.477	4.588	4.686
50	3.416	3.758	4.002	4.190	4.344	4.473	4.584	4.681

<i>Critical Values of Studentized Range Distribution(q) for Familywise ALPHA = .05.</i>								
Denominator	Number of Groups (a.k.a. Treatments)							
DF	3	4	5	6	7	8	9	10
51	3.414	3.756	3.999	4.187	4.340	4.469	4.580	4.677
52	3.412	3.753	3.996	4.184	4.337	4.465	4.576	4.673
53	3.410	3.751	3.994	4.181	4.334	4.462	4.572	4.669
54	3.408	3.749	3.991	4.178	4.331	4.459	4.569	4.666
55	3.406	3.747	3.989	4.176	4.328	4.455	4.566	4.662
56	3.405	3.745	3.986	4.173	4.325	4.452	4.562	4.659
57	3.403	3.743	3.984	4.170	4.322	4.449	4.559	4.656
58	3.402	3.741	3.982	4.168	4.319	4.447	4.556	4.652
59	3.400	3.739	3.979	4.165	4.317	4.444	4.553	4.649
60	3.399	3.737	3.977	4.163	4.314	4.441	4.550	4.646
61	3.397	3.735	3.975	4.161	4.312	4.439	4.548	4.643
62	3.396	3.734	3.973	4.159	4.309	4.436	4.545	4.641
63	3.395	3.732	3.972	4.157	4.307	4.434	4.542	4.638
64	3.393	3.730	3.970	4.155	4.305	4.431	4.540	4.635
65	3.392	3.729	3.968	4.153	4.303	4.429	4.538	4.633
66	3.391	3.727	3.966	4.151	4.301	4.427	4.535	4.630
67	3.390	3.726	3.965	4.149	4.299	4.425	4.533	4.628
68	3.389	3.725	3.963	4.147	4.297	4.423	4.531	4.626
69	3.387	3.723	3.962	4.146	4.295	4.421	4.529	4.624
70	3.386	3.722	3.960	4.144	4.293	4.419	4.527	4.621
71	3.385	3.721	3.959	4.142	4.291	4.417	4.525	4.619
72	3.384	3.719	3.957	4.141	4.290	4.415	4.523	4.617
73	3.383	3.718	3.956	4.139	4.288	4.413	4.521	4.615
74	3.382	3.717	3.954	4.138	4.286	4.411	4.519	4.613
75	3.382	3.716	3.953	4.136	4.285	4.410	4.517	4.611
76	3.381	3.715	3.952	4.135	4.283	4.408	4.515	4.610
77	3.380	3.714	3.951	4.133	4.282	4.406	4.514	4.608
78	3.379	3.713	3.949	4.132	4.280	4.405	4.512	4.606
79	3.378	3.712	3.948	4.131	4.279	4.403	4.511	4.604
80	3.377	3.711	3.947	4.129	4.278	4.402	4.509	4.603
81	3.377	3.710	3.946	4.128	4.276	4.400	4.507	4.601
82	3.376	3.709	3.945	4.127	4.275	4.399	4.506	4.600
83	3.375	3.708	3.944	4.126	4.274	4.398	4.504	4.598
84	3.374	3.707	3.943	4.125	4.272	4.396	4.503	4.597
85	3.374	3.706	3.942	4.123	4.271	4.395	4.502	4.595
86	3.373	3.705	3.941	4.122	4.270	4.394	4.500	4.594
87	3.372	3.704	3.940	4.121	4.269	4.392	4.499	4.592
88	3.372	3.704	3.939	4.120	4.268	4.391	4.498	4.591
89	3.371	3.703	3.938	4.119	4.266	4.390	4.496	4.590
90	3.370	3.702	3.937	4.118	4.265	4.389	4.495	4.588
91	3.370	3.701	3.936	4.117	4.264	4.388	4.494	4.587
92	3.369	3.700	3.935	4.116	4.263	4.387	4.493	4.586
93	3.368	3.700	3.934	4.115	4.262	4.386	4.492	4.585
94	3.368	3.699	3.934	4.114	4.261	4.384	4.491	4.583
95	3.367	3.698	3.933	4.114	4.260	4.383	4.489	4.582
96	3.367	3.698	3.932	4.113	4.259	4.382	4.488	4.581
97	3.366	3.697	3.931	4.112	4.258	4.381	4.487	4.580
98	3.366	3.696	3.930	4.111	4.257	4.380	4.486	4.579
99	3.365	3.696	3.930	4.110	4.257	4.379	4.485	4.578
100	3.365	3.695	3.929	4.109	4.256	4.379	4.484	4.577

Reference List

- Abraham, R., Simha, J. B., & Iyengar, S. S. (2006, December). A comparative analysis of discretization methods for medical data mining with Naïve Bayesian classifier. *9th IEEE International Conference on Information Technology*, 235-236.
- Acuña, E., & Rodríguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. *Classification, Clustering, and Data Mining Applications*, 639-647. Springer.
- Adamo M., Dickie, L., & Ruhl, J. (2015, January). SEER program coding and staging manual 2015. *National Cancer Institute*, Bethesda, MD.
- Agarwal, N., & Yiliyasi, Y. (2010, November). Information quality challenges in social media. *International Conference on Information Quality*, Little Rock, Arkansas.
- Ahmadi, F., & Abadi, M. E. S. A. (2013). Data mining in teacher evaluation system using WEKA. *International Journal of Computer Applications*, 63(10), 12-18.
- Al-Aidaros, K. M., Bakar, A. A., & Othman, Z. (2012). Medical data classification with Naïve Bayes approach. *Information Technology Journal*, 11(9), 1166-1174.
- Al-Badrashiny, M. & Bellaachia, A. (2016). Breast cancer survivability prediction via classifier ensemble. *World Academy of Science, Engineering and Technology, International Science Index 113, International Journal of Computer, Electrical, Automation, Control and Information Engineering*, 10(5), 799-803.
- Al-Bahrani, R., Agrawal, A., & Choudhary, A. (2013, October). Colon cancer survival prediction using ensemble data mining on SEER data. *IEEE International Conference on Big Data*, 9-16.
- Alkharboush, N. A. H. (2013). A data mining approach to improve the automated quality of data (Doctoral dissertation). Retrieved from Queensland University of Technology. http://eprints.qut.edu.au/65641/1/Nawaf%20Abdullah%20H_Alkharboush_Thesis.pdf
- Azevedo, A. I. R. L., & Santos, M. F. (2008). KDD, SEMMA, and CRISP-DM: A parallel overview. *Proceedings of the IADIS European Conference in Data Mining*, 182–185.
- Bellaachia, A., & Guven, E. (2006, April). Predicting breast cancer survivability using data mining techniques. *Proceedings of the Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining*. 1-4.

- Benitez, K., & Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association*, 17(2), 169-177.
- Bhuvaneswari, T., Prabakaran, S., & Subramaniaswamy, V. (2015). An effective prediction analysis using J48. *Asian Research Publishing Network (ARPN) Journal of Engineering and Applied Sciences*, 10(8), 3474-3480.
- Blake, R., & Mangiameli, P. (2011). The effects and interactions of data quality and problem complexity on classification. *ACM Journal of Data and Information Quality*, 2(2), 1-28.
- Bostwick, D. G. & Burke, H. B. (2001). Prediction of individual patient outcome in cancer. *Cancer*, 91, 1643–1646. doi:10.1002/1097-0142(20010415)91:8+<1643::AID-CNCR1177>3.0.CO;2-I
- Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7), 1145-1159.
- Bramer, M. (2007). *Principles of data mining*. London: Springer.
- Brenner, H., Gefeller, O., & Hakulinen, T. (2002). A computer program for period analysis of cancer patient survival. *European Journal of Cancer*, 38(5), 690-695.
- Buja, A., & Lee, Y. S. (2001, August). Data mining criteria for tree-based regression and classification. *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, 27-36.
- Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004, July). Ensemble selection from libraries of models. *Proceedings of the ACM 21st International Conference on Machine Learning*, 18-27.
- Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial Intelligence in Medicine*, 26(1), 1-24.
- Cox, D. R., & Oakes, D. (1984). *Analysis of survival data* (21). CRC Press.
- Dash, M. & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis* 1(3), 131–156.
- Delen, D., Walker, G., & Kadam, A. (2005, June). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*, 34(2), 113-27.

- Demšar, J., Zupan, B., Aoki, N., Wall, M. J., Granchi, T. H., & Beck, J. R. (2001). Feature mining and predictive model construction from severe trauma patient's data. *International Journal of Medical Informatics*, 63(1), 41-50.
- Dhir, P. & Garg, S. (2017). Survey on Cloud Computing and Data Masking Techniques. *International Journal of Innovation & Advancement in Computer Science*, 6(4), 1-7.
- Dimitoglou, G, Adams J.A., & Jim C.M. (2012). Comparison of the C4.5 and a Naïve Bayes classifier for the prediction of lung cancer survivability. arXiv preprint arXiv:1206.1121.
- Duggan, M. A., Anderson, W. F., Altekruze, S., Penberthy, L., & Sherman, M. E. (2016). The surveillance, epidemiology, and end results (SEER) program and pathology: Toward strengthening the critical relationship. *The American Journal of Surgical Pathology*, 40(12), 94-102.
- Endo, A., Shibata, T., & Tanaka, H. (2008). Comparison of seven algorithms to predict breast cancer survival. *Journal of the Biomedical Fuzzy Systems Association*, 13(2), 11-16.
- Fan, W. (2008, June). Dependencies revisited for improving data quality. *Proceedings of the 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, Vancouver, Canada, 159-170.
- Farhangfar, A., Kurgan, L., & Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, 41(12), 3692-3705.
- Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, 31(1), 1-38.
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *Artificial Intelligence Magazine*, 17(3), 37-54.
- Fletcher, S., & Islam, M. Z. (2015). Measuring information quality for privacy preserving data mining. *International Journal of Computer Theory and Engineering*, 7(1), 21-28.
- Fogarty, J., Baker, R. S., & Hudson, S. E. (2005, May). Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. *Proceedings of Canadian Human-Computer Communications Society Graphics Interface*. 129-136. Waterloo, Ontario.

- Freund, Y., & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. *Proceedings of the 13th International Conference of Machine Learning*, 96, 148-156.
- G. K., R. K., Rabi, B. J., & TN, M. (2012). A study on dynamic data masking with its trends and implications. *International Journal of Computer Applications*, 38(6), 19-24.
- Gao, L. (2015). Analysis of employment data mining for university student based on WEKA platform. *Journal of Applied Science and Engineering Innovation*, 2(4), 130-133.
- García-Molina, H., Labio, W., & Yang, J. (1998). Expiring data in a warehouse. *Proceedings of the 24th International Conference on Very Large Databases*, New York City, New York, USA, 500-511.
- Grimmer, U., & Hinrichs, H. (2001, November). A methodological approach to data quality management supported by data mining. In E. M. Pierce and R. Katz-Haas (Eds.) *Proceedings of the Sixth International Conference on Information Quality*, Boston, Massachusetts, USA, 217–232.
- Gupta, A., Mumick, I. S., & Subrahmanian, V. S. (1993, May). Maintaining views incrementally. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Washington, DC, USA, 157-166.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1), 10-18.
- Han, J., Cai, Y., & Cercone, N. (1993). Data-driven discovery of quantitative rules in relational databases. *IEEE Transactions on Knowledge and Data Engineering*, 5(1), 29-40.
- Hankey, B. F., Ries, L. A., & Edwards, B. K. (1999, December). The surveillance, epidemiology, and end results program: A national resource. *Cancer Epidemiology Biomarkers Prevention*, 8(12), 1117-1121.
- Hart, C. (1998). *Doing a literature review: Releasing the social science research imagination*. London, UK: Sage Publications.
- Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Second Edition. New York, USA: Springer.
- Haug, A., & Arlbjørn, J. S. (2011). Barriers to master data quality. *Journal of Enterprise Information Management*, 24(3), 288-303.

- Hipp, J., Güntzer, U., & Grimmer, U. (2001, May). Data quality mining - Making a virtue of necessity. *Proceedings of the Sixth ACM SIGMOD International Conference on Management of Data and Workshop on Research Issues in Data Mining and Knowledge Discovery*, Santa Barbara, California, USA, 52-57.
- Hu, H., Li, J., Plank, A., Wang, H., & Daggard, G. (2006, November). A comparative study of classification methods for microarray data analysis. *Proceedings of the fifth Australasian conference on data mining and analytics*, 61. 33-37.
- IBM Knowledge Center (2016). Using data masking with table space cloning. http://www.ibm.com/support/knowledgecenter/SSAUUE_3.1.0/com.ibm.db2tools.ckz31.doc.ug/ckzucon_tscloning_masking.htm
- Islam, M. Z., Barnaghi, P., & Brankoviz, L. (2003). Measuring data quality: Predictive accuracy vs. similarity of decision trees. *Proceedings of the Sixth International Conference on Computer and Information Technology*, 2, Dhaka, Bangladesh, 457-462.
- Iyer, V. (2013). *Ensemble stream model for data-cleaning in sensor networks* (Doctoral dissertation). Retrieved from Florida International University. <http://search.proquest.com.ezproxylocal.library.nova.edu/docview/1496775131?accountid=6579>. (1496775131)
- Jeusfeld, M. A., Quix, C., & Jarke, M. (1998). Design and analysis of quality information for data warehouses. *Conceptual Modeling*, 349-362.
- Kalariya, D. C., Shah, V., & Vala, J. (2015). Association rule hiding based on heuristic approach by deleting item at RHS side of sensitive rule. *International Journal of Computer Applications*, 122(8), 25-28.
- Keshavamurthy, B. N., Khan, A. M., & Toshniwal, D. (2013). Privacy preserving association rule mining over distributed databases using genetic algorithm. *Neural Computing and Applications*, 22(1), 351-364.
- Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI'95)*, 14(2), 1137-1145.
- Kononenko, I., Bratko, I., & Kukar, M. (1997). Application of machine learning to medical diagnosis. In Bratko, I., Michalski, R. S., & Kubat, M. (1999). *Machine learning and data mining: methods and applications*, 389, 408. Chichester: John Wiley & Sons, 1999.
- Kumari, M., & Godara, S. (2011). Comparative study of data mining classification methods in cardiovascular disease prediction. *International Journal of Computer Science and Technology*, 2(2), 304-308.

- Kurgan, L. A., & Musilek, P. (2006). A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review*, 21(01), 1-24.
- Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information & Management*, 40(2), 133-146. <http://web.cba.neu.edu/~ywlee/publication/aimq.pdf>
- Levy, Y., & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science: International Journal of an Emerging Transdiscipline*, 9(1), 181-212.
- Lowd, D., & Domingos, P. (2005, August). Naive Bayes models for probability estimation. *Proceedings of the 22nd ACM International Conference on Machine Learning*, 529-536.
- Lu, W. & Miklau, G. (2008, August). Auditguard: A system for database auditing under retention restrictions. *Proceedings of the Very Large Database Endowment*, 1(2), Auckland, New Zealand, 1484-1487.
- Luebbers, D., Grimmer, U., & Jarke, M. (2003, September). Systematic development of data mining-based data quality tools. *Proceedings of the 29th International Conference on Very Large Databases*, 29, Berlin, Germany, 548-559.
- Luengo, J., García, S., & Herrera, F. (2012). On the choice of the best imputation methods for missing values considering three groups of classification methods. *Knowledge and Information Systems*, 32(1), 77-108.
- Malazizi, L., Neagu, D., & Chaudhry, Q. (2006). A data quality assessment algorithm with applications in predictive toxicology. *Proceedings of the International Multiconference on Computer Science and Information Technology*, 131-140.
- Mazalu, R., Cechich, A., & Martin, A. (2013). Automatic profile generation for visual-impaired users. *Proceedings of the 14th Argentine Symposium on Software Engineering*, 142-153.
- McGarry, K. (2005). A survey of interestingness measures for knowledge discovery. *Knowledge Engineering Review*, 20(1), 39-61.
- Motiwalla, L., & Li, X. B. (2013). Developing privacy solutions for sharing and analyzing healthcare data. *International Journal of Business Information Systems*, 13(2), 10.1504/IJBIS.2013.054335. <http://doi.org/10.1504/IJBIS.2013.054335>
- Murugan, S. A., & Kannan, M. (2013, July). A novel approach for analysis of breast cancer and mental health using various data mining tools. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(7), 2554-2558.

- Mutter, S., Hall, F., & Frank, E. (2004, December). Using classification to evaluate the output of confidence-based association rule mining. *Proceedings of the 17th Australian Joint Conference on Artificial Intelligence*, Cairns, Australia, 538-549.
- Narayanan, A., & Shmatikov, V. (2006). How to break anonymity of the Netflix prize dataset. arXiv preprint cs/0610105.
- Ogigau-Neamtiu, F. (2016). Tokenization as a data security technique. *Zeszyty Naukowe AON*, 2(103), 124-135.
- Palepu, R. B., & Rao, D. K. S. (2012). Meta data quality control architecture in data warehousing. *International Journal of Computer Science, Engineering and Information Technology*, 15-24. doi: 10.512/ijcseit.2012.2402
- Patil, T. R., Sherekar, S. S. (2013). Performance analysis of Naïve Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications*, 6(2), 256-261.
- Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, 45(4), 211-218.
- Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37-63.
- Prandini, M., Campi, A., Marzolla, M., & Melis, A. (2014). Data sanitization in a clearing system for public transport operators (Doctoral dissertation). Retrieved from Alma Mater Studiorum, Universita Di Bologna. http://amslaurea.unibo.it/7248/1/andrea_melis_tesi.pdf
- Prokosch, H. U., & Ganslandt, T. (2009). Perspectives for medical informatics. *Methods in Information Medicine*, 48, 38-44.
- Quinlan, R. (1996). Bagging, boosting, and c4.5, *AAAI/AAI*, 1, 725-730.
- Quinlan, J. R. (2014). *C4.5: programs for machine learning*. Elsevier.
- Ramakrishnan, T., Jones, M. C., & Sidorova, A. (2011). Factors influencing business intelligence (BI) data collection strategies: An empirical investigation. *Decision Support Systems*, 52(2), 486-496.
- Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 41(2), 79-82.

- Rosenberg, J., Chia, Y. L., & Plevritis, S. (2005). The effect of age, race, tumor size, tumor grade, and disease stage on invasive ductal breast cancer survival in the US SEER database. *Breast cancer research and treatment*, 89(1), 47-54.
- Roski, J., Bo-Linn, G. W., & Andrews, T. A. (2014). Creating value in health care through big data: Opportunities and policy implications. *Health Affairs*, 33(7), 1115-1122.
- Salama, G. I., Abdelhalim, M., & Zeid, M. A. E. (2012). Breast cancer diagnosis on three different datasets using multi-classifiers. *International Journal of Computer and Information Technology*, 1(1), 36-43.
- Salkind, N. J. (2012). *Exploring research*. Eighth Edition. New York City, New York, USA: Pearson.
- Scalzo, B., Burleson, D., K., Fernandez, C., Ault, M., & Kline, K. (2007). *Database benchmarking: Practical methods for Oracle & SQL server (IT In-focus series)*. North Carolina, USA: RampantTech,
- Sekaran, U., & Bougie, R. (2013). *Research methods for business: A skills building approach*. Sixth Edition. Chichester, West Sussex, UK: John Wiley & Sons.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage Learning.
- Shafique, U., & Qaiser, H. (2014). A comparative study of data mining process models (KDD, CRISP-DM and SEMMA). *International Journal of Innovation and Science Research*, 12(1), 217-222.
- Shankaranarayanan, G. A. N. E. S. A. N., & Cai, Y. (2006). Supporting data quality management in decision-making. *Decision Support Systems*, 42(1), 302-317.
- Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008, August). Get another label? Improving data quality and data mining using multiple, noisy labelers. *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Las Vegas, Nevada, USA, 614-622.
- Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of data warehousing*, 5(4), 13-22.
- Shmueli, G., Patel, N. R., & Bruce, P. C. (2010). *Data mining for business intelligence: concepts, techniques, and applications in Microsoft Office Excel with XLMiner*. John Wiley & Sons.

- Sidi, F., Shariat Panahy, P. H., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012, March). Data quality: A survey of data quality dimensions. *Proceedings from the 2012 IEEE International Conference on Information Retrieval & Knowledge Management (CAMP)*, Kuala Lumpur, Malaysia, 300-304. doi: 10.1109/InfRKM.2012.6204995
- Student. (1927). Errors of routine analysis. *Biometrika*. 19 (1/2), 151–164. doi: 10.2307/2332181
- Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2013), National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2016, based on the November 2015 submission. <http://seer.cancer.gov>
- Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05), 557-570.
- Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Person Education. Inc., New Delhi.
- Tiwari, M., Jha, M. B., & Yadav, O. (2012). Performance analysis of data mining algorithms in WEKA. *International Organization of Scientific Research Journal of Computer Engineering*, 6(3), 32-41.
- U.S. Department of Health and Human Services, National Institutes of Health. (2015) *HIPAA privacy rule and its impact on research*. Retrieved from https://privacyruleandresearch.nih.gov/pr_08.asp
- U.S. Department of Health and Human Services. 2013 Research. Accessed April 14, 2016.
- Vinogradov, S., & Pastyak, A. (2012). Evaluation of data anonymization tools. *Proceedings from the International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA)*, 163-168.
- Wang, R. (2009, September). Raising the bar on corporate iq: \$1 million at a time. *The Total Data Warehousing Institute (TDWI) Meeting*.
- Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers, *Journal of Management Information Systems* 12(4), 1996, 5-34.
- Widom, J. (1995, November). Research problems in data warehousing. *Proceedings of the Fourth International Conference on Information and Knowledge Management*, Baltimore, Maryland, USA, 25-30.

- Wilson, R. L., & Rosen, P. A. (2003). Protecting data through perturbation techniques: The impact on knowledge discovery in databases. *Journal of Database Management*, 14(2), 14-26.
- Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical machine learning tools and techniques*. Third Edition. Burlington, Massachusetts, USA: Morgan Kaufmann.
- Wu, M. C., & Buchmann, A. P. (1997, March). Research issues in data warehousing. *Proceedings of the International Conference on Databases in Office, Engineering and Science*, Ulm, Germany.
- Xue, B., Zhang, M., & Browne, W. N. (2012, January). Single feature ranking and binary particle swarm optimization based feature subset ranking for feature selection. *Proceedings of the 35th Australasian Computer Science Conference*, 122, Melbourne, Australia, 27-36.
- Yang, Q., & Wu, X. (2006). 10 challenging problems in data mining research. *International Journal of Information Technology & Decision-Making*, 5(04), 597-604.
- Zaki, M. J., & Meira Jr, W. (2014). *Data mining and analysis: Fundamental concepts and algorithms*. Cambridge University Press.
- Zelič, I., Kononenko, I., Lavrač, N., & Vuga, V. (1997). Induction of decision trees and Bayesian classification applied to diagnosis of sport injuries. *Journal of medical systems*, 21(6), 429-444.
- Zlotnik, A., Gallardo-Antolín, A., & Martínez, J. M. M. (2015, February). Calculating classifier calibration performance with a custom modification of WEKA. *Proceedings of the Fourth International Conference on Integrated Information*, 1644, 128-132.